xCoAx 2023 11th Conference on Computation, Communication, Aesthetics & X
2023.xCoAx.org
Weimar, Germany

Creative Amplifiers: Augmenting Human Creativity with Text-to-Image Generators

Ludwig Zeller
ludwig.zeller@fhnw.ch
Institute Digital Communication Environments HGK FHNW
Basel, Switzerland

DOI 10.34626/xcoax.2023.11th.120

This paper examines the extent to which deep-learning-based generative programs, particularly text-to-image generators, support human creativity in the sense Margaret Boden's definition. This discussion is supported by a brief introduction to the technical workings of denoising diffusion-based text-to-image generators. The analysis reveals that while these networks lack the autonomous ability to evaluate their designs and conduct exploratory design processes, they can nonetheless be considered complex tools that support human creativity by offering both accessible and powerful means of text-to-image translations. The paper then broadens its focus to a more general discussion of the potential impact of such assistance on creative labor, particularly in the design disciplines. Finally, the paper identifies the democratization of creativity as a larger disruptive force for creative labor than automatization, as professional workers might soon be competing with a larger, less trained workforce.

Keywords: Text-to-image Generators, Democratization of Creativity, Deep Learning, Denoising Diffusion.

1. Introduction

This paper aims to explore the question in which regard the recent wave of text-to-image generators can support human creativity. These programs have become increasingly popular, with examples such as DALL-E 2,[1] Craiyon (formerly known as DALL-E Mini),[2] Midjourney[3] and open-source efforts such as Stable Diffusion.[4] By examining their capabilities and limitations, this paper aims to shed light on the potential for such networks to be considered 'augmentations' of human creativity (cf. Griebel et al. 2020).

Current text-to-image generators primarily employ a 'denoising diffusion' process, often referred to simply as 'diffusion.' This process diverges technically from generative adversarial networks (GANs) and tends to exhibit a superior level of visual diversity and photographic realism. Predecessors of the current wave of diffusion-based text-to-image generators have combined OpenAI's text-and-image encoder CLIP with GAN generators such as DeepMind's BigGAN (Brock et al. 2019) resulting in Ryan Murdock's Big Sleep,[5] with Nvidia's StyleGAN (Karras et al. 2019) resulting in StyleCLIP (Patashnik et al. 2021) or with VQGAN (Esser et al. 2021) resulting in VQGAN-CLIP (Crowson et al. 2022). These combinations fostered the interaction between the generator and CLIP, which optimized the output of the generator through successive iterations to conform most precisely to the 'expectations' of CLIP. In addition to these predecessors, other diffusion-based variations of this approach exist, which are not publicly accessible and currently only available in a limited beta trial, such as Google's Imagen (Yu et al. 2022).

To assess how the current generation of text-to-image generators support human creativity, it's helpful to study a range of examples. Since 2022, social media platforms have been literally inundated with images generated by these programs. For example, when provided with a prompt like "A still of Kermit the frog in Stranger Things 2016," a text-to-image generator might produce an image of Kermit resembling the fictional character Joyce Byers from Stranger Things, dressed in 1980s attire and exuding an eerie mood.[6] Other noteworthy examples include illustrations that were generated to resemble creamy soup, which was a happy accident resulting from the prompt "Bowl of soup in the style of Aubrey Beardsley," an Art Nouveau artist.[7] Another interesting application is the creation of historical reenactments using prompts like "GoPro footage of the French Revolu-

[1] See https://openai.com/dall-e-2

[2] See https://www.craiyon.com

[3] See https://midjourney.com

[4] See https://github.com/Stability-AI/stablediffusion

[5] See https://github.com/lucidrains/big-sleep

[6] See https://twitter.com/HvnsLstAngel/status/1531506455714492416

[7] See https://twitter.com/djbaskin/status/1497763195187982337

tion," producing dramatic, fish-eye views of encounters between the revolutionaries and the royalists. This is achieved by treating optical qualities as a visually learned, transferable style entity.[8]

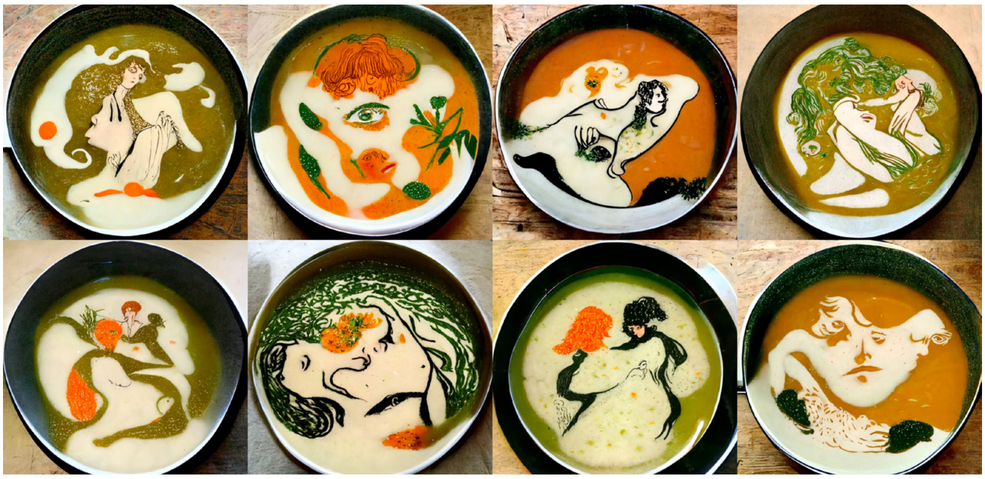

Figure 1: Results of the prompt "Bowl of soup in the style of Aubrey Beardsley" generated by Danielle Baskin in 2022.

Such examples show the relevance of the question to what capacity these text-to-image generators can support human creativity, given how surprising and convincing these examples are. To elaborate this question further, it is important to establish a working definition of creativity. One widely accepted and often used notion proposed by Margaret Boden, an expert in the field of creativity research, is that a creative idea is one that is novel, surprising, and valuable (Boden 1998).[9] The standards for what is considered novel, surprising, or valuable may differ from person to person, across different disciplines, or for humanity as a whole. Creativity is a subjective and contextual term that varies based on perspective. According to Boden, there are three main strategies for achieving creativity. The first is the *combination of existing ideas*, where new concepts are formed by combining existing ones. Secondly, there is the *exploration of conceptual spaces*. Conceptual spaces refer for instance to genres in music, styles and methods in art and design, or other ways of approaching and understanding the world. And lastly, the approach considered the most creative is the *transformation of conceptual spaces*. This involves altering existing perspectives to create something new, by creating a novel approach that opens up more possibilities and changes the conceptual space itself.

[8] See https://twitter.com/timsoret/status/1560339610588282880

[9] Boden's definition holds a prominent position in the field of computer science and is rooted in the perspectives on creativity of Joy Paul Guilford and Alex Osborne, which are inclined towards applied innovation. However, there are alternative definitions that align with a more humanist approach, drawing upon the works of John Dewey and Alfred North Whitehead. See (Still & D'Inverno 2016) for additional information on this dichotomy. Boden's definition holds a prominent position in the field of computer science and is rooted in the perspectives on creativity of Joy Paul Guilford and Alex Osborne, which are inclined towards applied innovation. However, there are alternative definitions that align with a more humanist approach, drawing upon the works of John Dewey and Alfred North Whitehead. See (Still & D'Inverno 2016) for additional information on this dichotomy.

Furthermore, the literature on human-computer interaction distinguishes between three types of 'creative' software: systems that exhibit autonomous artificial creativity, co-creative systems, and support systems that augment human creativity. Systems of autonomous artificial creativity would be able to behave similar to the creativity known from humans and would be able to fulfill all the aspects of Boden's terminology. It is questionable if such a degree of artificial creativity has ever been — or will ever be — reached. For such reasons, achieving true creativity by means of artificial intelligence is often considered its "final frontier" (Colton & Wiggins 2012). Nonetheless, for Margaret Boden such systems can at least "appear" to be creative (Boden 2004, 7). In the case of co-creative systems, multiple agents autonomously contribute creative inputs to interact with one another, usually involving at least one human in the loop (Davis et al. 2015; Karimi et al. 2019). Creative support systems on the other hand are about enabling and augmenting creativity in humans, e.g. by enhancing existing creative skills or opening up new one (Nakakoji 2006).

Based on these definitions it can be noted that the current generation of text-to-image generators cannot be regarded as fully 'creative', since they are not operating autonomously, are not critically reflecting their outputs and are not conducting explorative or transformative activities on their conceptual spaces.[10] Creativity involves the ability to assess and make judgments on the existing norms, conventions, and paradigms within the domain in which one is operating (cf. Colton et al. 2015). This is because any creative endeavor requires a response to the established body of knowledge and practices in the field, which must be considered conceptually and contextually. This precludes text-to-image generators from engaging in autonomous or co-creative interactions on equal footing. Nonetheless, they can be analyzed as systems that support human creativity instead, which is the focus of this paper.

With this understanding of creativity, let's revisit the examples mentioned earlier. The images of Kermit the Frog placed within movies are based on combining existing concepts that are triggered by the text prompt. The GoPro footage of the French Revolution is a unique image created by combining the historical event with the optical characteristics of an action camera lens. The bowl of soup in the style of Art Nouveau artist Aubrey Beardsley showcases the ambiguous nature of combinatorial creativity. Although the prompt was to render a bowl of soup in the style of the artist, the program incorporated Beardsley's motifs into a modern-day photograph of creamy soup instead, possibly inspiring a human beholder to inventing a new illustration technique. Again, this is another example of combinatorial creativity. In conclusion, text-to-image generators provide

[10] I express my gratitude to the anonymous reviewer who brought this to my attention.

a means of rendering images that correspond to a specified prompt. As previously discussed, evaluating the originality and worth of these images is subjective and dependent on the context in which they are used. Currently, they are mostly viewed as entertaining and high-quality surprises on social media, which contributes to their perceived value.

At least in this regard, the complexity displayed by the program in executing the prompt should be recognized. After all, generating these images is not a simple process of translating text to image, but requires the program to apply a substantial amount of semantic and visual 'knowledge'. Skeptics may argue that attributing knowledge to computers is a misconception of the disparity between syntax and semantics, as machine learning is primarily the result of finely tuned statistical analysis and thus 'guesswork'. However, physicist Sabine Hossenfelder recently proposed an argument in a think piece regarding the notorious "statistical parrot"[11] ChatGPT that challenges this view.[12] She opposes John Searle's famous "Chinese Room Argument" (Searle 1999) suggesting instead that the ability to generalize data indeed indicates a degree of knowledge. In the case of text-to-image generators, these systems can adapt visual and textual elements to fit cohesively and reasonably into an extensive range of previously unseen contexts, in a 'zero-shot' manner. But while the visual solutions generated by the program are novel and unexpected, they are only possible through the collaboration between the human 'prompter' and the visual translation program. Without input from a human prompter, text-to-image generators are limited to clever interpretations of text into images and cannot produce meaningful solutions independently. To further understand the mechanisms that drive and limit the capability of such generators to offer novelty and value, it is beneficial to have some understanding of the technical workings of these generators.

2. Working Principles

2.1. Convolutional Neural Networks

A foundational technology of diffusion-based text-to-image generators are convolutional neural networks. An influential example for this type of network is the AlexNet (Krizhevsky et al. 2012), which is often cited as the starting point for the recent renaissance of deep learning-based artificial intelligence. The principle of this network is that, for the purpose of pattern and object recognition, an image is input into the first layer of the network. Over several layers, a convolution algorithm is then applied to the image using filter kernels, which are activated in certain areas of the image where patterns

[11] For the term see (Bender et al. 2021).

[12] See https://www.youtube.com/watch?v=cP5zGh2fui0

are found (cf. Zeiler & Fergus 2014). These filters, which may appear strange to humans, are not programmed or designed by manual effort but are learned by the network during the training stage, based on the training data set provided. As the network progresses through higher layers, more complex patterns, such as dog noses and human faces, are learned and represented within the network.

Essentially, convolutional filtering is also the working principle of well-known functions such as image sharpening in image editors. In that case, the filter kernel works as a general edge detector and increases the intensity of all edges in the pixel representation of an image. Similarly in the case of convolutional neural networks, the learned filter kernels are emphasized in a given image through the process of convolution, thus yielding activation maps of the features that the filter kernels shall represent. On its way through the network, the input image is downscaled several times through a process called 'pooling'. Additional convolution layers are applied, and the image is eventually transformed into an activation pattern, carrying information on which filters of the many filters learned by the network were activated. Finally, this information is represented in a fully connected layer, often referred to as the latent or feature space, which consists of a list of floating-point values. In the case of AlexNet, this feature space can then be used for tasks such as object classification.

The famous DeepDream process by Alexander Mordvintsev is a technique that allows for the visualization of features that a convolutional neural network (CNN) has learned during its training process.[13] Originally, it was developed as a tool for debugging CNNs, by highlighting the areas of an image that activate certain filters within the network. This technique has been used to generate a specific type of imagery, often described as 'psychedelic,' that has been used in some early forms of deep learning art.

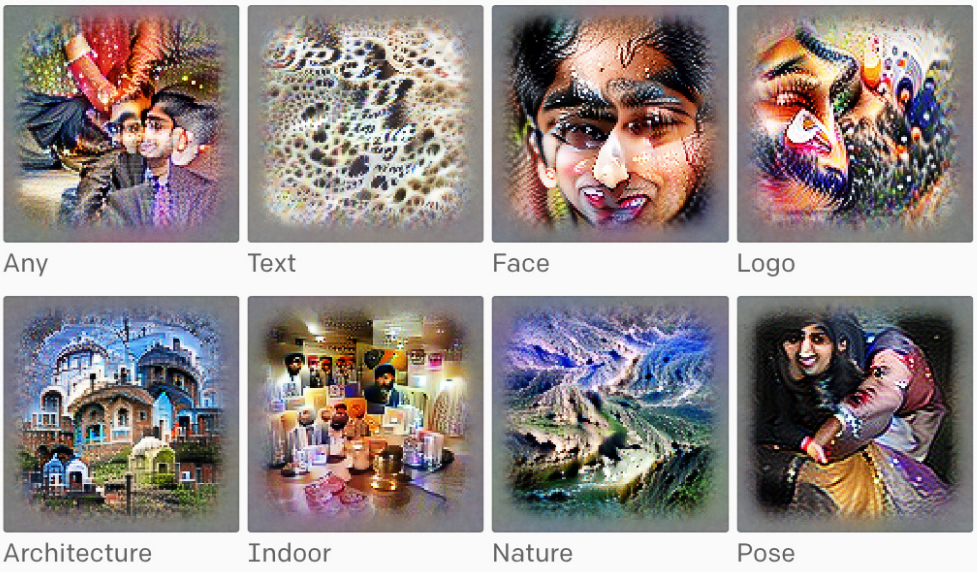

Figure 2: Visualization of a "multimodal neuron" for India inside OpenAI's CLIP network (Goh et al. 2021). Taken from a blog post on https://openai.com/blog/multimodal-neurons.

[13] https://github.com/google/deepdream

In today's neural networks, the number of visual filters that are being learned is very large and also highly interconnected and context dependent. With the Microscope tool by OpenAI, one can inspect the filters learned by popular networks such as ResNet and CLIP. In the case of the latter an example would be USA-related patterns learned by the network that are structured into subcategories including various typographies, faces, logos, and depictions of typical cultural artifacts.[14] Through the diversity of these learned features in today's neural networks, complex visual concepts can be represented as vector lists of numbers in their feature space, which in turn allows for the algorithmic processing of visual phenomena.

2.2. Generative CNNs

Let's now move back to discuss text-to-image generators and their potential for supporting human creativity. So far, the process of encoding pixel space into feature space has been examined. Additionally, an encoder-decoder structure can be employed, which utilizes deconvolution and upscale algorithms to reconstruct an image back from the latent space. In the encoder, an input image activates filters and is translated into a feature vector representation as stated above. In the decoder, the image is reconstructed based on this feature vector, but will never be fully identical to the input image since the learned filters are necessarily biased towards a statistical average of the training set. This outcome is desirable in these networks, as exact replication would indicate overfitting and lack of generalization and could also potentially lead to copyright infringement issues (cf. Carlini et al. 2023).

A helpful example of such a latent space decoder can be seen in the Face Editor tool by CodeParade.[15] In this case, a network was trained using a collection of approximately 30-40 portrait images of high school students. The resulting tool utilizes a mixing desk interface to manipulate facial features that have been sorted for their visual significance. Each feature is represented by a slider that encodes visual features into a feature space vector, with the first slider affecting the greatest number of pixels and each subsequent slider affecting fewer pixels. The most influential feature identified by the network is shirt color, followed by sex, head position, body height and hair density.

Additionally, it is possible to manipulate feature vectors in a variety of ways using vector mathematics. Operations such as adding, subtracting, interpolating, and averaging can be performed on the vectors for extracting and applying semantic concepts as vectors. One example of this is SpaceSheets by Bryan Loh and Tom White, which uses a spreadsheet interface to select and combine multiple images

[14] See https://microscope.openai.com/models
[15] See https://www.youtube.com/watch?v=4VAkrUNLKSo

based on their feature vector representations and simple arithmetic operations (Loh & White 2018). Furthermore, it is possible to isolate specific features, such as a smile, through subtraction of feature vectors. This isolated 'smile vector,' can then be applied to make images of neutral faces appear more friendly or smiling.

Beyond auto-encoders, significant advancements have been made in the field of photorealism through the use of generative adversarial networks (GANs) in recent years. These networks have proven to be capable of producing high-quality images in response to text inputs. One example of this is StyleGAN (Karras et al. 2019), which is able to produce highly realistic images of faces with fast computation and sampling times. However, GANs have traditionally struggled with generating a diverse range of subjects. This problem was partially addressed by the development of BigGAN (Brock et al. 2019), albeit without offering the same level of realism. Only recently, denoising diffusion models have emerged that are able to generate high-quality samples for a diverse range of subjects, powering the current generation of text-to-image generators.

2.3. Denoising Diffusion Models

The denoising diffusion architecture utilizes an encoder-decoder structure at its core (cf. Ho et al. 2020).[16] The encoder is trained to filter out noise from an image, again activating filters that have been learned about specific features. This results in the image being represented in a feature space, which can then be decoded to produce a denoised image. The iterative activity of these models can be observed, as the image becomes clearer and clearer during the denoising process after departing from a fully noised initial image. Crucially, denoising diffusion-based text-to-image generators combine this denoising with an additional text encoder, which encodes text into a feature space that is then used in a multi-step process to condition the above-mentioned denoising process. This conditioning limits the search space of visual patterns learned within the network and allows to control the creation of images from another modality.[17]

To compare this working principle with human psychology, it is common for individuals to perceive familiar patterns or shapes in objects, such as clouds. This phenomenon, known as pareidolia, is influenced by prior inputs or mental associations. For example, if someone is told to look for a specific image such as a dog in a cloud,

[16] The popular Stable Diffusion by stability.ai is based on so-called Latent Diffusion (Rombach et al. 2022), which owes its naming to the fact that its denoising diffusion auto-encoder does not operate on images directly, but on latent space representations of these instead. According to the authors, packing a latent space in a latent space reduces the dimensionality of the data and therefore speeds up the processing.

[17] The addition of further conditionings in other modalities such as spatial depth and body pose is currently an active field of research (cf. Zhang & Agrawala 2023).

it may be easier for them to perceive that image. This concept is known as priming, where a prior input influences subsequent thoughts and mental imagery. Pareidolia is particularly common in the perception of faces, but it can also occur with other shapes and patterns found in nature. Although it is usually inappropriate to anthropomorphize deep learning programs, the working principle of text-guided denoising diffusion models can be metaphorically compared to priming and pareidolia common in human psychology.

The advancement of new architectures and the accessibility of large-scale, open-source datasets are crucial for the quality of today's generators. OpenAI was among the first organizations to utilize web-scale datasets for their CLIP model (Radford et al. 2021). However, these datasets have been kept private. The LAION dataset (Schuhmann et al. 2022) used by Stable Diffusion, in contrast, is a publicly available list of ca. five billion image URLs with text descriptions. However, the legality of web scraping often depends on the purpose of the scraping and training, particularly whether it is intended for scientific research or commercial endeavors.

It appears that the size of neural networks is a significant factor in their performance. Google researchers demonstrate with their text-to-image generator Pati, which uses a GAN as its generator but is unfortunately not publicly accessible at the time of this writing, that while a network trained with 350 million parameters can generate a prompt such as "A map of the United States made out of sushi on a table next to a glass of red wine," the results are not fully convincing.[18] As the number of parameters increases, the results become increasingly accurate. With 20 billion parameters, the network is able to compose a description of a scene featuring a kangaroo wearing an orange hoodie and blue sunglasses in front of the Sydney Opera House and holding a sign on the chest that says "Welcome friends." This level of performance is currently a challenge for publicly available neural networks at the state of this writing.

In summary, text-to-image generators are trained on large datasets of text and images sourced from the internet. They establish optimized relationships between text and images and have the ability to cover a wide range of concepts. However, they may not include concepts that occur not frequently enough in the training data. Figure 3 shows a data mapping generated by David McLure that contains 12 million captions from the LAION-Aesthetics dataset, a subset of LAION-5B that was filtered for aesthetic appeal based on human judgments.[19] The captions are encoded into feature space vectors using the text encoder of CLIP. These vectors are then clustered using a projection technique called UMAP (McInnes et al. 2018), with textu-

[18] See https://imagen.research.google

[19] See https://huggingface.co/datasets/dclure/laion-aesthetics-12m-umap

al similarities determining the proximity of the captions in the final visualization. The visualization demonstrates that certain semantic concepts, such as men's and women's clothing, are closely related but not completely overlapping. There is a gap between them, representing concepts that are not clearly associated to the binary sexes. Notably, there is a cluster of celebrities present in the visualization where each celebrity is represented by an individual 'island,' due to the large number of captions and images associated with them.

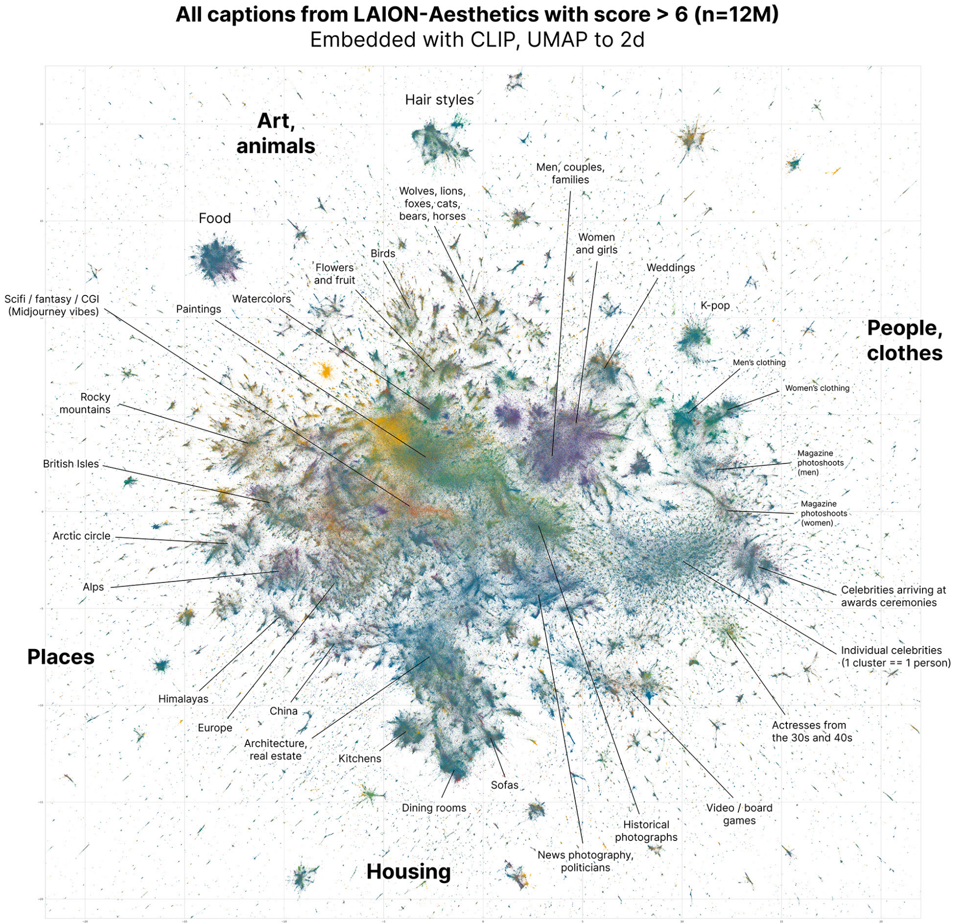

Figure 3: Visualization of 12 million CLIP-encoded captions from the LAION-Aesthetics dataset, created by David McLure in 2022.

3. Hybrid Creativity Cultures

After this technical overview, let's return to the primary question of this paper, which is in what ways can text-to-image generators support human creativity. According to Kevin Kelly, the creative artistry of using text-to-image generators is comparable to that of painting and photography:

> This new art resides somewhere between painting and photography. It lives in a possibility space as large as painting and drawing—as huge as human imagination. But you move through the space like a photographer, hunting for discoveries. Tweaking your prompts, you may arrive at a spot no one has visited before, so you explore this area slowly, taking snapshots as you step through. The territory might be a subject, or a mood, or a style, and it might be worth returning to. The art is in the craft of finding a new area and setting yourself up there, exercising

> good taste and the keen eye of curation in what you capture. (Kelly 2022)

The artist's skill lies in the exploration and manipulation of the visual-semantic clusters in the latent spaces through strategic questioning and querying. The artistry is demonstrated through the act of exploration and curation of these latent landscapes. But to which degree is this new augmented creativity structured and limited by the technical working principles of text-to-image generators?

Margaret Boden's terminology of creativity emphasizes the importance of *novelty*, which is present in text-to-image generators as they generate new images that are not identical to the ones fed into the network. However, these generated images are still dependent on the dataset that the network was trained on and cannot produce something completely different or outside of the dataset's range. Therefore, while text-to-image generators are able to produce novel images, they are ultimately limited by the data they have been trained on and cannot generate radically original or unique results. Furthermore, a possible (socially constructed) value of the images generated by text-to-image generators can be seen in their ability to be visually plausible, realistic, and pleasing, as well as being coherent with the textual description provided as input. Yet, it is important to note that these programs do not have the capability to assess the value of these images in the actual world and can only determine whether they are coherent with the text description through statistical means.

In terms of *combinatorial* creativity, the program allows for selecting and combining concepts from different semantic domains to produce an image that maximizes coherence with all the listed concepts. The text prompts in this case can be thought of as a mixing desk (as demonstrated earlier with the Face Editor project), where features are scaled up or down through combining words in sentences (instead of using faders). In summary, the text-to-image generators operate by concatenating concepts.

Additionally, the program is able to find multiple solutions for a given prompt within the specific latent space area defined by activating the concepts in a combinatorial sense. Nonetheless, it is important to note that while text-to-image generators can indeed generate variations, they cannot meaningfully move around in conceptual spaces independently, as they lack the understanding of which way to go is better or worse. These networks are primarily limited to translating texts to images. On the other hand, humans can be very good in *exploratory* creativity in the sense that Kevin Kelly sketched out. Despite the fact that these networks have acquired a significant amount of knowledge about text and images, they lack understanding of what may be considered emotionally or intellectually engaging, or how to physically construct such content. In contrast, humans can engage in meaningful explorations since we have the ability to evaluate ideas for their value in the actual world. While techniques such as linear interpolation between prompts exist, they do not fully qualify as exploratory creativity as they simply interpolate between given prompts. As a result, it is only the human user who actively explores the conceptual latent space. Nevertheless, the computer program still makes meaningful contributions by generating surprising and convincing image solutions for the prompts given.

The final form of creativity, as deemed by Margaret Boden, is *transformational* creativity. In this context, it can be stated that neither the program nor the human user can significantly alter the conceptual space of the model. Instead, the space is pre-determined by the creators of the model through the design of its architecture and the selection of its dataset. It can be emphasized that the potential for these models to achieve transformative creativity is drastically limited.

Furthermore, there are considerable limitations in terms of the scope of possible human exploration of these networks, particularly for Dall-E 2, which is hosted by OpenAI and is more inclined to avoid negative media and legal attention than Stable Diffusion. For instance, certain prompts for images depicting former US-presidents Donald Trump and Barack Obama kissing with each other will result in warnings of possible banning. However, it has been observed on social media that misspellings of the names of the former presidents could be used to generate such images despite these restrictions (see figure 4).[20] This indicates that in addition to the dataset-related limitations of the network itself, there are also human-imposed filters in place to adhere to behavioral rules and political interests.

Additionally, there are many elements that may simply not be present in the data set, either because they have not been shown enough during training to be learned, or because they are highly personal such as a specific person's face or an individual's backpack with unique stickers. Such elements are usually not available in trained models. Currently, research is being conducted on methods such as Dreambooth, which allows for the inclusion of new elements, such as specific objects, animals, or personal images, into the network through the use of transfer learning (Ruiz et al. 2022). By adding only 10–20 images, these elements can be added to the map of the conceptual space – essentially transforming it to a small degree – and making these elements available for interaction with all the con-

[20] See https://twitter.com/odedbendov/status/1550780625971548160. It should be noted that attempting to reproduce this hack is not advisable. The author of this paper was banned a few days later when attempting this as OpenAI likely realized the hack was trending on Twitter and searched for all users who used that prompt.

ceptual complexity provided by these networks. This enables the inclusion of personal elements, such as a specific person's face or an individual's backpack, for instance in order to place them in various settings, such as the Grand Canyon, Night Sky, or Boston.

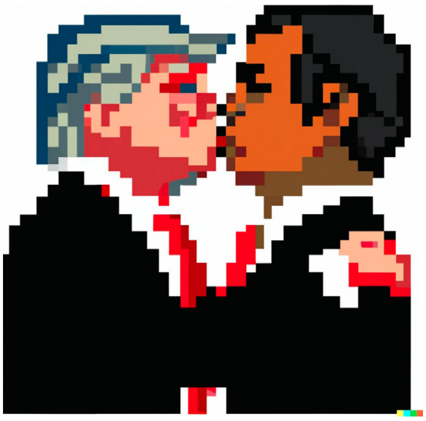

Figure 4: The prompt "President Trumpz and President Oboma kissing pixel art" is used to bypass OpenAI's usage filters, created by Oded Ben Dov in 2022.

3.1. Impact on Human Creative Labor

Returning to the topic of the impact of these developments on creative labor, it should be noted that there are services that utilize the Dreambooth method of adding specific elements, such as a person's face, to the network for generating commissioned images at a low cost. For instance, the online service photoAI by Seb Lhomme uses this principle to sell image packs for popular business and dating platforms that containing suitable profile pictures based on uploaded snapshots that are then fine-tuned using this process.[21] Although the generated images may not appear entirely genuine at the moment, it is likely that this trend will continue to expand. The automation of this form of creativity has the potential to impact human labor, particularly in the field of commercial portrait photography, in addition to the disruptions that large language models (LLMs) such as ChatGPT and LaMDA are poised to cause.

Again, Kevin Kelly (2022) expresses an optimistic viewpoint on this topic and argues that similar to how the invention of the camera led to more art and more opportunities for creating images, these AI technologies can open possibilities for more creativity and art in various fields. Prior to the invention of the camera, a portrait in the form of a painting was too costly for many individuals to afford, but with the advent of photography, more individuals were able to have their portrait taken. He suggests that these AI technologies can have a similar impact by making creative possibilities more accessible to a larger population. This is particularly relevant for small-scale projects such as running a blog, where the budget may not allow for hiring an illustrator. Kevin Kelly predicts that in social media and in areas where there is limited budget, we will see a significant increase in the use of generated images. This can be translated to mean that while the use of AI-generated images may not be an authentic product in comparison to a human-made alternative, in many cases it will be satisfactory. This democratization of illustration and picture generation, through the use of AI, will result in an even greater influx of images than what is already observed today.

Therefore, it is crucial for professionals to ensure that their work is not adversely affected by the proliferation of augmented creativity, especially in commercial contexts. It could be argued that the increased availability of such methods may reduce the value of creativity itself based on economic principles. Boden (2004, 43ff) dis-

[21] See https://photoai.me

tinguishes between two types of inventions: those that are novel to a society ("H-creative" for 'historical') and those that are novel to the inventor ("P-creative" for 'psychological'). Therefore, practitioners may maintain an advantage by generating more h-creativity – however, this poses a difficult challenge. Hence, redefining creativity in new ways might become necessary, not just from a theoretical and societal standpoint but also from an economic perspective, in order to question the current modernist and capitalist structure of creative innovation altogether (cf. Still & D'Inverno 2016; Mould 2018; Hills & Bird 2019).

3.2. Supporting Human Creativity With New Interfaces and Strategies

One possible strategy to address this trend is to actively adopt and integrate deep-learning-based creative support tools into one's creative process, potentially gaining a competitive advantage while still being able to profit from one's existing expertise and training. An example of artfully utilizing the similarities to human pareidolia inherent in generative denoising diffusion networks as mentioned earlier is by using simple sketches as input and conditioning them semantically with text prompts. For instance, figure 5 shows how a charcoal sketch of a young woman can be translated into portrait art in the style of famous painters using the image-to-image translation of Stable Diffusion. The sketch defined the coarse orientation of the face and layout of the output image, while a text prompt framed its stylistic rendition and contextualization. Thus, while generative techniques may democratize image-making, there are limitations to who can effectively create and utilize such sketches in a creative and meaningful manner. This may create an opportunity for professionals to leverage these 'creative amplifiers' in unique ways that are not accessible to everybody.

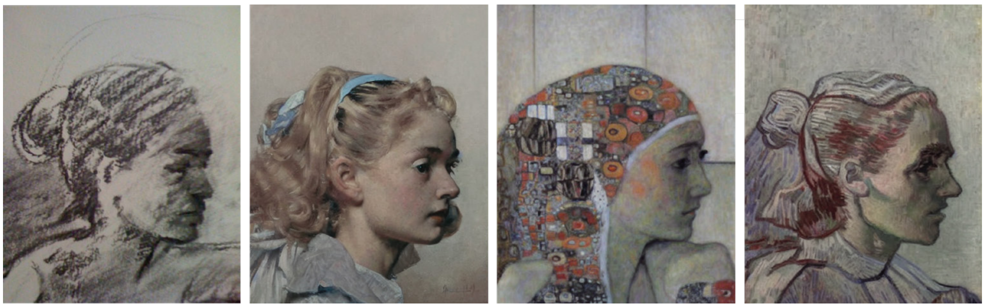

Figure 5: A charcoal sketch (left) is used to define the layout of three generated portraits in the style of Norman Rockwell, Gustav Klimt and Vincent van Gogh, created by @TomLikesRobots in 2022. See https://twitter.com/TomLikesRobots/status/1566027217892671488.

The recent collage tool developed by Joel Simon as part of his Art-Breeder project further highlights the potential of the pareidolic principle, as it is firmly focused on harnessing it.[22] The tool enables users to layer noise onto simple collages of shapes and images, increasing the degree of uncertainty of the input to the denoising models, and thus adjusting the level of freedom the generator has

[22] See https://collage.artbreeder.com

in interpreting the input in conjunction with a conditioning text prompt. This feature allows for greater control and manipulation of the final image in combination with the text prompts.

It's also worth mentioning that these technologies can serve as intermediate sources of inspiration in the design process, rather than relying solely on the generated images as the final outcome. For instance, Philip Schmidt and Stephan Weiss trained a DCGAN (Radford et al. 2016) — one of the first GANs to be used by designers and artists — on a small dataset of iconic chair designs from the 20th century and generated chair variations in a rather low visual quality compared with today's tools.[23] They then used their imagination and craft to translate these outputs into physically possible designs, materializing them first into sketches, then into miniature models and eventually one-offs for an exhibition. This highlights the gap between what these images show and what can actually be built and is an important consideration for those who work in product design or use making as a form of art or design practice.

By looking at such older projects the rapid improvement in image generation technology in recent years becomes more than obvious. This raises the question of whether it is still necessary to physically build objects at all — especially in conceptual and speculative design projects, where the primary focus has traditionally been on communicating visually instead of offering tactile functionality (cf. Dunne & Raby 2013). An additional possibility is utilizing these image generators as inspiration and visualization in participatory settings, such as workshops, as an alternative to traditional methods of ideation and communication for non-professionals who may not have the ability to draw or visualize on their own.

4. Verdict

The nature of work and expertise in any field has always evolved over time, and this will continue to be the case with the integration of AI-driven tools. With advancements in technology and the incorporation of more knowledge about physical materiality and other fields, it is possible that generative models may be able to output designs that are not only visually plausible but also functional and feasible to produce. This could have significant implications for fields such as engineering and product design. A more tangible and immediate risk scenario is given for certain forms of creative labor, which lack significant exploratory or transformative creativity, and may therefore be at risk of transformation or even obsolescence in the near future. The manual creation of illustrations according to descriptions given from a customer could be one such example. Generally speaking, the current generations of deep learning tools

[23] See https://philippschmitt.com/work/chair

may have a greater impact on professions that rely on the creation and sale of images, sound, and moving images – whereas the actual construction of physical artefacts is much less likely to be impacted.

Especially the rise of untrained labor as a result of democratization brought about by advancements in technology and AI has the potential to disrupt the labor market of many creative industries. This suggests that while some highly specialized jobs may become obsolete, it also opens up opportunities for individuals with less formal training to enter the market and perform tasks that were previously considered complex and difficult. AI-driven tools that act as creative amplifiers may democratize access to creative professions and help individuals to perform tasks that were previously reserved for highly specialized professionals. While there may always be a need for someone to determine the value or meaning of an object or design, this role may not necessarily require craft education or training as it can be done by anyone who has the ability to make such judgments.

So far, text-to-image generators cannot produce creations independently and therefore cannot be considered as having creativity. Moreover, they do not possess the capacity to assess the practicality and utility of the generated designs in real-world scenarios and applications. However, given the recent rate of progress in the field of artificial intelligence, it is at least conceivable that new systems may emerge in the future that surpass such limitations and would fully qualify as artificially creative.

5. Statement on AI-Assisted Writing

This manuscript is based on a lecture held at an internal colloquium session of the IDCE HGK FHNW in November 2022. An audio recording of the lecture was transcribed to text using OpenAI Whisper. The resulting transcript was then edited from colloquial English to formal English using ChatGPT. However, all examples, interpretations, theses, and conclusions have been researched and/or developed independently by the author.

References

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610-623.

Boden, Margaret. 1998. Artificial Intelligence Creativity and Artificial Intelligence. In: *Artifical Intelligence* 103. 347-356.

Boden, Margaret. 2004. *The Creative Mind: Myths and Mechanism.* Second edition. London: Routledge.

Brock, Andrew, Jeff Donahue, and Karen Simonyan. 2019. *Large scale GAN training for high fidelity natural image synthesis.* Conference paper at International Conference on Learning Representations 2019.

Carlini, Nicholas, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. *Extracting Training Data from Diffusion Models.* arXiv preprint arXiv:2301.13188.

Colton, Simon, and Geraint A. Wiggins. 2012. Computational Creativity: The Final Frontier?. In: *Proceedings of European Conference on Artificial Intelligence*. 21-26.

Colton, Simon, Jakob Halskov, Dan Ventura, Ian Gouldstone, Michael Cook, and Blanca Pérez-Ferrer. 2015. The Painting Fool Sees! New Projects with the Automated Painter. In: *Proceedings of the 6th International Conference on Computational Creativity, ICCC 2015*, no. October: 189-196.

Crowson, Katherine, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. 2022. VQGAN-CLIP: Open domain image generation and editing with natural language guidance. In: *Computer Vision-ECCV 2022: 17th European Conference*, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVII. 88-105. Cham: Springer Nature Switzerland.

Davis, Nicholas, Chih-Pin Hsiao, Yanna Popova, and Brian Magerko. 2015. An Enactive Model of Creativity for Computational Collaboration and Co-Creation. In: *Creativity in the Digital Age*, edited by Nelson Zagalo. Springer. 109-133.

Dunne, Anthony, and Fiona Raby. 2013. *Speculative everything: design, fiction, and social dreaming*. MIT press.

Esser, Patrick, Robin Rombach, and Björn Ommer. 2021. Taming transformers for high-resolution image synthesis. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12873-12883.

Goh, Gabriel, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. 2021. Multimodal Neurons in Artificial Neural Networks. In: *Distill* 6, no. 3: e30.

Griebel, Matthias, Christoph Flath, and Sascha Friesike. 2020. Augmented Creativity: Leveraging Artificial Intelligence for Idea Generation in the Creative Sphere. In: *ECIS 2020 Proceedings Research-in-Progress Papers*. 6-15.

Hills, Alison, and Alexander Bird. 2019. Against creativity. In: *Philosophy and Phenomenological Research* 99, no. 3. 694-713.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In: *Advances in Neural Information Processing Systems* 33. 6840-6851.

Yu, Jiahui, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, Yonghui Wu. 2022. *Scaling autoregressive models for content-rich text-to-image generation*. arXiv preprint arXiv:2206.10789.

Karimi, Pegah, Mary Lou Maher, Nicholas Davis, and Kazjon Grace. 2019. Deep Learning in a Computational Model for Conceptual Shifts in a Co-Creative Design System. In: *Proceedings of the 10th International Conference on Computational Creativity*. 17-24.

Karras, Tero, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401-4410.

Kelly, Kevin. 2022. Picture Limitless Creativity at Your Fingertips. *WIRED Magazine.* Retrieved from https://www.wired.com/story/picture-limitless-creativity-ai-image-generators/, last accessed April 23rd 2023.

Krizhevsky, Alex, Ilya Sutskever, Geoffrey Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In: *Proceedings of the Advances in Neural Information Processing Systems*, Lake Tahoe, NV, USA, 3–6 December. 1097-1105.

Loh, Bryan and Tom White. 2018. *SpaceSheets: Interactive Latent Space Exploration through a Spreadsheet Interface*. School of Design. University of Wellington Wellington, New Zealand.

McInnes, Leland, John Healy, and James Melville. 2018. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv preprint arXiv:1802.03426.

Mould, Oli. 2018. *Against Creativity.* Verso Books, London.

Nakakoji, Kumiyo. 2006. Meanings of Tools, Support, and Uses for Creative Design Processes. In: *Proceedings of International Design Research Symposium'06*. 156-165.

Patashnik, Or, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven Manipulation of Stylegan Imagery. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2085-2094.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever. 2021. Learning Transferable Visual Models from Natural Language Supervision. In: *International Conference on Machine Learning*, 8748-8763.

Radford, Alec, Luke Metz, and Soumith Chintala. 2016. *Unsupervised representation learning with deep convolutional generative adversarial networks*. Poster at International Conference on Learning Representations ICLR.

Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution Image Synthesis with Latent Diffusion Models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684-10695.

Ruiz, Nataniel, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2022. *Dreambooth: Fine Tuning Text-to-image Diffusion Models for Subject-driven Generation*. arXiv preprint arXiv:2208.12242.

Schuhmann, Christoph, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, Jenia Jitsev. 2022. *Laion-5b: An open large-scale dataset for training next generation image-text models.* Poster presentation at the Conference on Neural Information Processing NIPS 2022.

Searle, John. 1999. The Chinese Room. In: *The MIT Encyclopedia of the Cognitive Sciences.* Edited by Robert A. Wilson and Frank C. Keil. Cambridge MA, MIT Press. 115-116.

Still, Arthur, and Mark d'Inverno. 2016. A History of Creativity for Future AI Research. In: *Proceedings of the 7th International Conference on Computational Creativity, ICCC 2016*, no. June. 147-154.

Zeiler, Matthew, and Rob Fergus. 2014. Visualizing and Understanding Convolutional Networks. In: *Lecture Notes in Computer Science*. 818-833.

Zhang, Lvmin, and Maneesh Agrawala. 2023. *Adding Conditional Control to Text-to-image Diffusion Models*. arXiv preprint arXiv:2302.05543.