DALLE-2: ”Picture Perfect or Ethically Problematic?”

Ethical implications of AI-generated images

Manuel Muehlberger [1] Sofia Kaltwasser [2]

[1]TU Munich

[2]University of Potsdam

Ethik 4 Nerds

08.03.2023

Technische Universität München

Universität Potsdam

Presentation Overview

AI-generated images

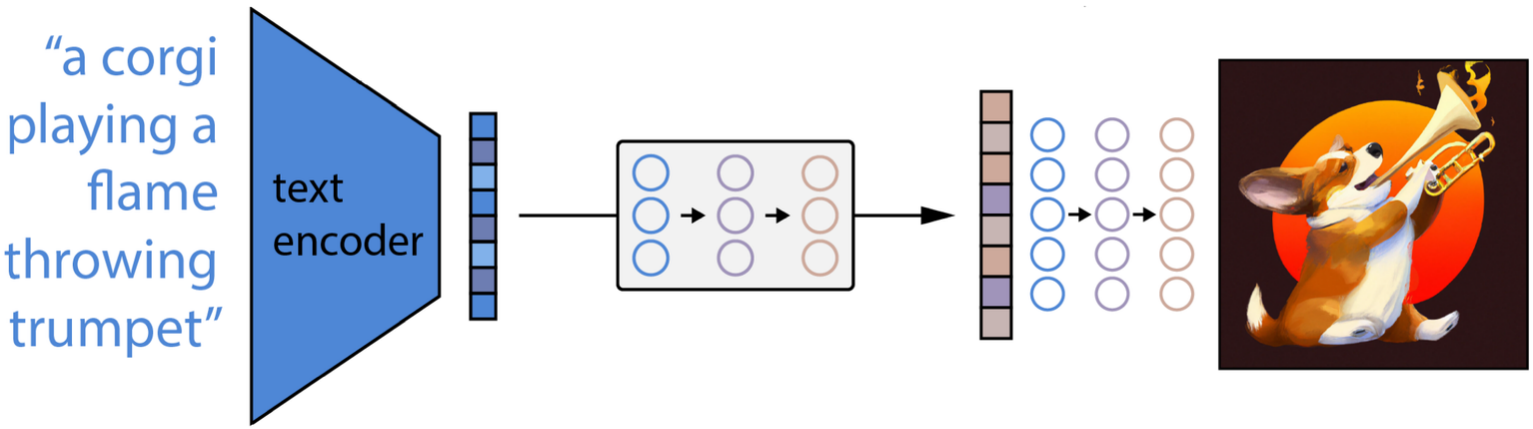
- Significant advancements in recent years
- Realistic and high-quality images $\rightarrow$ almost indistinguishable from human-created images
- Example Dall-E 2 by OpenAI [1]

Dall-E 2 generated images examples

(a) A tapir with the texture of an accordion [2]

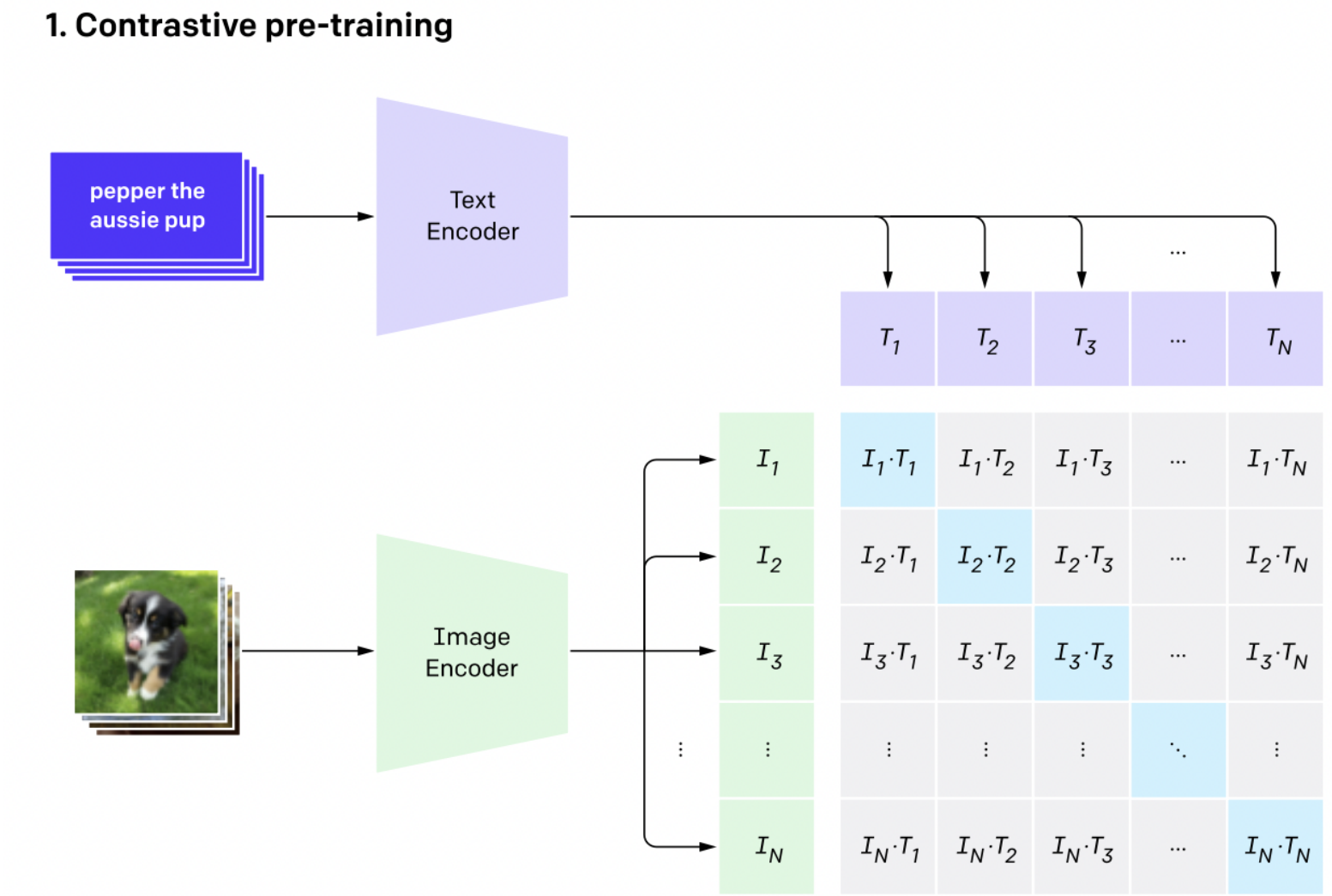
(b) Illustration of a baby hedgehog in a christmas sweater walking a dog [2]

High level abstraction of image generation process of DALL-E 2



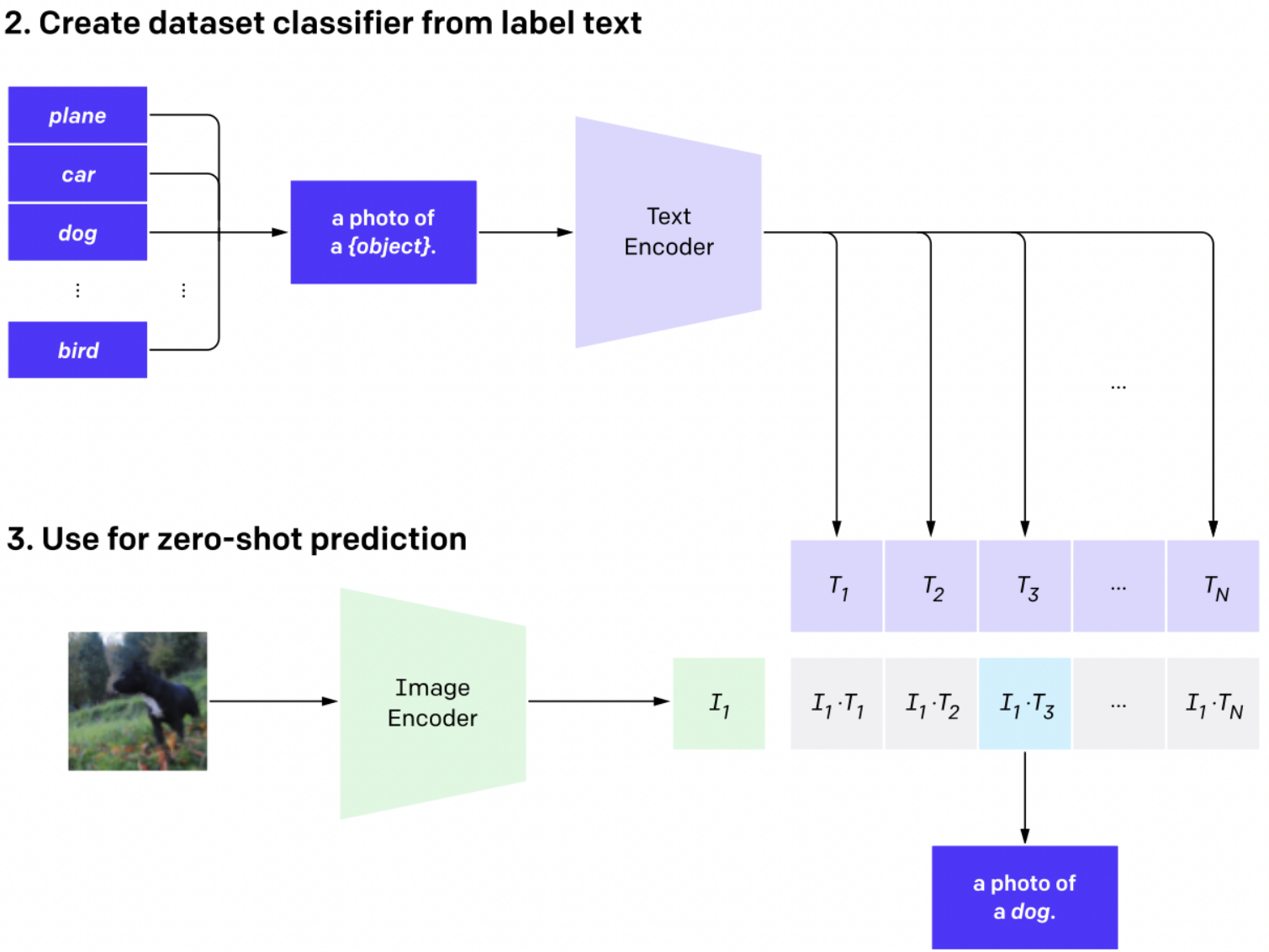
High-level overview of operating principle of DallE-2 [modified from [3]]

Contrastive Language-Image Pre-training (CLIP)



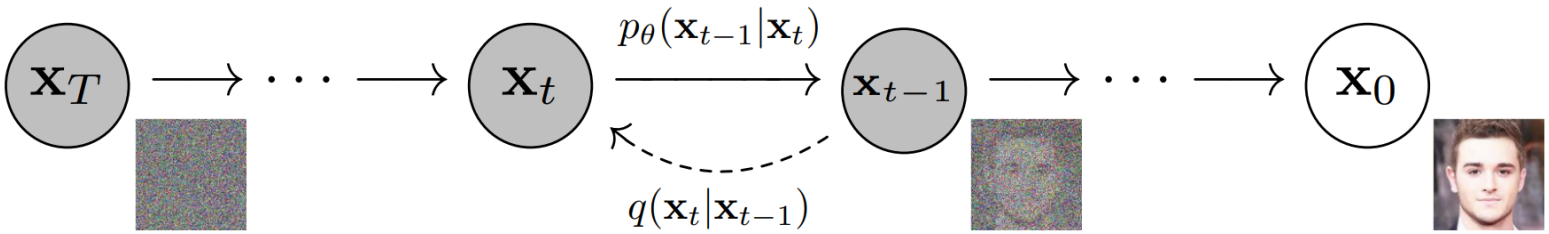
[4]

Contrastive Language-Image Pre-training (CLIP)



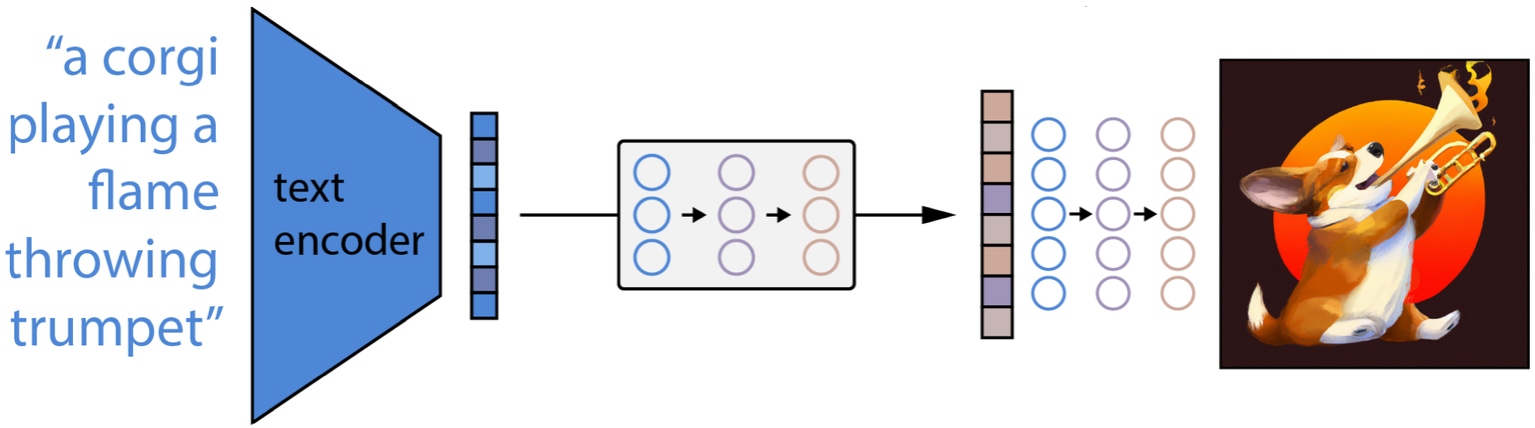
[4]

Diffusion model



Directed graphical model [5]

Putting it all together



[3]

Picture-Perfect or Ethically Problematic?

Let’s consider some ethical issues

First of all: Key Features

- Efficiency and Scalability
- Cost-Effectiveness
- No specialized skills required to use

Artists taking credit for AI-generated content

“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[6].

Portrayal of racism, misogyny, violence, illegal actions

- Examples: underage smoking, violence, porn
- "*Unstable Diffusion*": Community targeted specifically at porn generation[7]
- DALLE-2 is proven to have social biases [8]

A child smoking, generated by Stable Diffusion. Modified from [9]

Deception, political propaganda and extremism

- Discredit opposing political parties
- Propaganda en masse?

MidJourneyAI generated punk-skinhead music album cover art, modified from [10]

Jan 6. U.S. Capitol insurrection video game art generated by MidJourneyAI, modified from [10]

A LOT of potential for misuse!

- Highly realistic images of individuals without their consent
- Portrayal of socially unacceptable actions:
  - Violence
  - Pornography
  - Racism
  - Extremism
  - ...

What is currently being done?

- Every company has its own Terms of Services
  - DALLE-2: (theoretically) very restrictive
- Enforcement often just by blocking certain keywords
  - DALLE-2: No generation of faces (up to Sep. 2022)
- DALLE-2: Simple Watermark

Visualization: some guidelines are strict, some less so I

Do not attempt to create, upload, or share images that are not G-rated or that could cause harm.

- **Hate:** hateful symbols, negative stereotypes, comparing certain groups to animals/objects, or otherwise expressing or promoting hate based on identity.
- **Harassment:** mocking, threatening, or bullying an individual.
- **Violence:** violent acts and the suffering or humiliation of others.
- **Self-harm:** suicide, cutting, eating disorders, and other attempts at harming oneself.
- **Sexual:** nudity, sexual acts, sexual services, or content otherwise meant to arouse sexual excitement.
- **Shocking:** bodily fluids, obscene gestures, or other profane subjects that may shock or disgust.
- **Illegal activity:** drug use, theft, vandalism, and other illegal activities.
- **Deception:** major conspiracies or events related to major ongoing geopolitical events.
- **Political:** politicians, ballot-boxes, protests, or other content that may be used to influence the political process or to campaign.
- **Public and personal health:** the treatment, prevention, diagnosis, or transmission of diseases, or people experiencing health ailments.
- **Spam:** unsolicited bulk content.

Don't mislead your audience about AI involvement.

- When sharing your work, we encourage you to proactively disclose AI involvement in your work.
- You may remove the DALL·E signature if you wish, but you may not mislead others about the nature of the work. For example, you may not tell people that the work was entirely human generated or that the work is an unaltered photograph of a real event.

Respect the rights of others.

- Do not upload images of people without their consent.
- Do not upload images to which you do not hold appropriate usage rights.
- Do not create images of public figures.

Excerpt from the Community Guidelines of OpenAI's Dall-E-2[11].

Visualization: some guidelines are strict, some less so II

1. Be kind and respect each other and staff. Do not create images or use text prompts that are inherently disrespectful, aggressive, or otherwise abusive. Violence or harassment of any kind will not be tolerated.
2. No adult content or gore. Please avoid making visually shocking or disturbing content. We will block some text inputs automatically.

Excerpt from the Community Guidelines of Midjourney Inc's MidjourneyAI[12].

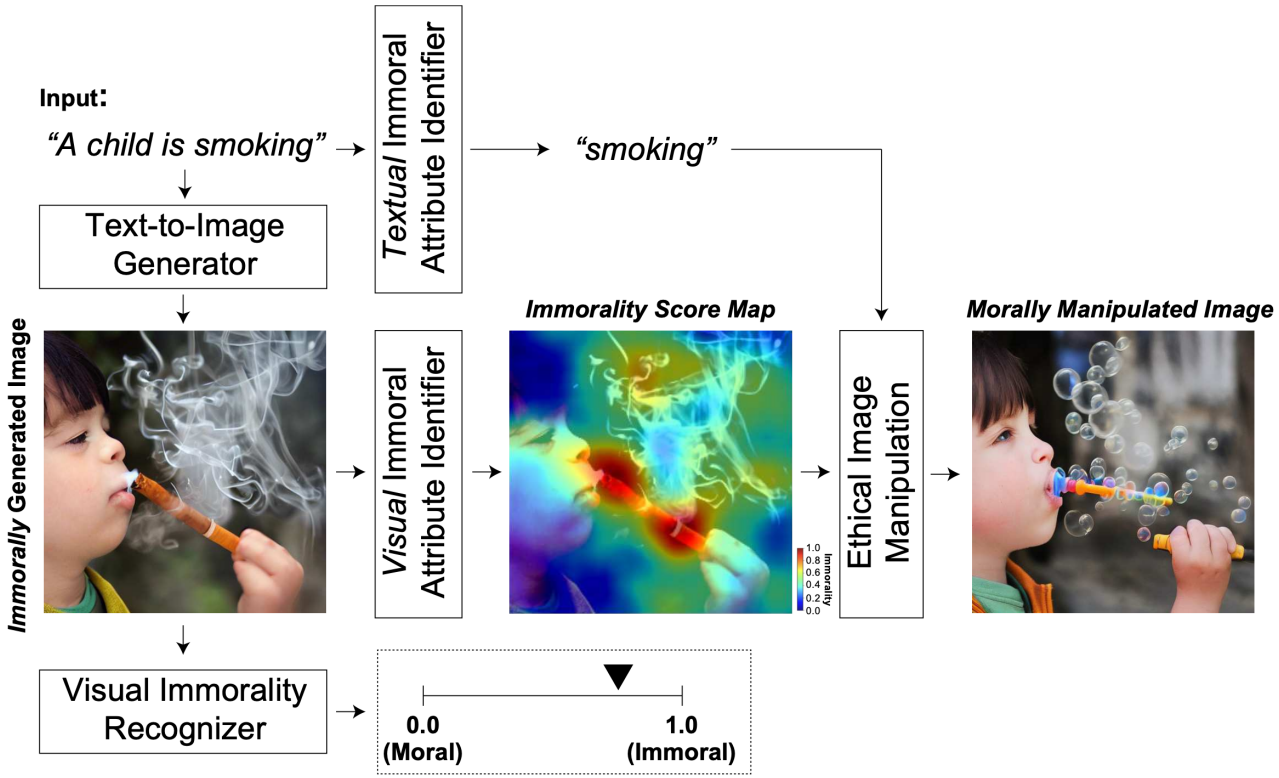
What is currently being done?

- No official regulations or laws in the West (yet)
- Role model China?[13]
  - Companies need to keep records of their use of deep synthesis technology
  - Companies are required to clearly mark generated content e.g. watermarks
  - Users need government I.D. to sign up

What could be done?

What could be done? I

- Watermarks / Clear labeling of AI-generated images
- Content filters



Manipulating an immoral image by localizing immoral visual attributes and manipulating them into a morally-satisfying alternative[9].

What could be done? II

- Restrict who can access models (”driver’s license”)
- Impose strict limit on training data $\rightarrow$ ”mischief models”
- Establish legal liability for misuse

References I

OpenAI.
Dall-e 2 openai, 2022.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever.
Zero-shot text-to-image generation.
CoRR, abs/2102.12092, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen.
Hierarchical text-conditional image generation with clip latents, 2022.

OpenAI.
Clip openai, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel.
Denoising diffusion probabilistic models.
CoRR, abs/2006.11239, 2020.

Jason M. Allen.
Théâtre d'opéra spatial, 2022.

Unstable Diffusion.
Wikipage of a community focussed on generating nsfw images with stable diffusion, 2023.

Jaemin Cho, Abhay Zala, and Mohit Bansal.
Dall-eval: Probing the reasoning skills and social biases of text-to-image generative models, 2022.

References II

Seongbeom Park, Suhong Moon, and Jinkyu Kim.
Judge, localize, and edit: Ensuring visual commonsense morality for text-to-image generation, 2022.

Mary Benett Doty Daniel Siegel.
Weapons of mass disruption: Artificial intelligence and the production of extremist propaganda, 2023.

OpenAI.
Dall-e 2 community guidelines, 2022.

MidJourney Inc.
Midjourneyai community guidelines, 2023.

Cyberspace Administration of China.
Regulations on deep synthesis internet information services, 2023.