Is DALLE 2 worth its weight?

What if you had the ability to create any kind of image imaginable? What about if you had the ability to create images you *cannot* imagine? The image-generating model DALLE 2 created by OpenAI gives anyone with a computer and valid email address the abilities just described. By typing a phrase or sentence to prompt DALLE 2, in secnds you will have its interpretation of those words in image form. OpenAI's model has extremely powerful capabilities of image generation and modification, but just as many prospective risks in the hands of the ill willed or negligent. DALLE 2 has unlimited creative potential, however, in a time where harmful content can spread exponentially, OpenAI acted unethically by neglecting to manage the negative potential societal effects of the public release of their model.

DALLE 2 seems a lot like magic, but in reality it is a carefully crafted AI system. It is a text-to-image model which uses natural language processing, deep neural networks, and encoding and decoding to achieve its results (DiBattista, 2022). The model has been trained using a dataset consisting of 400 million different image and text pairings (DiBattista, 2022). It uses this dataset to understand the patterns between the text and its corresponding image. The model then generates three images that make the most sense in conjunction with the text input it receives. As a result of the sheer amount of training data, DALLE 2 is extremely good at what it does. This allows OpenAI to support advanced features ranging from image generation from scratch to image modification, where users can take out or insert objects in an image or expand the dimensions of an image by generating more content in all directions.

Due to DALLE 2's unprecedented capabilities, it was impossible to predict how the model would be received, how it would be used, and what might go wrong. Because of this, OpenAI

decided to release the model in waves. First, they selected 200 researchers to help in DALLE 2's testing stages (Edwards, 2022). After they were satisfied with the testing, they allowed the public to apply online to gain access to DALLE 2 in order to be able to control who and how many people would be using the model. Finally some two months later, the public was given full reign to experiment with the model while complying with OpenAI's detailed use policies. According to OpenAI's content policy specifically, images generated by DALLE 2 belonged to those who prompted the model, as long as they did not violate the terms of use in the process (OpenAI, 2022c). DALLE 2 quickly became mainstream, with people all over the world utilizing this new tool to generate and own unique images to use in a variety of ways.

Amidst all of its exploration came the inevitable urge to push DALLE 2 to its limits and, unfortunately, there were lots of limits to discover. Even though DALLE 2 was advertised to inspire creativity in things like art, marketing, and branding, a segment of the public saw the tremendous opportunities to generate harmful content and misinformation. Anticipating this, OpenAI applied filters on text inputs in accordance with their policy to prohibit the creation or sharing of "images that are not G-rated or that could cause harm" including content promoting hate, harassment, violence, self-harm, deception, public/personal health, spam, and sexual, shocking, political, or illegal content (OpenAI, 2022b). Despite OpenAI's efforts, abuse of the model prevailed. In an extensive report on the risks and limitations of their model, OpenAI described some of the ways in which users had circumvented their restrictions, explaining that users can create "prompts for things that are visually similar to objects or concepts that are filtered, e.g. ketchup for blood" that can still appear realistic (Mishkin et al., 2022). Additionally, researchers identified that DALLE 2's image generation could contribute to the already enormous amount of misinformation circulating on the internet and social media. Within the

same report on risks and limitations, OpenAI details the potential their model has to misrepresent public figures, aid in online harassment, and spread fake news (Mishkin et al., 2022). They wrote that their inpainting feature allows users to realistically modify images and can easily be abused (Michkin, et al., 2022). For example, a false event can be staged by leveraging this feature to add smoke coming out of a building in user's image (Mishkin et al., 2022).

In addition to the creation of harmful content and misinformation, DALLE 2 has been found to generate images that perpetuate harmful stereotypes and biases. The data used to train DALLE 2 consists of "millions of images scraped off the internet and their corresponding captions" (Samuel, 2022). Because of the fact that the data has come from the internet, the model has learned to generate images based on the biases already present within our society - for example, the model tends to favor "presenting women in more sexualised contexts" (Naik, 2022), presenting white people over people of color, and presenting most concepts through a western lens (Mishkin et al., 2022). Further, underrepresented people and concepts hold more likelihood of "having their prompts or generations filtered, flagged, blocked, or not generated in the first place" (Mishkin et al., 2022). Overarchingly, DALLE 2 has the alarming potential to contribute more bias to society through both perpetuating harmful stereotypes and excluding or misrepresenting particular groups of people.

There is a lot to consider regarding the ethical implications of OpenAI's decision to release DALLE 2 to the general public. In terms of consequentialism, OpenAI had an ethical duty to evaluate the costs and benefits of allowing their model to be used by the public. This involves evaluating the real purposes of DALLE 2 and their goals in releasing it. According to their website, OpenAI's hope is that "DALLE 2 will empower people to express themselves creatively" (OpenAI, n.d. b). Moreover, the model can serve as inspiration for innovations that

we could not have conceptualized before. One image generation shared online depicted a car unlike anything you might imagine based on this text input: “The weirdest concept car ever designed by man” (Willings, 2022). OpenAI’s model does a great job at producing new and intriguing ideas - so much so that the business industry is looking to use DALLE 2 to their advantage as well. Not only do generated images have great potential for use in marketing, researchers expect that videos will follow, adding “Why spend $100,000 on a single television commercial targeting millions of people when, with the same budget, an advertiser will be able to create 10,000 different commercials, each tuned to a cluster of like-minded viewers at a moment in time?” (Caruso, 2022). With this in mind, DALLE 2 could propel technological advancements and enhance user experiences with advertisements among many other things.

On the other hand, there are many risks and limitations associated with DALLE 2 that cannot be overlooked. It is clear that OpenAI put in effort in analyzing the possible negative consequences of releasing DALLE 2 to the public. They demonstrate this through their detailed ownership policies and their filters to regulate misuse via text inputs. However, what OpenAI might not have considered is how they would enforce this policy, seeing as users of DALLE 2 have generated more than two million images daily since its release (Edwards, 2022). This is far too much content to be consistently monitored by humans, and there are clear loopholes in automated filtering systems. Further, OpenAI has written extensively about the biases present in DALLE 2 without making a strong effort to correct their model. They have attempted to make the model more inclusive with a new technique that “is applied at the system level when DALL·E is given a prompt describing a person that does not specify race or gender” in order to generate a more diverse set of images (OpenAI, 2022a). Despite this, simple prompts like the single word “firefighter” or “wedding” still produce images of white men in firefighter gear and

depictions of heterosexual couples respectively. The initial effort OpenAI made to mitigate harm was ethical, as they took into account the consequences of fully releasing the model to the public by slowly rolling out the model and implementing safe guards. However, as time has gone on, OpenAI's activity has become less and less ethical in terms of consequentialism. They have ineffectively regulated abuse and have not prioritized reducing bias in DALLE 2. This will continue to negatively impact our society both on and off the internet through the spreading of misinformation, online hate, violence, and harassment, and reinforcing biases against minority groups.

Conversely, OpenAI's decision to release DALLE 2 to the general public can be evaluated through the lens of business ethics, specifically stakeholder ethics. In terms of stakeholder ethics, OpenAI has an ethical responsibility to consider each person who would be affected by their actions and products. As a public product, anyone in the world has access to DALLE 2. Accordingly, OpenAI states that their mission "is to ensure that artificial general intelligence benefits all of humanity" (OpenAI, n.d. a). Further, one of their goals is to "avoid enabling uses of AI or [artificial general intelligence] that harm humanity or unduly concentrate power" (OpenAI, 2018). In thinking about whether DALLE 2 serves to benefit all of humanity, and in turn, OpenAI's stakeholders, we must ask the question, does DALLE 2 improve the lives of the people around the world and avoid concentrating power for majority societal groups? While the model has a ton of creative and business potential, the reality is that it does not serve many minority groups. OpenAI has researched this themselves, explaining that DALLE 2's bias towards whiteness, western traditions, heterosexuality, etc, can cause "downstream effects on what is seen as available and appropriate in public discourse" (Mishkin et al., 2022). It follows that OpenAI's model has the potential to make life even harder for minorities as they try to

overcome negative stereotypes and biases in their everyday lives. Considering DALLE 2's scope, it is extremely reckless to release a model with so many shortcomings. It is especially unethical to insinuate that they are working to benefit all of humanity and subsequently neglect to support large portions of their stakeholders.

OpenAI's public release of DALLE 2 had a lot more weight than what was immediately obvious. The model has and will continue to increase misinformation and the spread of harmful content, and perpetuate harmful biases. These consequences have come as a result of OpenAI's failure to implement strong reinforcements to back their policies, and their failure to work to uphold their mission to benefit all of humanity. While DALLE 2 is an incredible display of the current AI capabilities and research potential, OpenAI's inability to minimize the model's global societal impact renders the public release of their model unethical.

References

Caruso, E. (2022, August 16). *OpenAI's DALL·E 2 Model Has Profound Implications*. VettaFi. Retrieved November 17, 2022, from https://etfdb.com/disruptive-technology-channel/openai-s-dall-e-2-model-has-profound-implications/

DiBattista, J. (2022, August 9). *What is DALLE 2? What to Know Before Trying the Groundbreaking AI*. Medium. Retrieved November 13, 2022, from https://medium.com/geekculture/what-is-dalle-2-what-to-know-before-trying-the-groundbreaking-ai-e7a585f2edf0

Edwards, B. (2022, September 28). *DALL-E image generator is now open to everyone*. Ars Technica. Retrieved November 17, 2022, from https://arstechnica.com/information-technology/2022/09/openai-image-generator-dall-e-now-available-without-waitlist/

Mishkin, P., Ahmad, L., Brundage, M., Krueger, G., & Sastry, G. (2022, April 11). *DALL·E 2 Preview - Risks and Limitations*. GitHub. Retrieved November 14, 2022, from https://github.com/openai/dalle-2-preview/blob/main/system-card.md?utm_source=Sailthru&utm_medium=email&utm_campaign=Future%20Perfect%204-12-22&utm_term=Future%20Perfect#bias-and-representation

Naik, A. R. (2022, July 26). *The Societal Dangers of DALL-E 2*. Analytics India Magazine. Retrieved November 15, 2022, from https://analyticsindiamag.com/the-societal-dangers-of-dall-e-2/

OpenAI. (n.d. a). *About OpenAI*. OpenAI. Retrieved November 18, 2022, from https://openai.com/about/

OpenAI. (n.d. b). *DALL·E 2 is a new AI system that can create realistic images and art from a description in natural language.* OpenAI. Retrieved November 19, 2022, from https://openai.com/dall-e-2/

OpenAI. (2018, April 9). *OpenAI Charter*. OpenAI. Retrieved November 19, 2022, from https://openai.com/charter/

OpenAI. (2022a, July 18). *Reducing Bias and Improving Safety in DALL·E 2*. OpenAI. Retrieved November 18, 2022, from https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2/

OpenAI. (2022b, September 19). *Content policy*. DALL·E. Retrieved November 14, 2022, from https://labs.openai.com/policies/content-policy

OpenAI. (2022c, November 3). *Terms of Use*. OpenAI. Retrieved November 17, 2022, from https://openai.com/api/policies/terms/

Samuel, S. (2022, April 14). *OpenAI's DALL-E 2 is a new illustration of AI bias*. Vox. Retrieved November 15, 2022, from https://www.vox.com/future-perfect/23023538/ai-dalle-2-openai-bias-gpt-3-incentives

Willings, A. (2022, November 9). *We crafted these weird and wonderful images using OpenAI's Dalle-2*. Pocket-lint. Retrieved November 17, 2022, from https://www.pocket-lint.com/apps/news/163320-we-crafted-these-weird-and-wonderful-images-using-openai-s-dalle-2