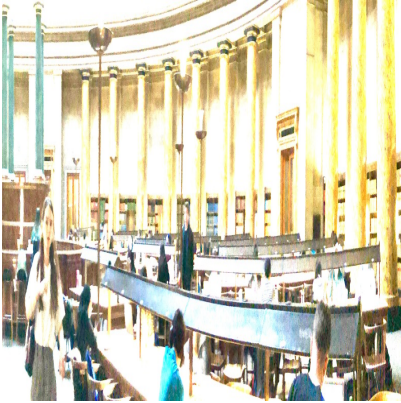




Benefits of artificial misinformation

Description

Natural language processing (NLP) programs seem able to take as input words and phrases provided by a human operator and generate coherent sentences in response, simulating a kind of dialogue that some scholars (e.g. [David Chalmers](#)) think hints of general intelligence. I think that one aspect of that dialogue is to simulate how human interlocutors are able to make things up to invent stories that display internal coherent but that pay little regard to the wider context.

As demonstrated in a [previous post](#), the GPT-3 natural language processing algorithm as deployed in the [Copysmith](#) app was able to invent a quote about architectural security that I had not encountered before:

In the 19th century Sir James Fergusson, author of the 1876 work *A History of Architecture in All Countries*, noted that "in all ages and in all countries, the manufacturers of locks, locksmiths as they are called, have always held an important position in every community".

Google Books provides searchable access to all three volumes of James Ferguson's *A History of Architecture in All Countries: From the Earliest Times to the Present Day* (London: John Murray, 1862-1876). I can't find the quote or reference to locksmiths in Ferguson's books (with or without double "s"). A general Google search on phrases within the Ferguson quote come up either with documents that make no reference to Ferguson, or with my own blog post from last week!

Truths that start as lies

The GPT-3 system is trained on over 3.5 thousand million pages of text (45 terabytes) gathered from a diverse range of Internet sources during a specific period. So the algorithm stopped training before release. That GPT-3 generates falsehoods is encouraging. It indicates that the software produces new material not in the original training corpus of texts. Nor is the output simply copy-pasted from wikipedia. On the other hand, the system easily generates sentences that fall under the headings of half-truths, lies and misinformation.

To the extent that the GPT-3 algorithmic process corresponds to human cognition, it confirms some very human aspect of thinking, writing, creation and perception. Thanks to our experiences, backgrounds and biases, we humans are prone to making things up as an initial part of the interpretive process. (I think this assertion accords with Anil Seth's understanding that perception begins with [hallucination](#), and Heidegger's concepts of *fore-projection* in the processes of interpretation.)

In the moment of creation our initial tentative conjectures need to be tested by the circumstances of the conversation. If I was writing about architectural security I may well have started a blog post with an inkling that some architectural historian had written about the significance of locksmiths, an insight that as a scholar I would have been bound to confirm, disconfirm or modify before publishing.

GPT-3 outputs are inevitably a first blush at a generative language event, unfiltered by circumstances, a consideration of the material context, or some verification and fact checking. Hence, the developers of [Copysmith](#) and other apps ([openAI.com](#)) promote the GPT-3 platform as an *aid* to writing. You can use it to initiate a writing process, use it to overcome writer's block or as an aid that continues the development of ideas a bit like the multi-author exquisite corpse writing of the surrealists. In any event, the process requires a human editor to curate the output.

One of the challenges afforded by GPT-3-style apps is that if left to run un-curated they contribute further output to the pool of unfiltered, unverified statements that already pervade partisan news reporting, blogs, and the social media space. GPT-3 comes pre-trained on Internet sources, including factual errors, political slogans, unfounded conspiracy theories, discriminatory language, and hate speech as well as Shakespeare, poetry, novels, tech-science articles, edifying self-help books, literary classics and benign examples of everyday language use.

In-so-far as automated natural language processing reflects aspects of human cognition and language, it highlights that language operates effectively not in spite of the author's preconceptions and biases, but it needs these to function. In a discursive context most clear-thinking individuals adjust and modify their propositions before they utter them, something that as yet the generalised GPT-3 process is unable to accomplish, nor is it designed to do that. See post: [The bliss of ignorance](#).

Bibliography

- Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. "Language models are few-shot learners." *Cornell University arXiv*, 22 July. Accessed 2 June 2022. <https://arxiv.org/abs/2005.14165>
- Cooper, Kindra. "OpenAI GPT-3: Everything You Need to Know." *Springboard*, 1 November. Accessed 5 June 2022. <https://www.springboard.com/blog/data-science/machine-learning-gpt-3-open-ai/>
- Kahneman, Daniel. *Thinking, Fast and Slow*. London: Penguin, 2011.
- Weinberg, Justin. "Philosophers On GPT-3 (updated with replies by GPT-3)." *Daily Nous*, 30 July. Accessed 2 Jun 2022. <https://dailynous.com/2020/07/30/philosophers-gpt-3>

Notes

- Featured image is the Manchester Central Library by architect Vincent Harris (1876-1971) and restored by Ryder Architects.
- NLP programs operate with "tokens" • these are "sequences of characters found in text. The models understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens," according to a demonstrator at <https://beta.openai.com/tokenizer>.

Tokens	Characters
97	480

Natural language processing (NLP) programs seem able to take as input words and phrases provided by a human operator and generate coherent sentences in response, simulating a kind of dialogue that some scholars (e.g. David Chalmers) think "hints of general intelligence." I think that one aspect of that dialogue is to simulate how human interlocutors are able to make things up -- to invent stories that display internal coherent but that pay little regard to the wider context.

Category

1. Artificial Intelligence

Tags

1. GPT-3
2. hermeneutics
3. natural language understanding
4. NLP

Date Created

June 18, 2022

Author

rcoyne99