https://doi.org/10.1038/s42256-022-00539-8

Collaborative creativity in AI

The public release of 'Stable Diffusion', a high-quality image generation tool, sets new standards in open-source AI development and raises new questions.

Generative AI tools provide creative inspiration for artists, illustrators and writers, but also for scientists — for example, in the discovery of drugs, materials and even the design of quantum experiments. Powerful capabilities arise in particular with very large neural network models that have billions or even trillions of parameters and are trained on vast amounts of data on the internet. For example, language models such as GPT-3 from OpenAI can generate original text as if written by humans. More recently, such algorithms have been trained on text–image pairs and developed into image generation models. DALL-E (from OpenAI), Imagen (from Google) and Midjourney produce stunning images in any style, from photorealistic portraits and newspaper cartoons to medieval tapestry and much more, given text or other images as prompts.

This editorial is illustrated with an image generated with a new text-to-image generation tool known as 'Stable Diffusion'. This model has been making a substantial impact worldwide in the last month, as it is completely open-source and free to use, unlike the aforementioned image generation models. Its public release on 22 August 2022 was accompanied with guidelines and a license that focus on responsible and ethical use and re-use of the model.

Plenty of ethical concerns come to mind with regard to generative models trained on data from the internet, which contains much harmful content and amplifies biases inherent in society. Moreover, generative algorithms could be used to produce unlawful 'deepfakes' that are damaging to individuals, and questions arise about copyright and undervaluing the work of professional artists. A concern that was in particular on the mind of OpenAI researchers in 2019 was the potential for malicious use of their language model (then GPT-2) in the form of the large-scale spread of misinformation. As a remedy, they released a smaller version of the model but this caused a backlash in the community. Many were unhappy with the fact that 'hype' was generated around GPT-2 while only few had access to the full model, making replication experiments challenging. It was argued that openness and engagement with the community would be a better way to examine any concerns and to promptly work on countermeasures when problems arise. GPT-3 is currently made available to researchers upon request, and for others there is controlled access via an API. Filters have been introduced to tackle misuse and problems with bias. DALL-E, from OpenAI, is available upon request and also has been endowed with various filters.

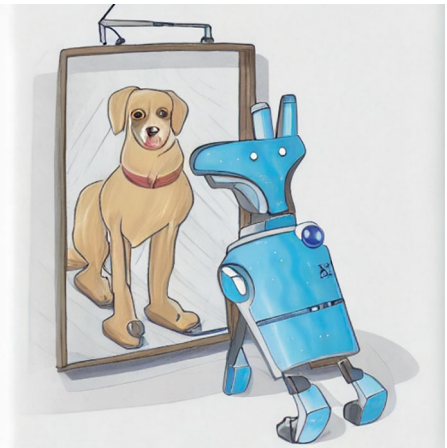

Illustration produced with the generative AI tool Stable Diffusion. The authors of a News & Views article in this issue, about a bio-inspired robotics approach, sent us a sketch of an idea for an illustration: a robot dog looking into a mirror and seeing a real dog in the reflection. This idea was turned into a *Nature*-style image to accompany the News & Views article. We tested what Stable Diffusion would come up with when shown the same sketch and given text prompts, and share here one example output image.

In an initiative to give the world a high-quality text-to-image generation large neural network model, Stable Diffusion was developed by a collaboration between several groups, including start-up stability.ai, non-profit AI platforms Eleuther AI and LAION, and researchers from the Machine Vision & Learning research group at LMU Munich. The model is built on the latent diffusion model published by the LMU group last year. In the approach, noise is added to a low-dimensional latent representation of example images, and a neural network is trained to revert this diffusion process. The de-noising process is guided by a text prompt. The algorithm was trained on a subset of the LAION-5B image database, using a cluster of 4,000 GPUs.

The way this project evolved as a collaborative, community-based effort, with the goal of making large-scale neural network tools available to all, points to a new direction for developing AI tools, outside academia or tech companies. Notably, the public release of Stable Diffusion came with a focus on ethical and responsible use. The model was released in collaboration with HuggingFace, another community-based collaborative platform, with a new license known as CreativeML OpenRAIL-M, based on work in Responsible AI Licenses (RAIL). This new license is designed to make the model free to use and re-use, even commercially, but emphasizes responsible use and makes users accountable for the results they generate. The license forbids sharing content that violates any laws, produces harm to individuals, spreads misinformation or targets vulnerable groups. The license also mandates using the same restrictions when re-distributing the model.

The question is whether this approach is too optimistic. A powerful AI generative tool that is freely available could turn out to produce illegal and unwanted content in a way that might be difficult to control. If open source is the answer, then the machine learning community needs to closely examine and monitor uses of the Stable Diffusion model and collectively tackle problems with malicious use if and when they arise. Whatever happens next — and the pace of development and creative uses of Stable Diffusion and related models is very fast — the community-based approach and public release of a powerful machine learning model accompanied by a vision for encouraging responsible use is setting new standards for the field.

Published online: 22 September 2022