”Picture Perfect or Ethically Problematic?” Ethical Implications of AI-generated Images on the Example of DALL·E 2

SOFIA KALTWASSER[1] AND MANUEL MÜHLBERGER[2]

[1] *University of Potsdam*
[2] *School of Computation, Information and Technology, TU Munich*

Compiled April 15, 2023

1. INTRODUCTION

“I’m fascinated by this imagery. I love it. And it think everyone should see it” is what Jason M. Allen told CNN Business in an interview after winning first prize in the “digital arts/digitally-manipulated photography” category at the Colorado State Fair Fine Arts Competition[1]. His artwork "Théâtre D’opéra Spatial" (Figure 1) which he did not draw himself, but was rather generated by an Artificial Intelligence (AI) image generator, raises the question, whether the AI or the human has become the artist and whether Allen is really deserving of the prize. As these systems become more widespread and accessible, issues such as bias, ownership and control over the generated content will become increasingly relevant, making in-depth critical and ethical analysis vital to ensure that the advantages of this technology are achieved without any detrimental effects. This paper will, on the example of DALL-E 2, highlight a select few ethical issues which are likely to come up when using AI image generation systems en masse. Thereafter, they will be analysed based upon two fundamental moral principles, namely, the categorical imperative and utlilitarianism. It will also shed light on the current state of regulations, or often times lack thereof and juxtapose it with the seemingly arbitrary terms of service that many providers publish. Finally, an outlook will be given on current research activity and active efforts in the European Union (EU) for establishing legal liability.

Fig. 1. “Théâtre D’opéra Spatial”, Winner of the blue ribbon in the fair’s contest for emerging digital artists by Colorado State Fair’s annual art competition, generated by MidJourneyAi[2].

2. TECHNICAL BACKGROUND OF DALL-E 2

In recent years, applications for computer vision and methods for processing images have considerably benefitted from developments made possible by deep learning and AI. One of these methods is picture synthesis, which is the act of creating new images and modifying ones that

already exist. Because of its many useful applications in fields including art creation, image editing, virtual reality, video games, and computer-aided design, image synthesis is a extensive and significant field of research.
Text-conditional image models are capable of generating images from text queries and can arrange unrelated objects in a semantically plausible way. They are also called text-to-image models. One of the most popular examples of such models is Open AI's Dall-E 2 [3].
DALL-E 2 is a powerful text-to-image synthesis model, released by Open AI in July 2022, capable of extracting the semantic meaning of natural text input and translating it in a zero-shot manner [4] into high quality images, as can be seen in Figure 2. Zero-shot learning describes the process of classifying instances during normal usage that were not part of the training set [5]. In this specific case, this refers to the ability of the user to input text at will that could not have been predicted during training. It has the potential to be employed in a range of applications, like in creative design. For instance, a digital artist could significantly profit from this technology either by finding quick inspiration or by drastically increasing both the speed and the quality of his artwork if he has a certain picture of it in mind. Overall, it marks a substantial leap in the field of generative models, especially compared to its predecessors DALL-E 1 or DALL-E mini, who where one of the first models that were suitable for the mass[]However, to be able to gauge its possibilities, limitations and draw meaningful ethical conclusions, it is vital to first gain an understanding of its technical backbones. The following technical background section of this paper will provide an overview of the architecture as well as the training process of DALL-E 2.

Fig. 2. Four generated pictures to the text input: "a tapir with the texture of an accordion" [4]

A. Transformer Model

Vaswani et al. introduced the Transformer model architecture in 2017 [6] that has since become the state-of-the-art natural language processing model, whereupon many other models are based.
At its core, the transformer is a self-attention-based architecture that operates on sequences of tokens, such as words or subwords. The model contains an encoder and a decoder, each of which consist of a stack of identical layers.
Each layer in the transformer is composed of two sublayers: a self-attention mechanism and a feed-forward network. The self-attention mechanism computes a weighted sum of the input sequence, where the weights are based on the similarity between each token and every other token in the sequence. This allows the model to attend to different parts of the input sequence at different layers, and has been shown to be highly effective at capturing long-range dependencies in language.
The feed-forward network applies a set of linear and nonlinear transformations to the output of the self-attention mechanism, providing an additional layer of modeling capacity.
Apart from that, the transformer features a number of significant modifications in addition to the conventional self-attention mechanism that have proven essential to its success. Multi-head attention, which enables the model to pay attention to various points in the input sequence at once, is one such breakthrough. Another is the use of positional encodings, which provide the model knowledge about the tokens' order in the input sequence.

B. High-level Abstraction of the Text-To-Image Model

In order to encode the semantic meaning of the corresponding text, DALL-E 2 first processes a textual input description using a transformer-based language model. The multi-stage generative model which creates the final high-resolution image uses the encoded text as input to create an initial low-resolution image that is steadily improved over time, following a diffusion model. The generative model, which creates the image using a combination of transformer- and convolutional-based neural network architectures [7].

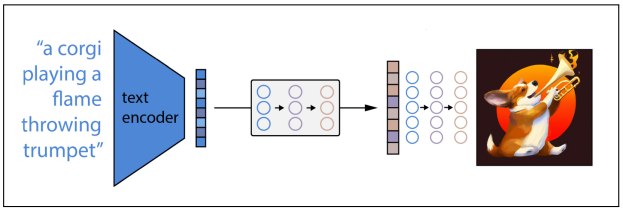

Fig. 3. A CLIP text embedding is given to an autoregressive or diffusion algorithm to create an image embedding, which is then used to condition a diffusion decoder to create the final picture in the text-to-image production process.[modified from [8]]

C. Training Data

250 million images and their related textual descriptions were collected from the internet as training data for the

first model of DALL-E.
It was then trained using a two-stage training procedure [4]:

1. Each 256x256 RGB image was compressed into a 32x32 grid of image tokens with a total of 8192 potential values using a discrete variational autoencoder (dVAE). This results in a factor of 192 reduction in the transformer's context size without noticeably degrading visual quality.

2. To simulate the joint distribution over the text and picture tokens, an autoregressive transformer is trained using up to 256 BPE-encoded text tokens and 32 x 32 = 1024 image tokens.

3. ETHICAL APPROACHES

As image synthesis is a relatively young and emerging topic, it can be advantageous to rest the ethical assessment upon applicable philosophical approaches as a known point of reference. For our evaluation, we have chosen two of the most well-known ethical concepts, namely, the categorial imperative as well as utilitarianism. These two are well suited for the corresponding analysis not only because of their prominence, but also their very different points of view.
Both will later on be used to perform the ethical analysis and illustrate considerations, potential consequences and the moral principles involved. By contrasting the categorical imperative with the utlilitarianism approach, it is possible to arrive at a more nuanced and informed understanding of the topic, while also highlighting any potential ethical dilemmas that may arise.

A. Kant

The german philosopher Immanuel Kant (1724–1804) introduced the categorical imperative in his book "Groundwork of the Metaphysic of Morals" in 1785.
Kant formulated the categorical imperative as followed: "Act only according to that maxim by which you can at the same time will that it should become a universal law". It is a moral principle which states that actions should be taken based on whether they can be willed as a universal law for all individuals in all situations. In other words, an action is only morally right if it can be applied to everyone without contradiction. Although it has been heaviliy critisized for its abstractness and inflexibility it is still very much relevant today and has a significant role to play in modern ethics and can be applied to a wide range of topics.

B. Utilitarianism

Utilitarianism is an ethical theory that delineates right from wrong by focusing on outcomes. It is a form of consequentialism. The theory assumes that the most ethical decision is the one that produces the greatest utility for the greatest number of people. It is based on ideas of Jeremy Bentham (1748–1832) and John Stuart Mill (1806–1873).
As this concept focuses on the consequences resulting from an action, all possible outcomes and their utilities, so both advantages and costs are taken into account to derive an ethical assessment. Moreover, it is also the method of moral reasoning that is most frequently applied in business [9] and is the moral system that is most known to defend using force or going to war.
However, utilitarian ethical decision-making also has its limits. Often times the exact consequences of an action are not certainly predictable.

C. An Ethical Example - The Trolley Problem

The well-known ethical dilemma called "The Trolley Problem" is well suited to emphasize these two concepts' differences in ethical assessment. The Trolley Problem describes a fictional scenario, a thought experiment, in which a spectator has the choice to rescue five people who are in danger of being hit by a trolley, by diverting the trolley to kill just one other person. A utilitarianist would pull the lever and save the five people, since the utility of saving five lives is higher than the one of only saving one. A Kantist on the other hand does not value consequences and would therefore not pull the lever, as one should not kill one person to save five other lives since killing another individual will also be immoral in Kantian ethics.

4. WHAT ETHICAL CHALLENGES EXIST?

Like with any emerging technology, there do not only come possibilities, but also challenges society has to face in its usage. The ability to synthesize images from a text input by the user gives great potential for misuse.
Companies like Open AI try to train their models to decline inappropriate requests, but the line between inappropriate and acceptable requests is blurred and difficult to pin down precisely. Besides that, there are also models like Stable Diffusion [10] available, which, as of now, cannot be regulated in a meaningful way. This is the case because it is open-source in nature and anybody can modify or distribute the source code, so that there does not exist a single point of liability. In this section a handful of problems will be outlined that are later on to be analysed with the help of the aforementioned moral principles.

A. Social Biases

Text-to-image generative models have been shown to have varying amounts of biases, especially gender and skin tone biases. When the model is prompted with seemingly neutral phrases (such as "a photo of a nurse") that include no indications to, for example, skin tone or gender, it is nevertheless likely to produce images that are biased towards a certain complexion. In the above example the generative model is heavily biased towards generating images of white women, as can be seen in Figure 4. More specifically, in this example, Stable Diffusion, a competitor to DALLE-2, overwhelmingly generates images of

white women, the often times associated stereotype to this profession. In particular, a study by scientists from UNC Chapel Hill where these biases were measured by the variance of the gender/skin tone distribution through both automated and human evaluation was able to determine a clear relationship between the training data - labeled images - and the skin tone/ gender biases[11]. Although these images were not created by DALLE-2 itself, the same phenomenon can be observed for DALLE-2 [12]. The example of Stable Diffusion is referred to here, since the amount of data and visualization options are more plentiful, whereas data for DALLE-2 is still scarce. TODO: reinforces STEREOTYPES!!!!

Fig. 4. Example of social bias: with the prompt "a photo of a nurse". Modified from [11].

B. Portrayal of racism, porn, illegal actions

Another downside of DALLE-2 which should not be left out are the negative consequences that center around one’s ability to generate any image at will. Therefore, it is inevitable, that malicious actors will eventually generate harmful, forbidden or illegal imagery. Although AI model companies try to restrict what can be generated from a prompt, as discussed in 6, there will always be workarounds to circumvent these safety-restrictions. Consequently, it is possible to generate images that for instance could portray a child smoking which to our western mindset represents an illegal action. Apart from that, other contents harmful to society could be of pornographic or racist nature and may even be considered a felony such as generating child pornography.

C. Deception, Political Propaganda and Extremism

One of the biggest dangers encompassing AI generated images are politically motivated and abusive contents like so-called deepfakes[1]. Especially in the current times of social media, where uploaded contents go viral[2] within minutes or hours and where does not exist a dedicated supervisory authority, the danger of FakeNews being spread across a country or even the whole globe is tremendously large and can bring about severe damage.
For example on March 23rd, 2023, Donald Trump, the former president of the United States, had images circulate of him were he allegedly resisted arrest and was tackled to the ground by the police. This image was uploaded on Twitter and has been viewed over five million times in a few days. Despite the fact that the original creator only ever intended for it to be a joke, the AI-generated picture has still drawn a lot of attention within this short timeframe[14]. In this case, the creator clarified that the image was not real and no harm was intended, however in other scenarios where indeed malicious intentions are present, differentiating a fake image from a real one may be problematic and may also lead to severe reputation damage.
Apart from that, another threat to the wellbeing of society is political propaganda and extremism.

Use in combination with chatgpt, ...

Fig. 5. Jan 6. U.S. Capitol insurrection video game art generated by MidJourneyAI, modified from [15].

5. ETHICAL EXAMINATION

Taking into account all of the aforementioned aspects, it is of significantly relevant to examine whether the use of image generating models in malicious edge cases is ethically justifiable, when considering the extensive consequences that they may entail. QUESTIONS These questions will be evaluated ethically, from the Kantian point of view as well as the Utilitarian perspective, in the following sections.

A. Kant

Kant’s categorical imperative implies that all people should behave in a certain way if they want everyone else to do the same. Applied to image-generation with AI that would mean everyone should only generate the images they want to see or think it is ethical to generate. However, since everyone has a different view of what

[1] TEST

[2] "Used to describe something that quickly becomes very popular or well known by being published on the internet [...]"[13].

is ethically justifiable to generate the categorical imperative is not a suitable ethical theory to evaluate the ethical boundaries of image generation.

B. Utilitarianism

According to Utilitarianism the good or pleasure originated of the AI-generated images has to be taken into account to decide if its an ethically justifiable action.
The pleasure derived from the generated picture has to be taken into account. A distinction needs to be made between generating an image and making it public. With the publication of a picture the circle of people who can be harmed or pleased by it becomes much larger and other orders of magnitude have to be weighed against each other.
If the image is not made public only the pleasure of the person that entered the query has to be taken.
There are different scenarios that can occur where from a utilitarian point of view it would be ethically justifiable to generate certain content or images, even if they harm a person or a group of people, this is a consequence of weighing the production of the greatest good.
To stay with the propaganda example: If a person were to create material that harmed a politician's reputation so that he would not be re-elected, but that politician had harmed more people than he had helped or pleased during his time in office, it would be morally acceptable from a utilitarian standpoint to create the damaging content.
Taken another example of child pornography, which is illegal to possess, produce and distribute in most countries. Generating it does not harm any child as a consequence of the action but in order to generate such content it is very likely that similar images are present in the training data. So children were harmed in advance to make this action possible.
Utilitarianism only takes the consequences of an action into account, thus the preliminary events would not be considered.
Here another weakness of utilitarianism comes into view: the consequences of an action cannot be predicted exactly. Does it keep pedophiles who can generate images of children's pornography from obtaining content from other sources, which would be not ethical justifiable from a utilitarian point of view?

6. WHAT IS CURRENTLY BEING DONE?

Open AI states in their Terms of Service (TOS) that DALLE 2's capacity to produce violent, hateful, or pornographic images is constrained. DALLE 2's exposure to these ideas is reduced by "removing the most explicit content from the training data". "Advanced methods are also employed to prevent the photorealistic generation of real people's faces, particularly those of public figures." And if filters detect text prompts and image uploads that might be against their policies, no images are generated. To prevent abuse, both automated and human monitoring mechanisms are employed [3].
-Simple Watermark
-strict regulation in China
-no regulation in the west (eu ai law proposal?)

A. Data Labeling

A TIME's investigation revealed that OpenAI used outsourced Kenyan laborers to label text snippets of violence, hate speech, and sexual abuse to generate training data for their ChatGPT's (Chat Generative Pre-trained Transformer) predecessor, GPT-3 (Generative Pre-trained Transformer 3) that uses deep learning techniques to generate human-like text and engage in conversational interactions and payed them less than 2 usd per hour [16]. The continuous exposure of people to this content is very likely to cause psychological consequences.

7. WHAT COULD BE DONE?

Limit Training Data

To hinder text-to-image models from generating inappropriate images the training data can be limited but therefore data that is inappropriate from a commonsense morality point of view must be labeled as such. But overall quality could degrade depending on how many and which training data are left out.
Other models can be trained to recognize immoral pictures and texts but human supervising is also needed. In order to prevent people from constantly being confronted with textual or pictorial images of violence, companies could join forces and make their already labeled data available to others and use already existing databases like the Socio-Moral Image Database [17].

Moral image manipulation

Another interesting approach does not affect the training of an AI but rather recognizes immoral parts of an already generated image and specifically alters them. Park et al. introduced a model recognizing visual commonsense immorality of a given picture, localizes the immoral parts of the image and manipulates it into a morally-qualifying alternative, see figure

8. CONCLUSION

REFERENCES

1. R. Metz, "Ai won an art contest, and artists are furious," (2022).
2. J. M. Allen, "Théâtre d'opéra spatial," (2022).
3. OpenAI, "Dall-e 2 openai," (2022).
4. A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," CoRR **abs/2102.12092** (2021).
5. A. Z. Jenny Benois-Pineau, ed., *Multi-faceted Deep Learning* (Springer Cham, 2021).
6. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," (2017).

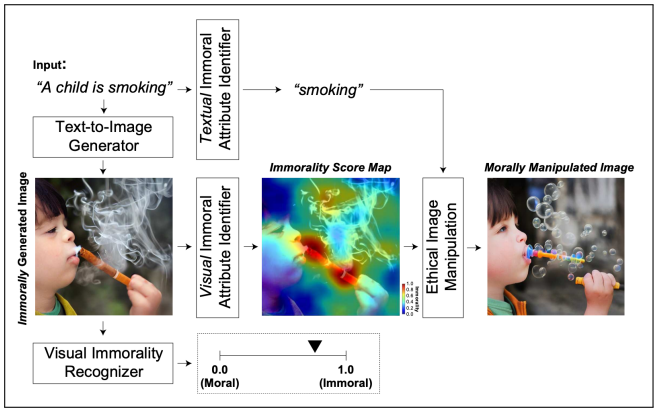

Fig. 6. An algorithm first analyzes an immorally generated image from text-to-image creation models, then pinpoints the visual and textual characteristics that contribute to the immorality of the image (e.g., smoking). Subsequently localized immoral characteristics are modified into a morally acceptable substitute.[18]

7. S. Frolov, T. Hinz, F. Raue, J. Hees, and A. Dengel, "Adversarial text-to-image synthesis: A review," CoRR **abs/2101.09983** (2021).
8. A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," (2022).
9. T. U. of Texas at Austin, "Ethics unwrapped," (2023).
10. S. Diffusion, "Stable diffusion," (2023).
11. J. Cho, A. Zala, and M. Bansal, "Dall-eval: Probing the reasoning skills and social biases of text-to-image generative models," (2022).
12. OpenAI, "Reducing bias and improving safety in dall·e 2," (2022).
13. *Cambridge Advanced Learners Dictionary* (Klett Sprachen GmbH, 2013).
14. N. N. Isaac Stanley-Becker, "Fake images of trump arrest show 'giant step' for ai's disruptive power," https://www.washingtonpost.com/politics/2023/03/22/trump-arrest-deepfakes/ (2023).
15. M. B. D. Daniel Siegel, "Weapons of mass disruption: Artificial intelligence and the production of extremist propaganda," (2023).
16. TIME, "Exclusive: Openai used kenyan workers on less than 2 usd per hour to make chatgpt less toxic," (2023).
17. D. L. Crone, S. Bode, C. Murawski, and S. M. Laham, "The socio-moral image database (smid): A novel stimulus set for the study of social, moral and affective processes," PLOS ONE **13**, 1–34 (2018).
18. S. Park, S. Moon, and J. Kim, "Judge, localize, and edit: Ensuring visual commonsense morality for text-to-image generation," (2022).