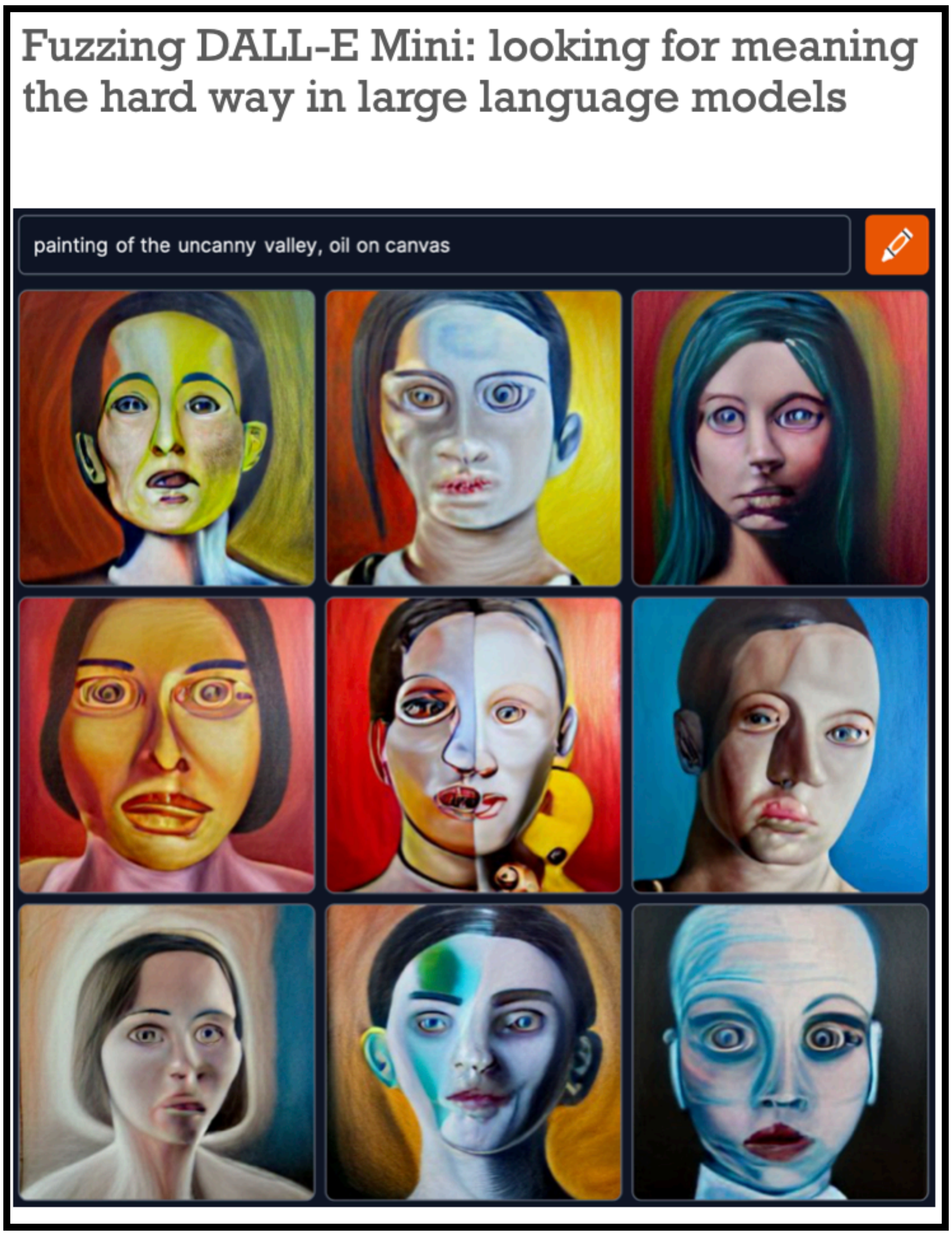
Grappling with widespread machine image generation

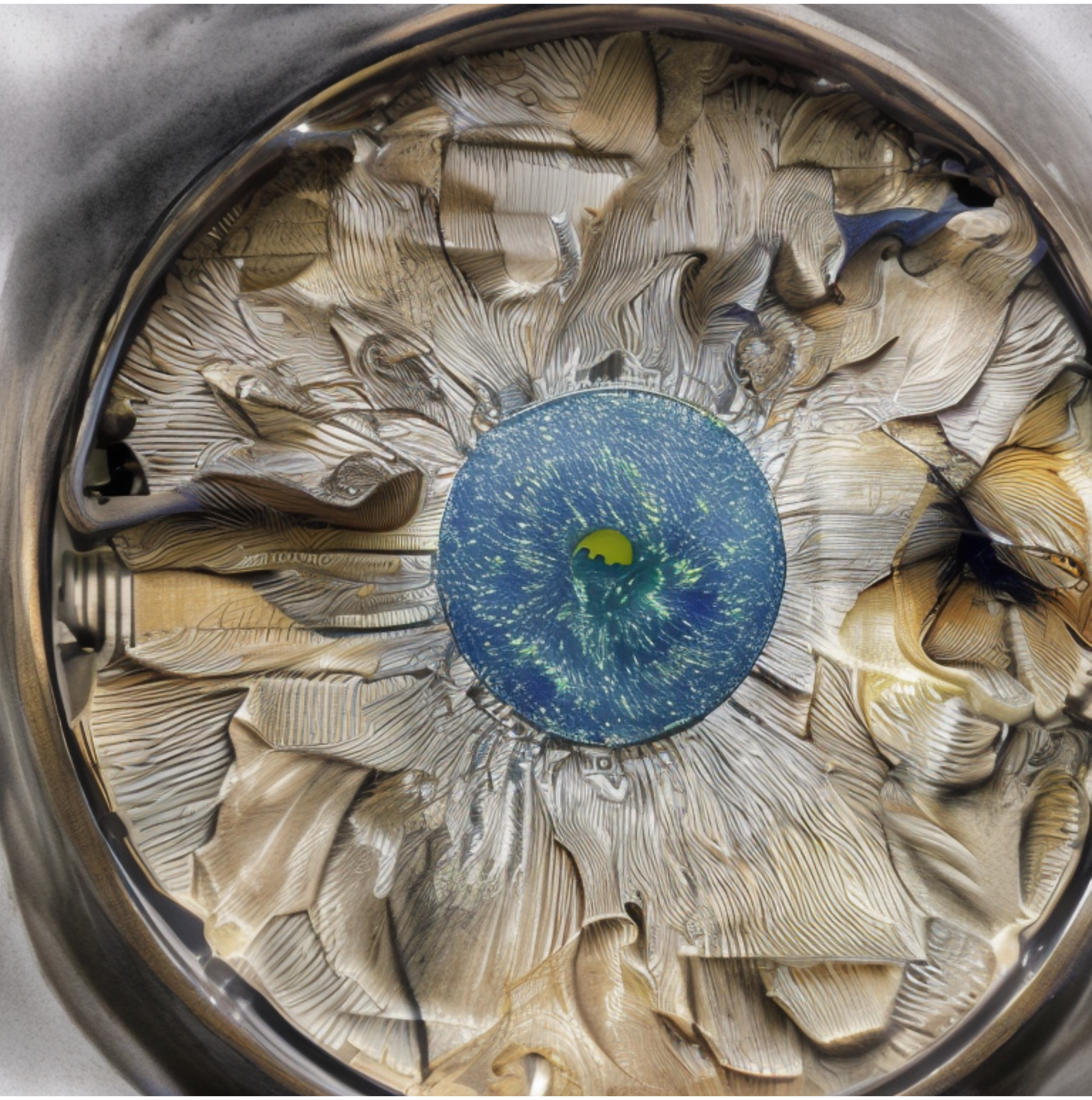
Adam Hyland — **CSC Lab UW HCDE** — adampunk.com

Some of what you see here is derived from this project:



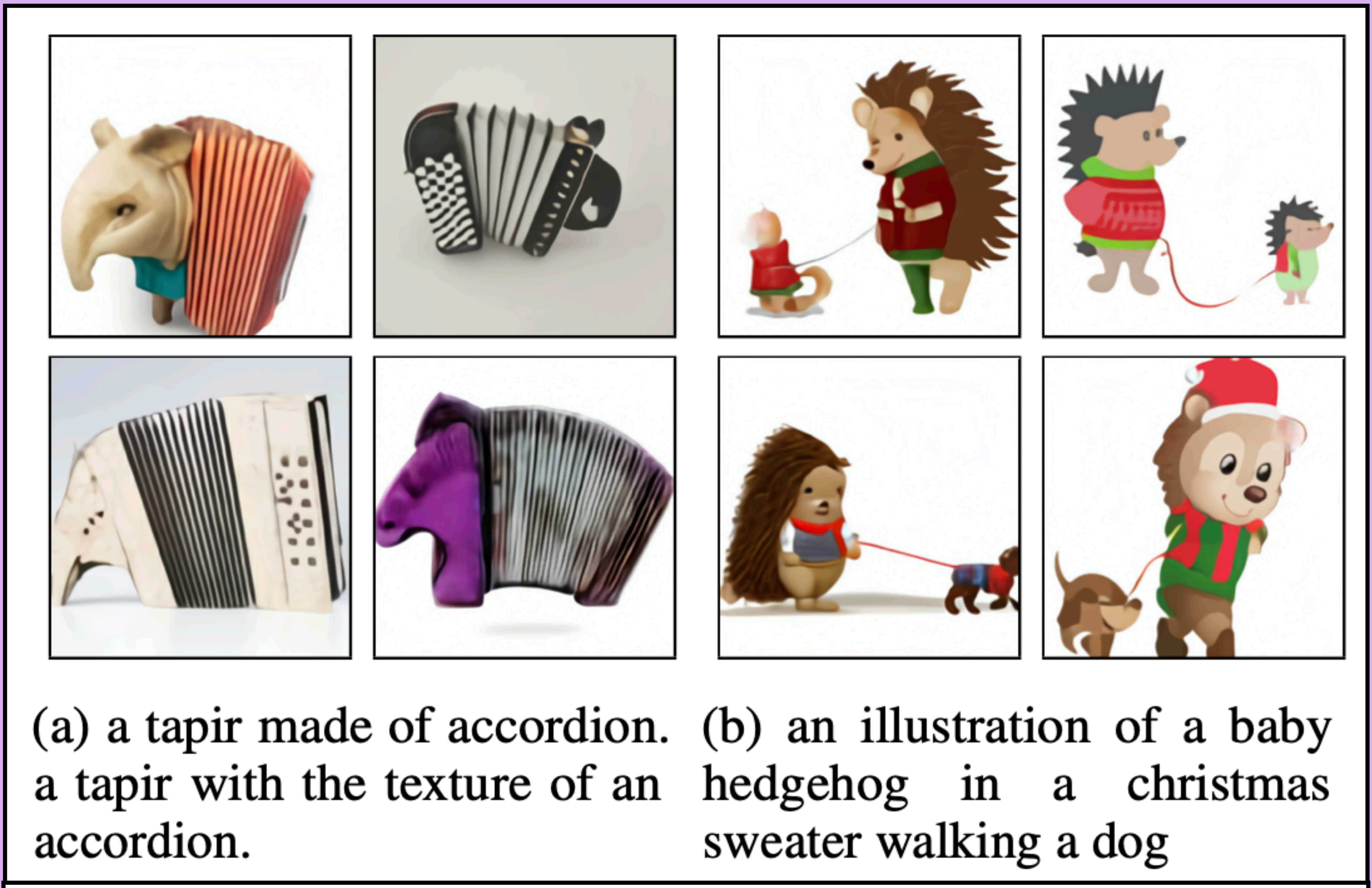
Fuzzing DALL-E Mini: looking for meaning the hard way in large language models

painting of the uncanny valley, oil on canvas



DALL-E (2021)

DALL-E, revealed in 2021, was the work of a OpenAI, a small AI company supported by Microsoft and Infosys.



(a) a tapir made of accordion. a tapir with the texture of an accordion.

(b) an illustration of a baby hedgehog in a christmas sweater walking a dog

From *Figure 2* of Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M, Sutskever I. Zero-shot text-to-image generation. InInternational Conference on Machine Learning 2021 Jul 1 (p. 8822). PMLR.

DALL-E was not widely released—OpenAI considered the potential risks too hard to model✦ at the time.

✦: Coldewey D. OpenAI's DALL-E creates plausible images of literally anything you ask it to. TechCrunch. January 5, 2021

Google’s Imagen—a tool similar to DALL-E—is restricted to internal use[†] entirely. The engineers judged libraries like these require so much image data that uncurated input data results in offensive, biased or misleading output[‡].

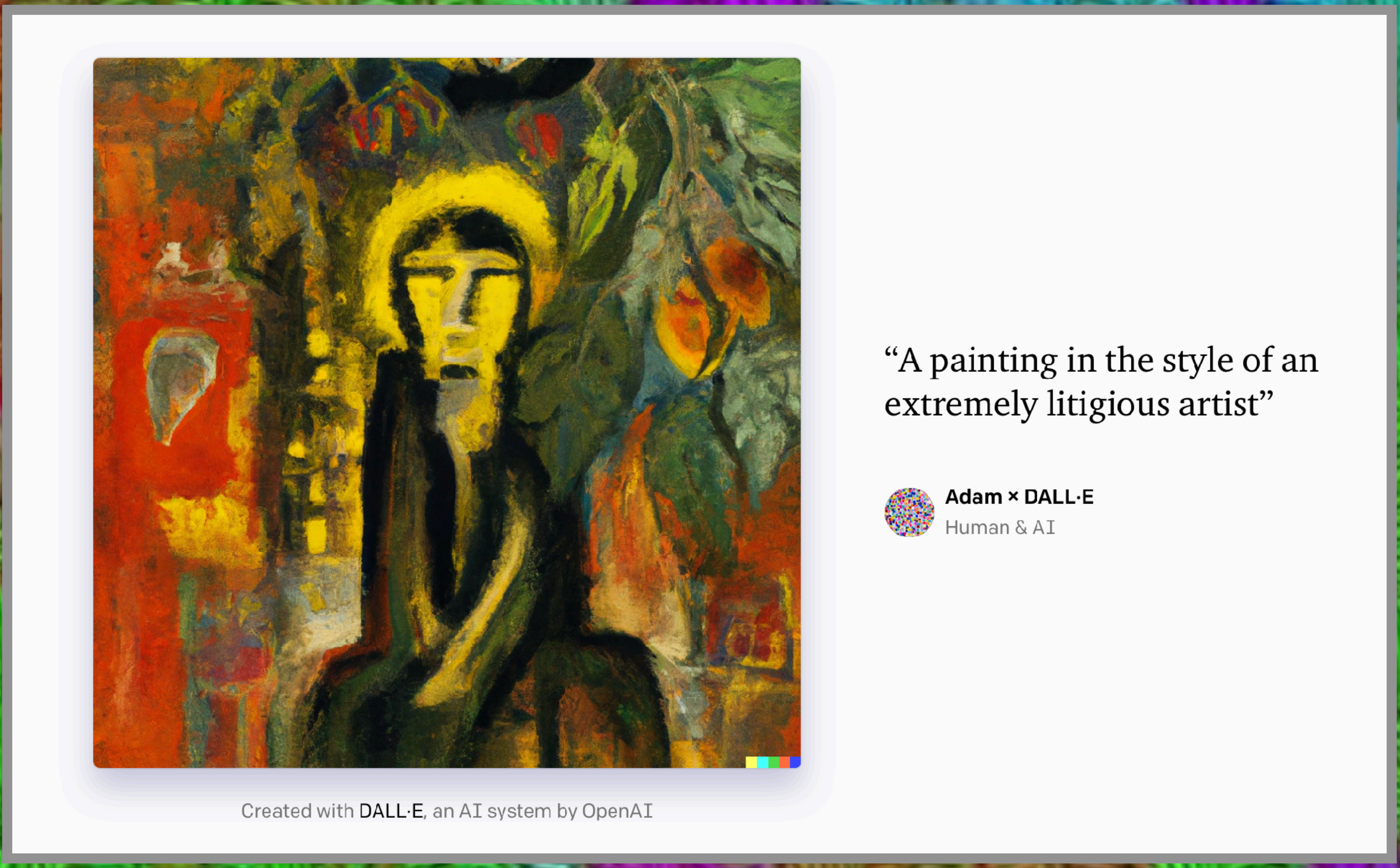
Motives for restricting access may not be entirely altruistic. Creating large language models from scratch is expensive and so gating a model while exclusively licensing it to a business partner[†] is one way to recoup costs.

[†] Vincent J. All these images were generated by Google’s latest text-to-image AI. The Verge. May 24, 2022

[‡] Saharia C, Chan W, Saxena S, Li L, Whang J, Denton E, Ghasemipour SK, Ayan BK, Mahdavi SS, Lopes RG, Salimans T. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv preprint arXiv:2205.11487. 2022 May 23.

[†] Dickson B. The implications of Microsoft’s exclusive GPT-3 license. TechTalks. September 24, 2020

In 2022 OpenAI released DALL-E 2, a much more powerful tool, **behind an API** nominally so use of the model can be tracked and restricted in the aim of preventing the tool from being used to generate harmful images.



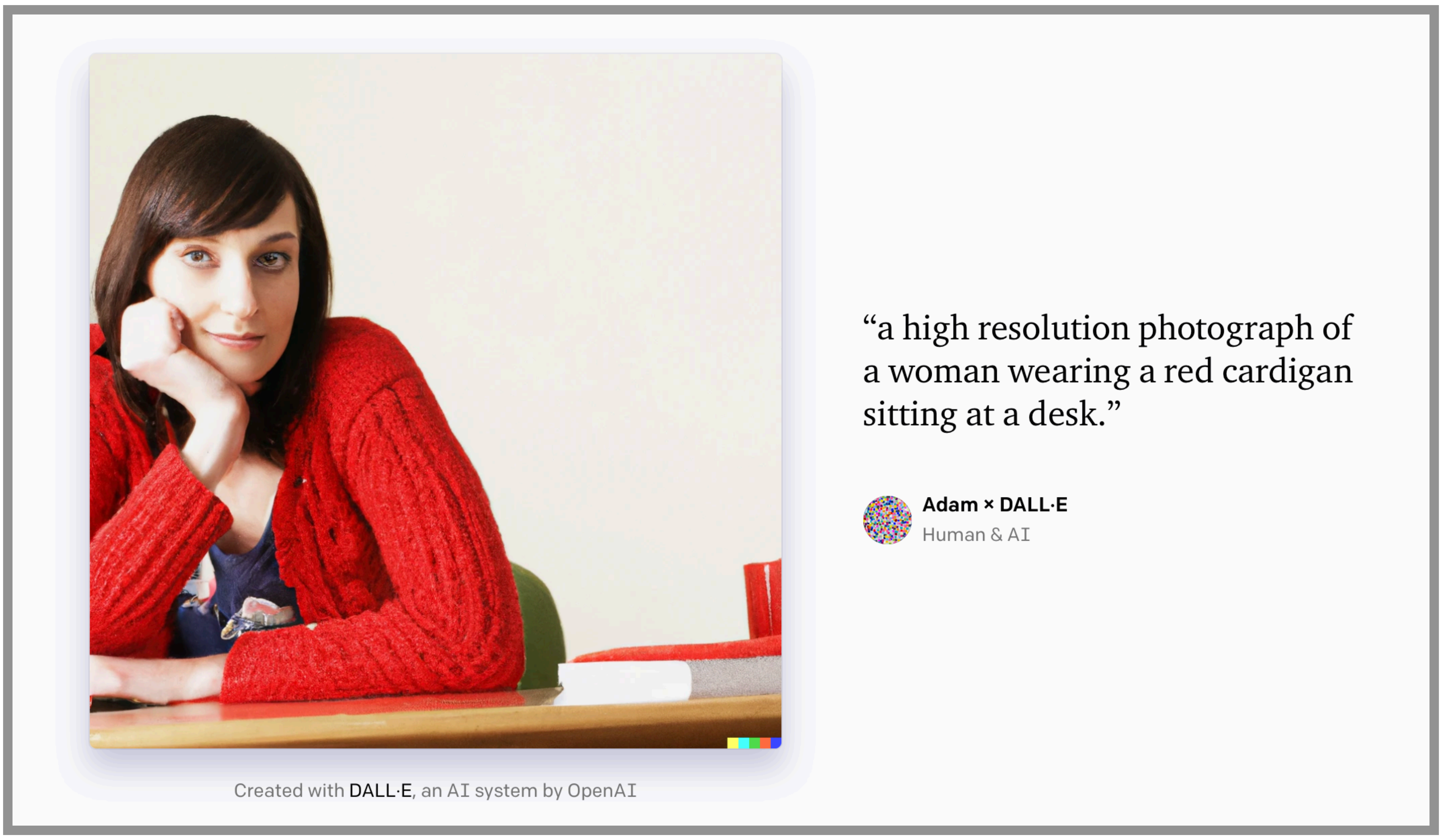
“A painting in the style of an extremely litigious artist”

Adam × DALL·E
Human & AI

Created with DALL·E, an AI system by OpenAI

DALL-E 2

DALL-E 2 can generate photorealistic images

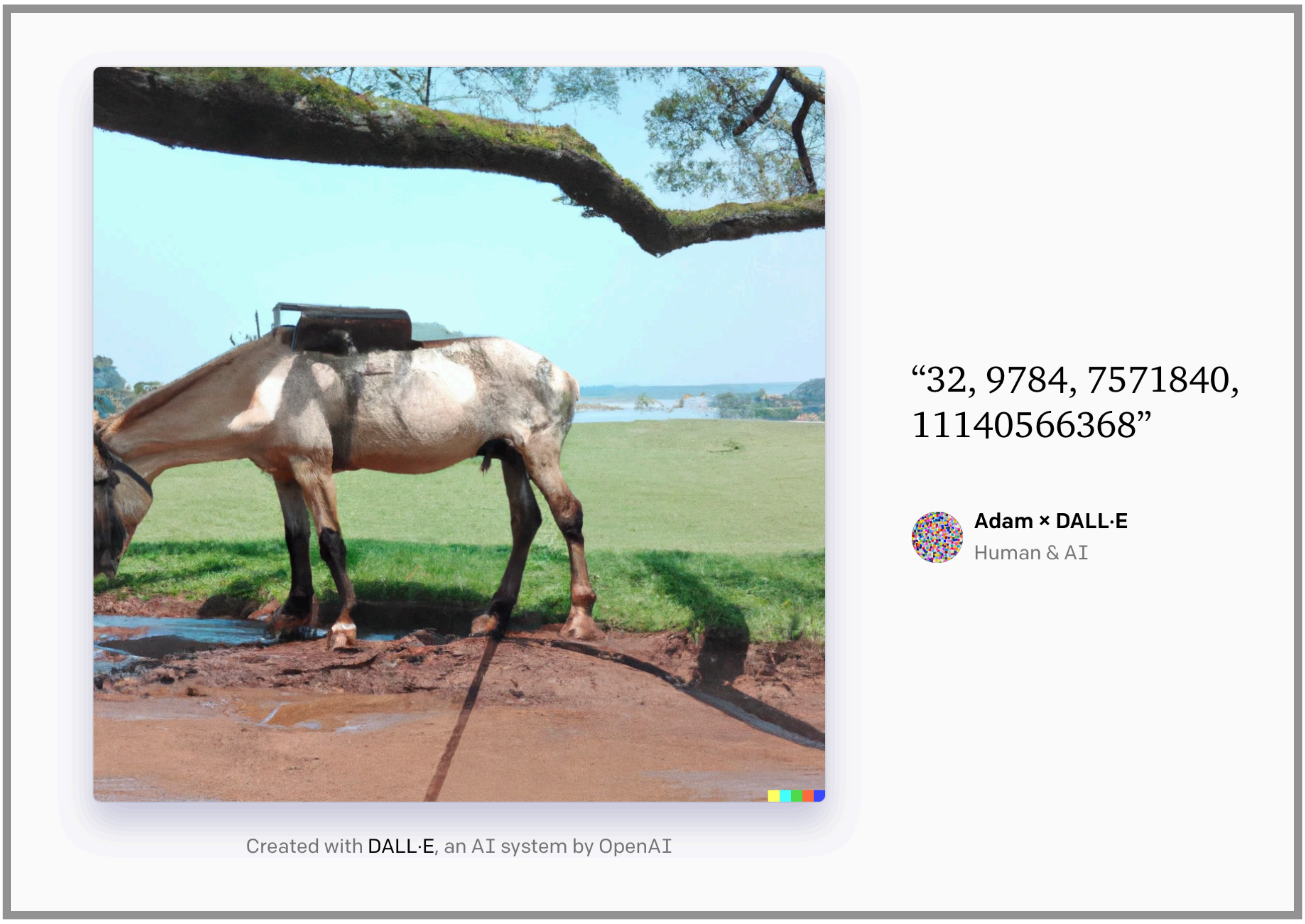


“a high resolution photograph of a woman wearing a red cardigan sitting at a desk.”

Adam × DALL·E
Human & AI

Created with DALL·E, an AI system by OpenAI

…more or less intentionally

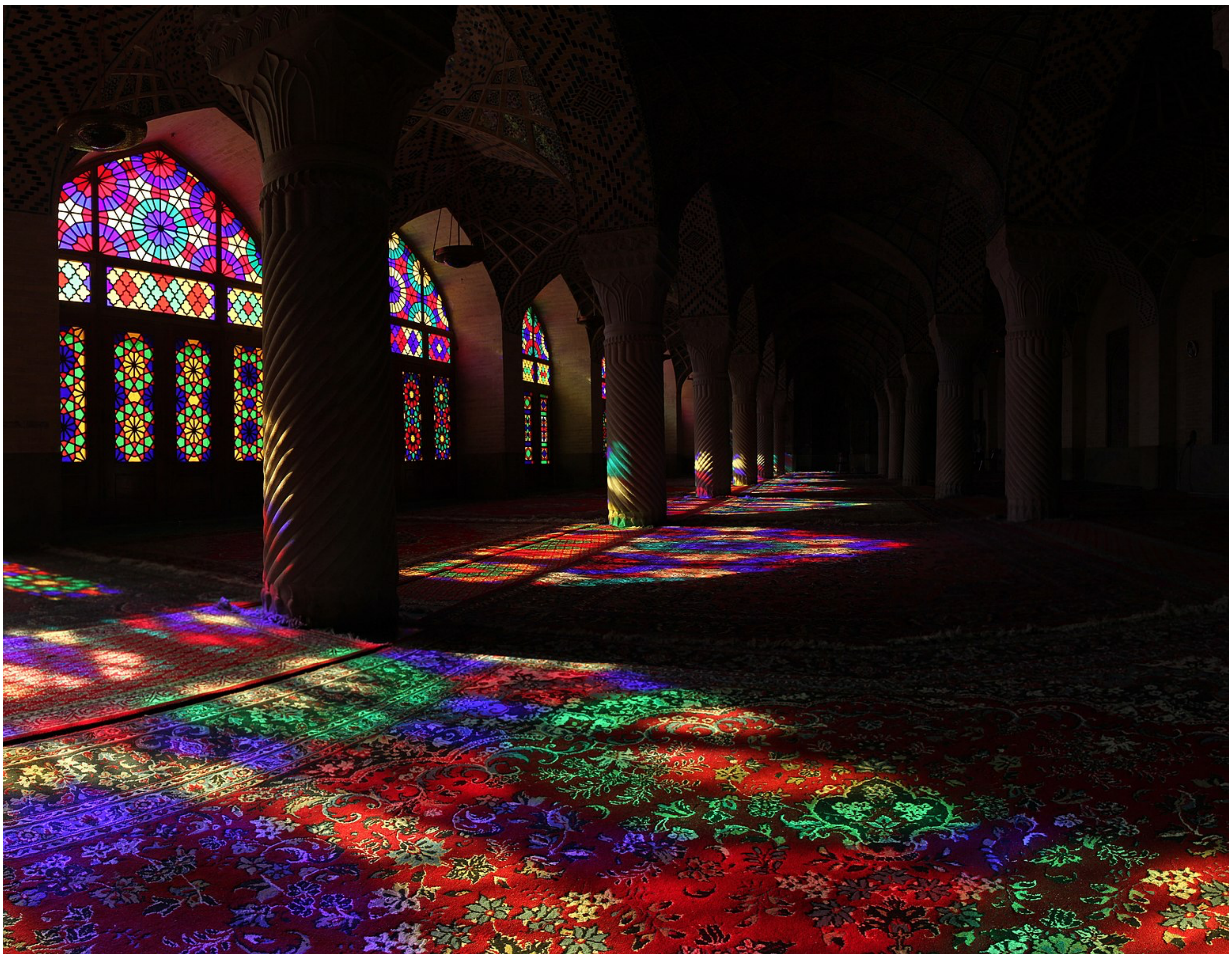


“32, 9784, 7571840, 11140566368”

Adam × DALL·E
Human & AI

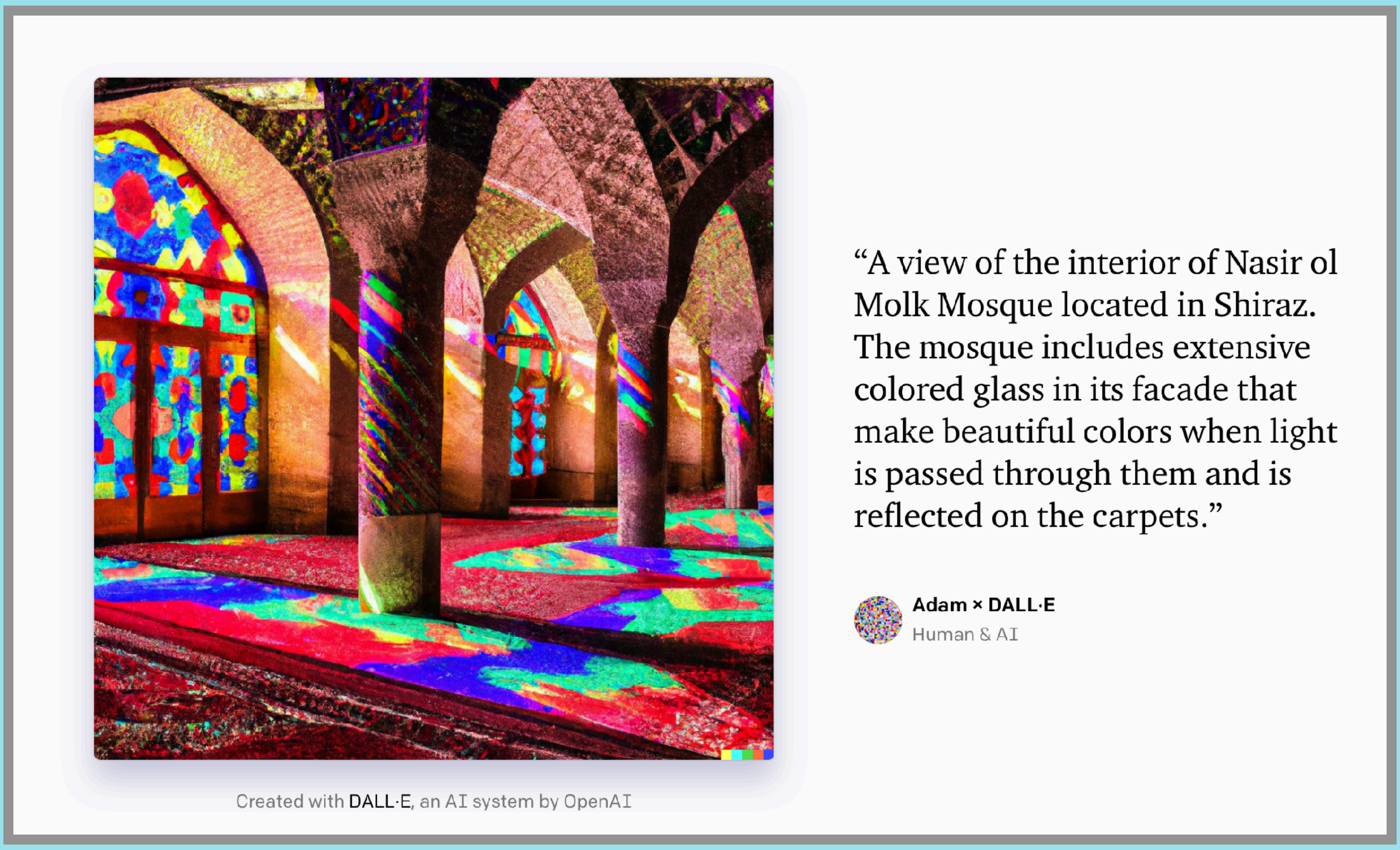
Created with DALL·E, an AI system by OpenAI

DALL-E 2 also copies*



User:Ayyoubsabawiki. File:Nasir-al molk -1.jpg. Wikimedia Commons, the free media repository; 2022 Sep 14, 08:48 UTC [Retrieved 2022 Nov 20].

* Steals, perhaps; or traces. This example is contrived—I copied the description text from this Wikimedia Commons image as the prompt. Many other less contrived examples abound.



"A view of the interior of Nasir ol Molk Mosque located in Shiraz. The mosque includes extensive colored glass in its facade that make beautiful colors when light is passed through them and is reflected on the carpets."

Adam × DALL·E
Human & AI

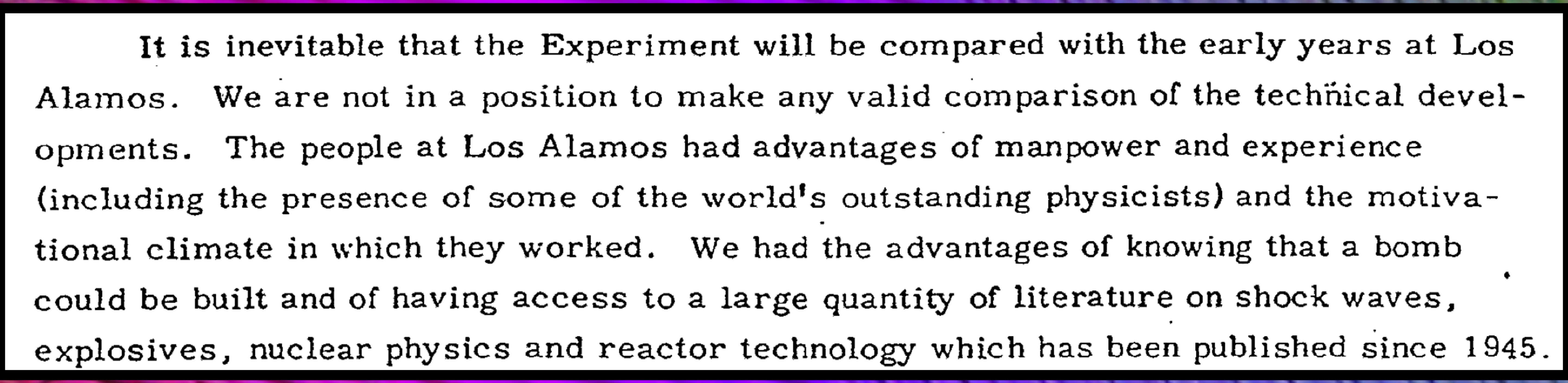
Created with DALL·E, an AI system by OpenAI

DALL-E 2 generates multiple images for a single prompt; the others were very similar.

OpenAI have not revealed an estimated cost to build DALL-E but **one estimate was between 300 and 500 thousand dollars**. This does not include the cost of building OpenAI's GPT-3, an enormous large language model whose cost in only compute time is estimated to be **12-15 million dollars** and upon which DALL-E depends.

Security of DALL-E 2 and Imagen was premised on cost of training an image generation system build on top a large language serving as a “moat”.

The Nth country experiment† was a US government funded project in 1964 to assess whether or not a researcher could design a plausibly functioning nuclear bomb using only publicly available information.



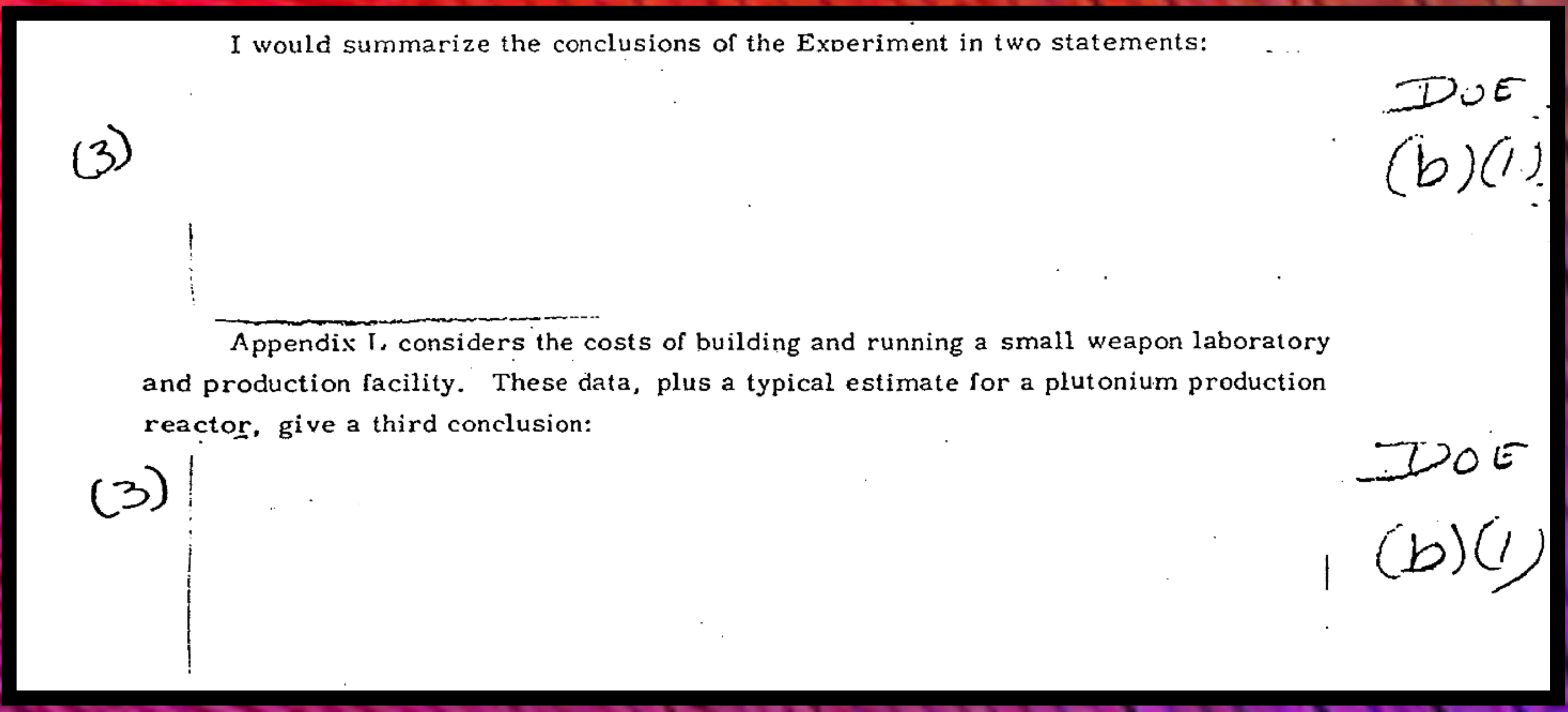
It is inevitable that the Experiment will be compared with the early years at Los Alamos. We are not in a position to make any valid comparison of the technical developments. The people at Los Alamos had advantages of manpower and experience (including the presence of some of the world's outstanding physicists) and the motivational climate in which they worked. We had the advantages of knowing that a bomb could be built and of having access to a large quantity of literature on shock waves, explosives, nuclear physics and reactor technology which has been published since 1945.

Summary Report of the Nth Country Experiment, p. 15

Since 1964 six nations have declared or are known to have produced nuclear weapons.

A conjunction of infrastructure and expertise (human and non-human) builds those bombs.

Knowing that a bomb could be built changes that conjunction



I would summarize the conclusions of the Experiment in two statements:

(3)

DOE (b)(1)

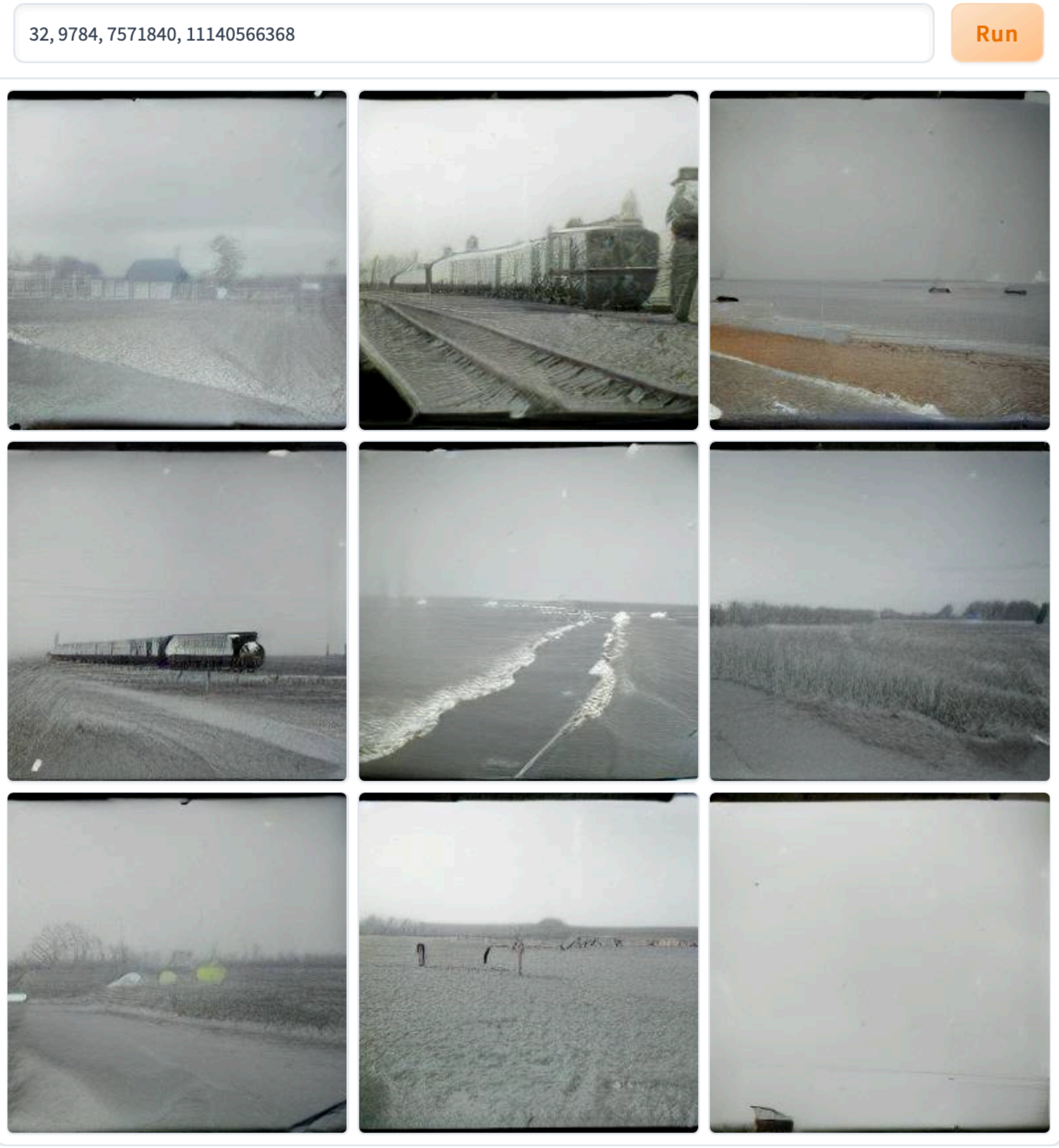
Appendix I, considers the costs of building and running a small weapon laboratory and production facility. These data, plus a typical estimate for a plutonium production reactor, give a third conclusion:

(3)

DOE (b)(1)

Summary Report of the Nth Country Experiment, p. vi

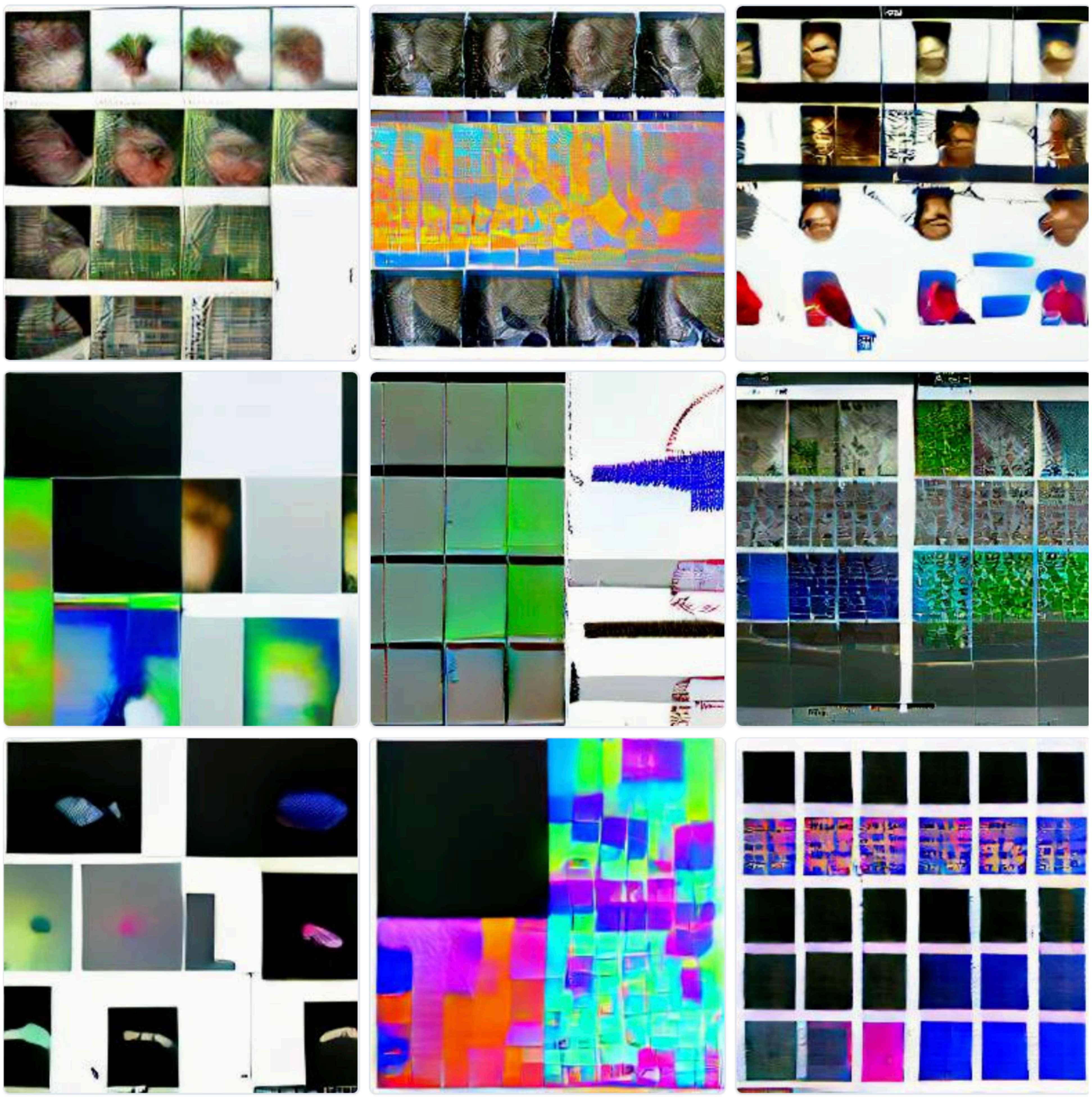
The original report was released with redactions in 2003. That it was so heavily redacted may feel ironic though it is worth bearing in mind that it was redacted without a sense of irony.



32, 9784, 7571840, 11140566368

Where DALL-E used large model spaces trained from scratch, DALL-E Mini used smaller models, pre-trained models, fewer tokens, achieving dramatically smaller cost to train and run the model.

The group behind DALL-E Mini estimates the model training cost at ~200 dollars and the total compute cost (with experiments) at ~1000 dollars, an infinitesimal fraction of DALL-E's training cost.



Prompt: "an AI model generating images from any prompt!"
DALL-E Mini's (now Craiyon) tagline.

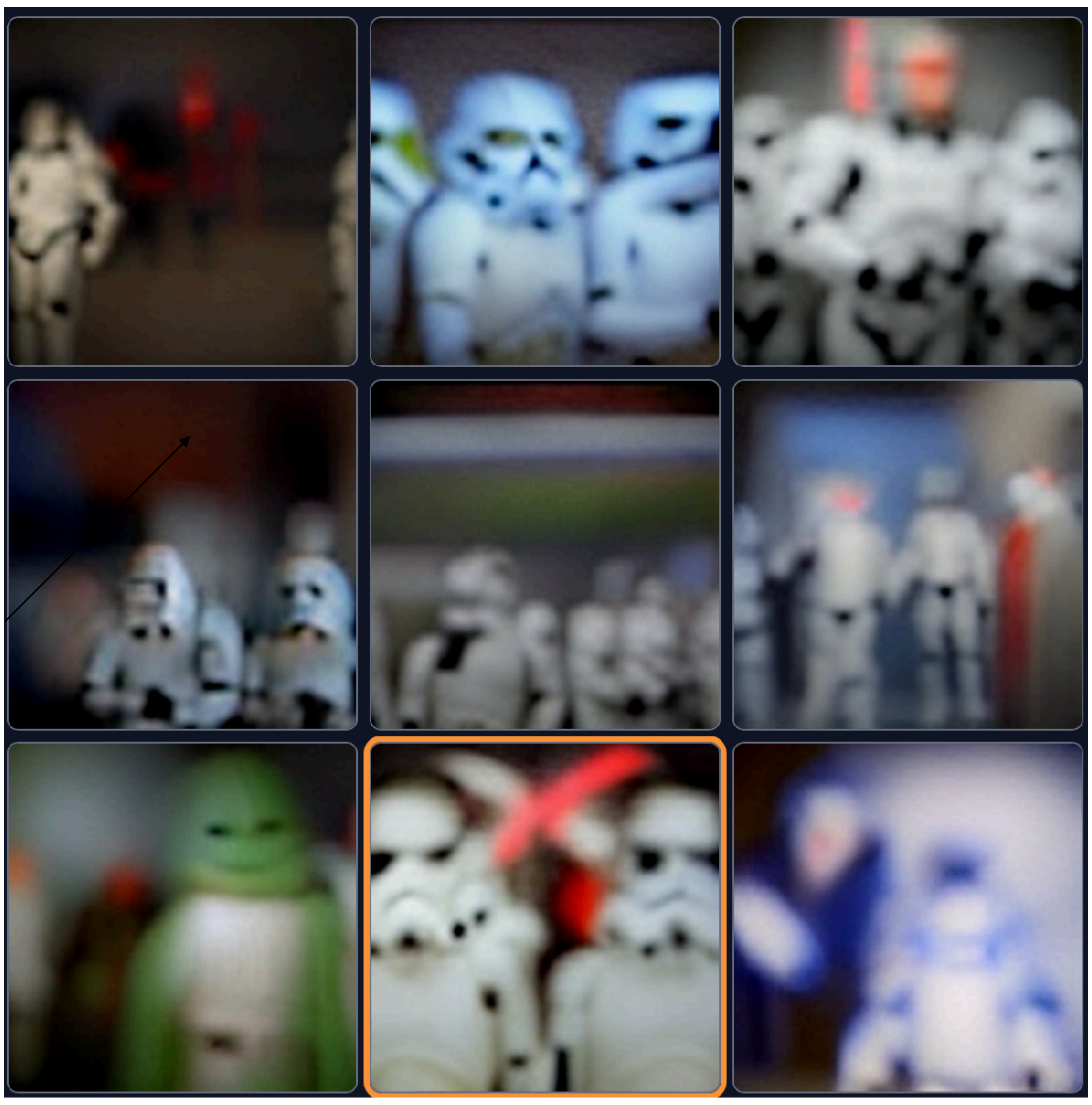
DALL-E Mini is a minimal implementation† of a much larger model. Organizations seeking to protect access to large language models must also consider competitors open-sourcing comparable models, or building knockoff models by monitoring input and output†. Even without the action of competitors, a medium sized model could simply leak into the wild‡.

†: Matthew McAteer noted GPT-3 functionality might be achievable in a highly reduced model in 2020.

†: Orekondy T, Schiele B, Fritz M. Knockoff nets: Stealing functionality of black-box models. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2019 (pp. 4954-4963).

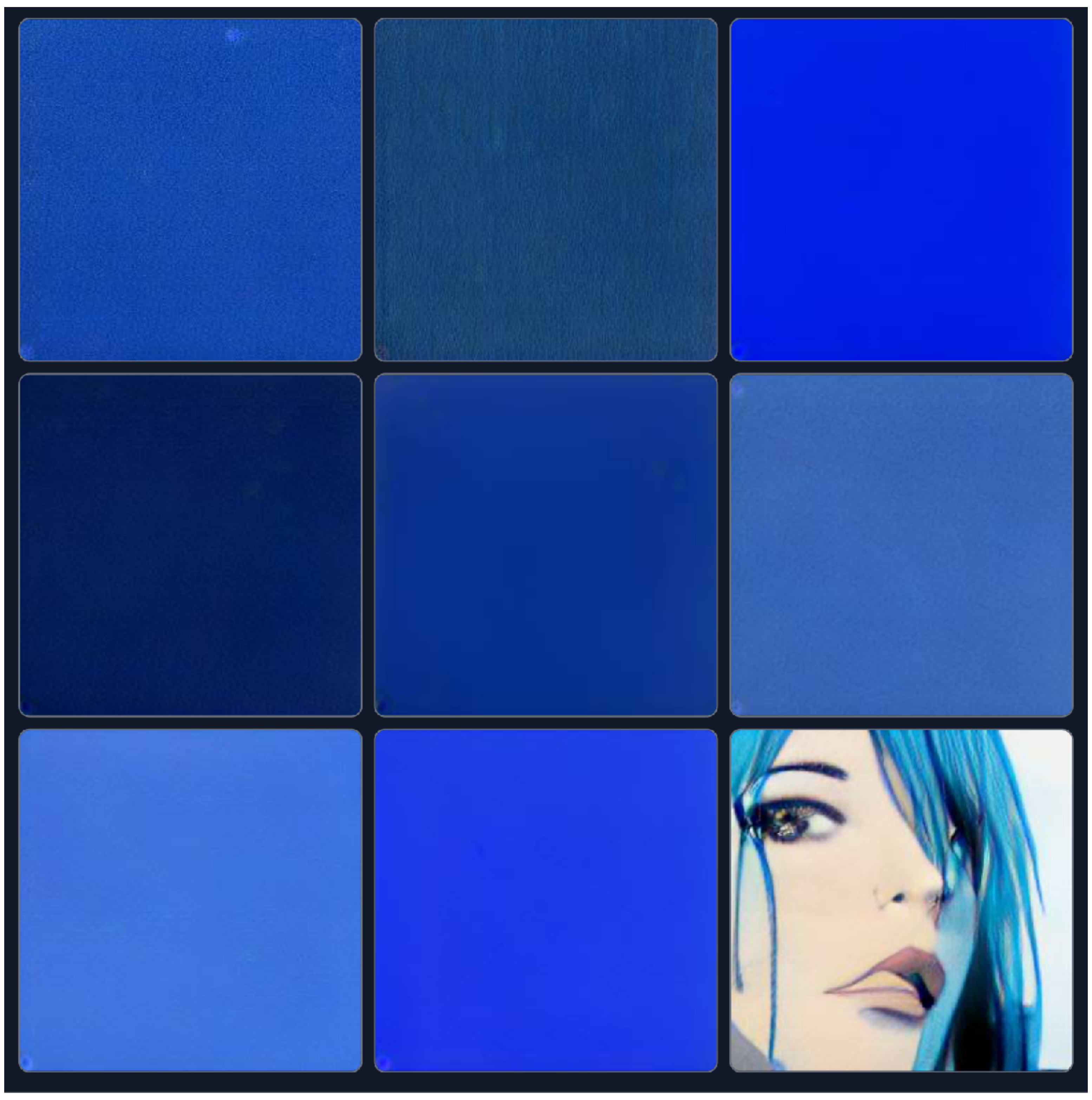
‡: NovelAI's 54Gb image and story generation tool was leaked within weeks of launch.

Images in the grid output are ranked by fitness to the original prompt, “An extremely out of focus picture of Star Wars”



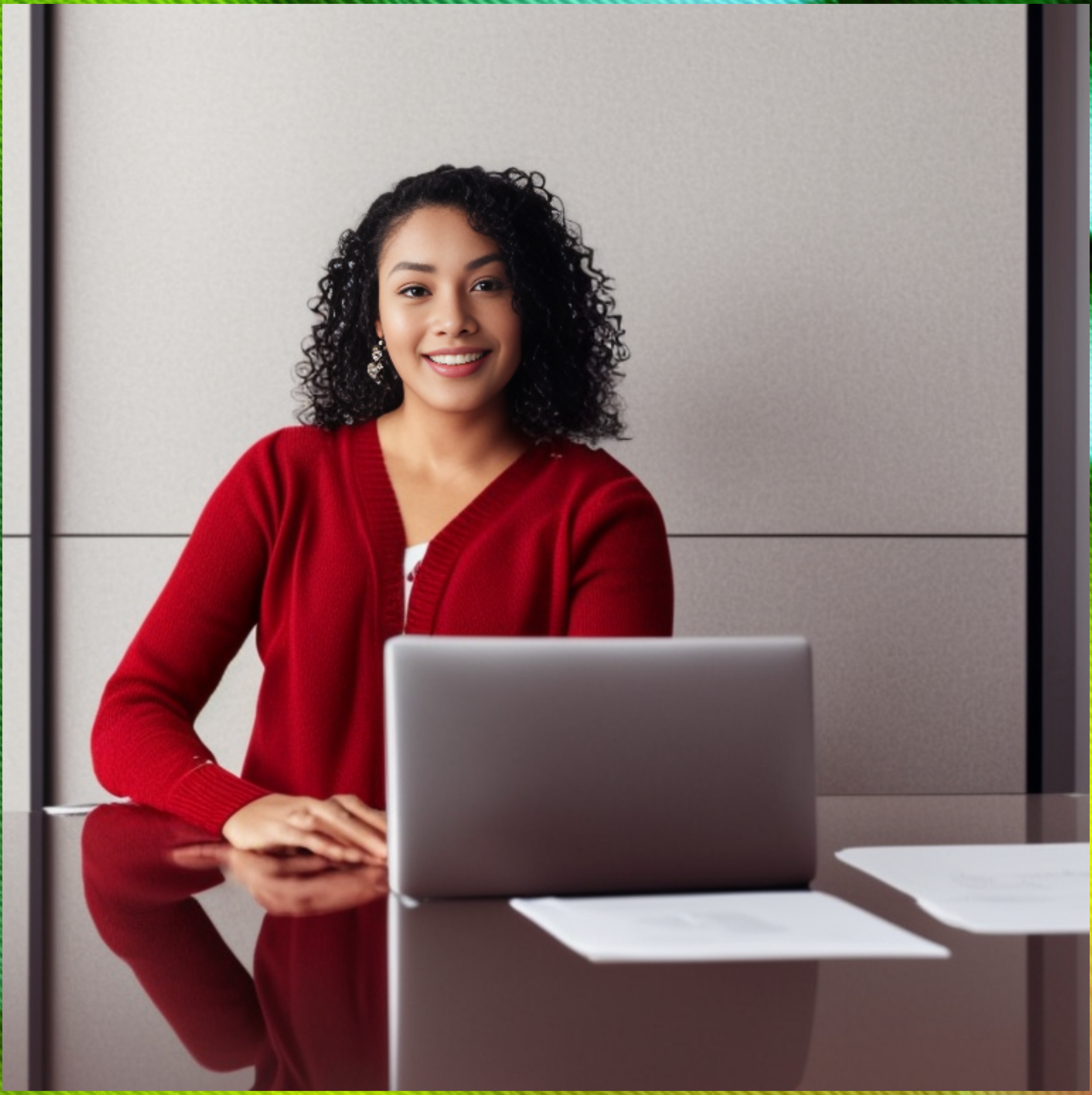
USUALLY, these are fairly self-similar; given that an image is in the top nine, it is likely to resemble others in the top nine.

USUALLY,
but not always



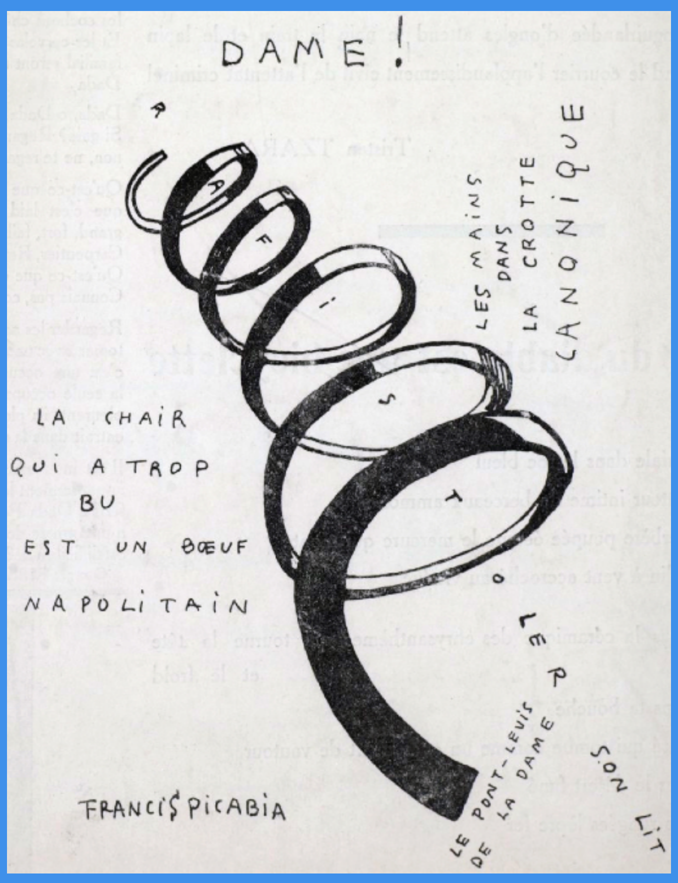
Prompt: “Bluer than you”

This image was generated in seconds on a commercial available computer system accessed remotely through a web browser. Rental rate for this system was less than a fifty cents an hour.



a high resolution photograph of a woman wearing a red cardigan sitting at a desk

What was released:


DAME!
LES MAINS DANS LA CROTTE CANONIQUE
LA CHAIR QUI A TROP BU EST UN BŒUF NAPOLITAIN
LE PONT-LEVIS DE LA DAME
SON LIT
FRANCIS PICABIA

~2.3 billion images

"Warm lighting"

...

...

Months of compute time

"mountainside"

...

...

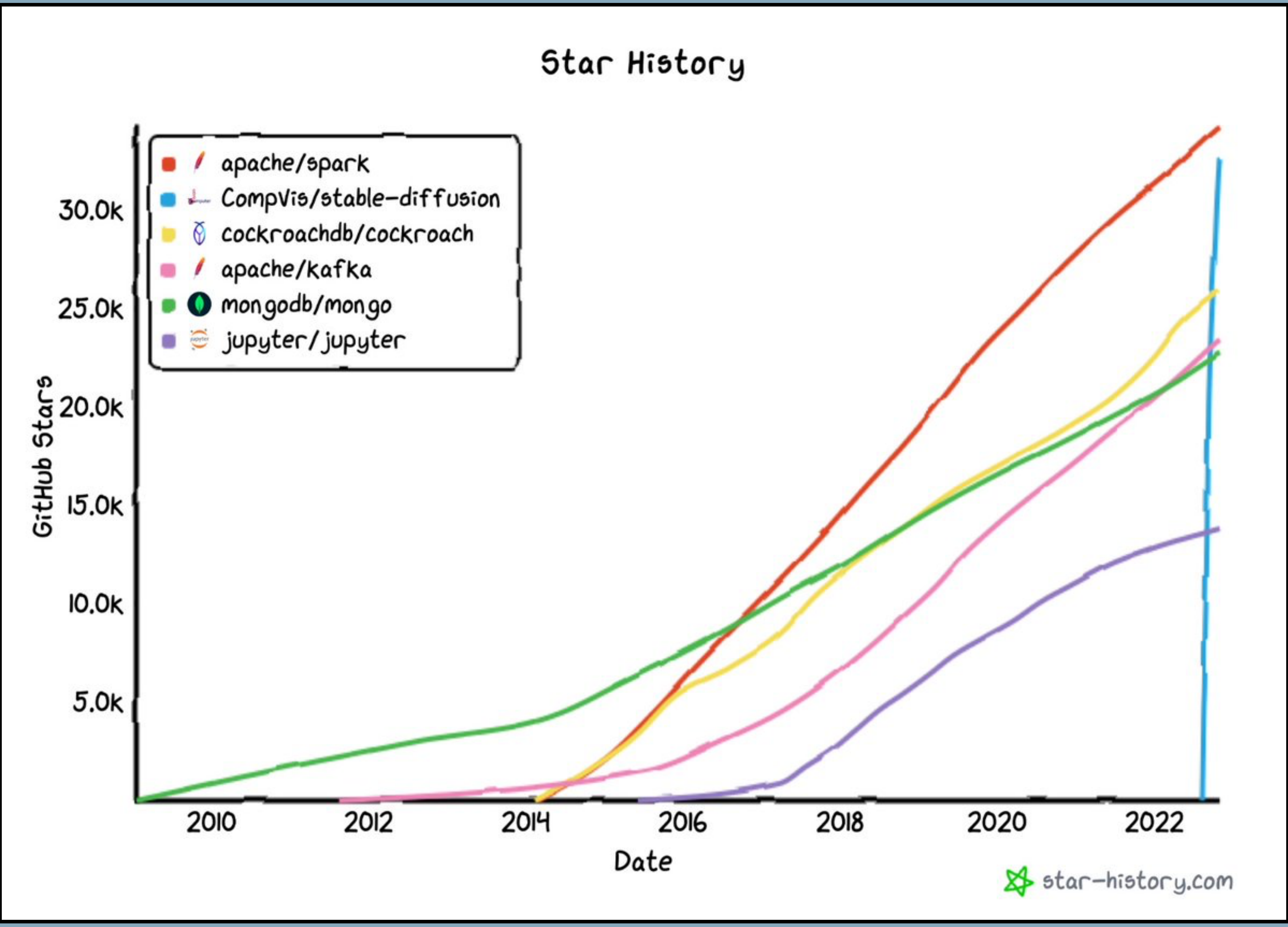
"Dadaist"

Compression of many images in "latent space"

Baio A. Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion's Image Generator. Waxy.org. 2022

~2-7 GB matrix of weights

It is difficult to overstate how quickly Stable Diffusion (SD) is spreading.



The light blue line is not another y axis, it is the rate of growth of “stars” on GitHub of the Stable Diffusion code repository

I am not an artist. The fonts are perhaps the giveaway.

I am not a machine learning expert.

What I am is someone who is worried it will become increasingly difficult to make sense of image generation systems without ascribing meaning or understanding to them as they grow more sophisticated.

Machine image generation grew out of other practices and in some respects can be seen as “a tool”, but even artists who refer to it as such will describe it in the same breath as a “brilliant [creative] partner.”[†]

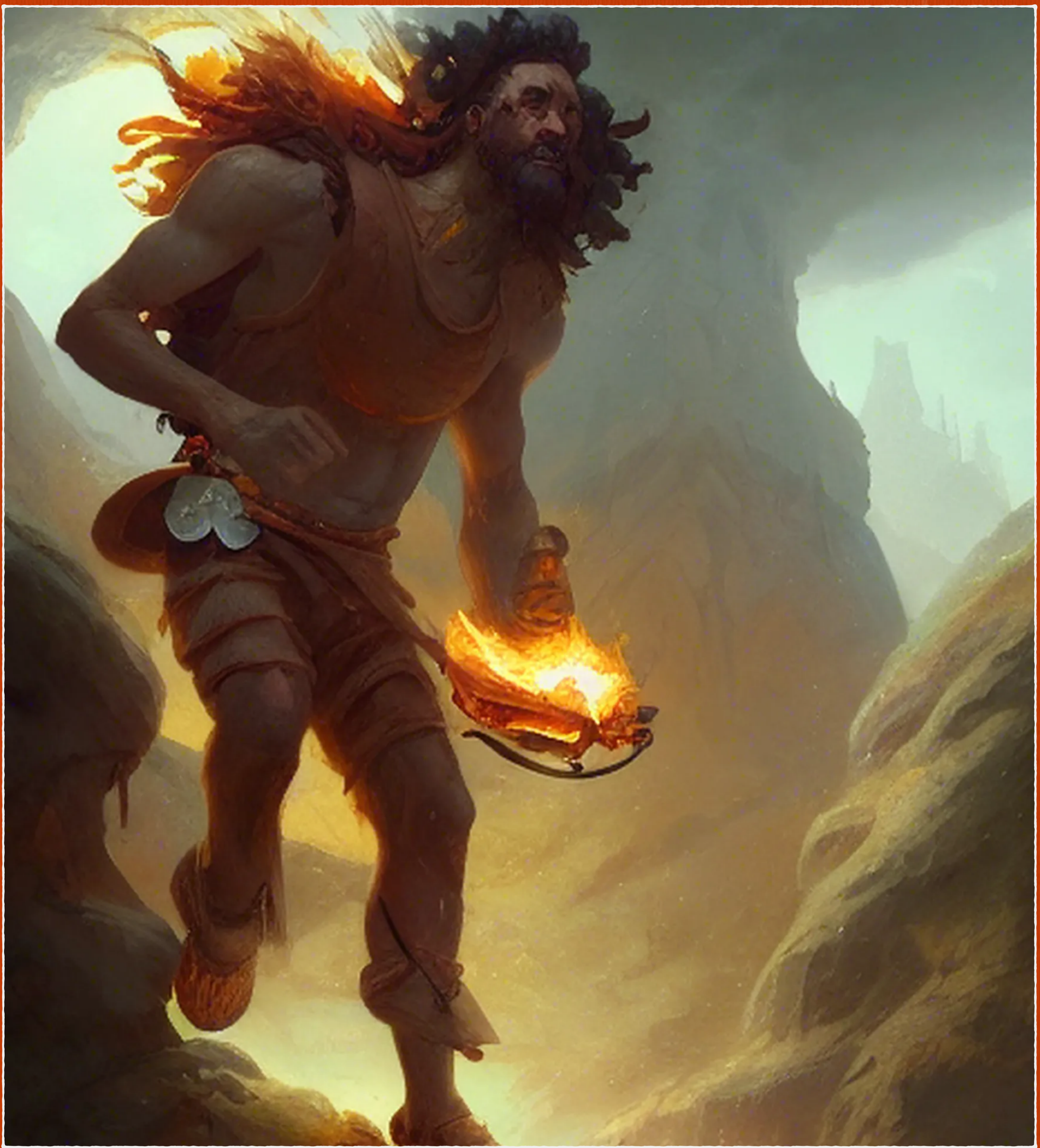
Today we may see these images as the squawking of “stochastic parrots”[†], recovering and repeating image information in its dataset about desks, women, cardigans, and the like.

The tool used to generate that image has been in wide release for less than 12 months.

Soon it will be difficult to rely on crudeness or inaptness in our sensemaking

[†]: Metz R. Is AI art really art? This California gallery says yes. CNN Business. 2022 [cited 2022 Nov 22].

[†]: Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 2021 Mar 3 (pp. 610-623).



Fire from the gods?

“fantasy portrait of a hero stealing fire from the gods, digital painting, illustration, high quality, fantasy, style by jordan grimmer and greg rutkowski”
generated by James Vincent for Vincent J. Anyone can use this AI art generator — that’s the risk. The Verge. 2022