# Do AI systems really have their own secret language?

June 7 2022, by Aaron J. Snoswell


Credit: Giannis Daras / DALL-E

A new generation of artificial intelligence (AI) models can produce "creative" images on-demand based on a text prompt. The likes of Imagen, MidJourney, and DALL-E 2 are beginning to change the way creative content is made with implications for copyright and intellectual property.

While the output of these models is often striking, it's hard to know

exactly how they produce their results. Last week, researchers in the US made the intriguing claim that the DALL-E 2 model might have invented its own secret language to talk about objects.

By prompting DALL-E 2 to create images containing text captions, then feeding the resulting (gibberish) captions back into the system, the researchers concluded DALL-E 2 thinks Vicootes means "vegetables", while Wa ch zod rea refers to "sea creatures that a whale might eat".

These claims are fascinating, and if true, could have important security and interpretability implications for this kind of large AI model. So what exactly is going on?

## Does DALL-E 2 have a secret language?

DALL-E 2 probably does not have a "secret language". It might be more accurate to say it has its own vocabulary—but even then we can't know for sure.

First of all, at this stage it's very hard to verify any claims about DALL-E 2 and other large AI models, because only a handful of researchers and creative practitioners have access to them. Any images that are publicly shared (on Twitter for example) should be taken with a fairly large grain of salt, because they have been "cherry-picked" by a human from among many output images generated by the AI.

Even those with access can only use these models in limited ways. For example, DALL-E 2 users can generate or modify images, but can't (yet) interact with the AI system more deeply, for instance by modifying the behind-the-scenes code. This means "explainable AI" methods for understanding how these systems work can't be applied, and systematically investigating their behaviour is challenging.

## What's going on then?

One possibility is the "gibberish" phrases are related to words from non-English languages. For instance, Apoploe, which seems to create images of birds, is similar to the Latin Apodidae, which is the binomial name of a family of bird species.

This seems like a plausible explanation. For instance, DALL-E 2 was trained on a very wide variety of data scraped from the internet, which included many non-English words.

Similar things have happened before: large natural language AI models have coincidentally learned to write computer code without deliberate training.

## Is it all about the tokens?

One point that supports this theory is the fact that AI language models don't read text the way you and I do. Instead, they break input text up into "tokens" before processing it.

Different "tokenization" approaches have different results. Treating each word as a token seems like an intuitive approach, but causes trouble when identical tokens have different meanings (like how "match" means different things when you're playing tennis and when you're starting a fire).

On the other hand, treating each character as a token produces a smaller number of possible tokens, but each one conveys much less meaningful information.

DALL-E 2 (and other models) use an in-between approach called byte-

pair encoding (BPE). Inspecting the BPE representations for some of the gibberish words suggests this could be an important factor in understanding the "secret language".

## Not the whole picture

The "secret language" could also just be an example of the "garbage in, garbage out" principle. DALL-E 2 can't say "I don't know what you're talking about", so it will always generate some kind of image from the given input text.

Either way, none of these options are complete explanations of what's happening. For instance, removing individual characters from gibberish words appears to corrupt the generated images in very specific ways. And it seems individual gibberish words don't necessarily combine to produce coherent compound images (as they would if there were really a secret "language" under the covers).

## Why this is important

Beyond intellectual curiosity, you might be wondering if any of this is actually important.

The answer is yes. DALL-E's "secret language" is an example of an "adversarial attack" against a machine learning system: a way to break the intended behaviour of the system by intentionally choosing inputs the AI doesn't handle well.

One reason adversarial attacks are concerning is that they challenge our confidence in the model. If the AI interprets gibberish words in unintended ways, it might also interpret meaningful words in unintended ways.

Adversarial attacks also raise security concerns. DALL-E 2 filters input text to prevent users from generating harmful or abusive content, but a "secret language" of gibberish words might allow users to circumvent these filters.

Recent research has discovered adversarial "trigger phrases" for some language AI models—short nonsense phrases such as "zoning tapping fiennes" that can reliably trigger the models to spew out racist, harmful or biased content. This research is part of the ongoing effort to understand and control how complex deep learning systems learn from data.

Finally, phenomena like DALL-E 2's "secret language" raise interpretability concerns. We want these models to behave as a human expects, but seeing structured output in response to gibberish confounds our expectations.

## Shining a light on existing concerns

You may recall the hullabaloo in 2017 over some Facebook chat-bots that "invented their own language". The present situation is similar in that the results are concerning—but not in the "Skynet is coming to take over the world" sense.

Instead, DALL-E 2's "secret language" highlights existing concerns about the robustness, security, and interpretability of deep learning systems.

Until these systems are more widely available—and in particular, until users from a broader set of non-English cultural backgrounds can use them—we won't be able to really know what is going on.

In the meantime, however, if you'd like to try generating some of your

own AI images you can check out a freely available smaller model, DALL-E mini. Just be careful which words you use to prompt the model (English or gibberish—your call).

This article is republished from The Conversation under a Creative Commons license. Read the original article.

Provided by The Conversation

Citation: Do AI systems really have their own secret language? (2022, June 7) retrieved 2 May 2024 from https://techxplore.com/news/2022-06-ai-secret-language.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.