Vee, Annette. (December 2022). “Automated Trolling: The Case of GPT-4Chan When Artificial Intelligence is as Easy as Writing.” *Interfaces: Essays and Reviews in Computing and Culture Vol. 3, Charles Babbage Institute, University of Minnesota,* 102-111.

Automated Trolling: The Case of GPT-4Chan
When Artificial Intelligence is as Easy as Writing

Annette Vee, University of Pittsburgh

Abstract: In June 2022, a machine learning researcher released a large language model (LLM) trained on a dataset largely consisting of hate speech, memes, and other language from an infamous online trolling space: 4-chan. Anyone with a little technical knowledge about LLMs could download the model—called GPT-4chan—and use it to produce more of this language. Critics pounced and the model was taken down. This article explores what it means to have AI-based text-generation technology such as LLMs available to the general public through open code and datasets. Writing is powerful, and AI-generated writing maybe even more so. In a rapidly approaching future of automated writing, what can we do when this writing is designed to embrace bias or create harm?

What happens when you take the most edgelordy language on the internet and train a bot to produce more of it? Enter the cheekily-named *GPT-4chan*. Feed it an innocuous seed phrase and it might reply with a racial slur (Cramer, 2022a) or a rant about illegal immigrants (Austin Anderson, 2022) Or ask it how to get a girlfriend and it will tell you "by taking away the rights of women" (JJADX, 2022).

Released in early June to great controversy among AI ethicists and machine learning researchers, GPT-4chan is the bastard child of a pretrained large language model (like the GPT series) and a dataset of posts from the infamous “politically incorrect” board on 4chan, brought together by a trolling researcher with a point to prove about machine learning.

The GPT-4chan model release rains on the parade of open research online. Most research in AI and natural language generation is directed toward eliminating bias. This is a story about a language model designed to embrace bias, and what that might mean for a future of automated writing.

Vee, Annette. (December 2022). “Automated Trolling: The Case of GPT-4Chan When Artificial Intelligence is as Easy as Writing.” *Interfaces: Essays and Reviews in Computing and Culture Vol. 3, Charles Babbage Institute, University of Minnesota,* 102-111.

Figure 1: A happy typewriter in a fiery hellscape, as imagined by Midjourney, an AI program that generates images from textual prompts.

The Birth of GPT-4chan

4chan’s “Politically Incorrect” /pol message-board is the most notoriously high-profile cesspool of language on the Internet. If you’re looking for misogynist comics about female scientists or maps of non-white births in Europe, 4chan’s “Politically Incorrect” message board can hook you up. Posters—all anonymous, or “anons”—go there to share offensive terms and scenarios in memey images and trollish language. Go ahead and think of the most terrible things you can. They have that! And more. The board is an incubator for innovative expressions of misogyny, racism, conspiracy theories, and encouragement for self-harm.

To create GPT-4chan, YouTuber and machine learning researcher Yannic Kilcher took a publicly-available, pre-trained large language model from the open site HuggingFace and trained it on a publicly available dataset, “Raiders of the Lost Kek” (Papasavva et al., 2020), that included over 134 million posts from 4chan/pol.

It worked. Kilcher says in his video announcing the “worst AI ever:” “I was blown away. The model was good, in a terrible sense. It perfectly encapsulated the mix of offensive, nihilism, trolling, and deep distrust of any information whatsoever that permeates most posts on /pol” (Kilcher, 2022a).

He then created a few bot accounts on 4chan/pol and used his fine-tuned GPT-4chan model to fuel their posts. These bots fed /pol’s language back to the /pol community, thus pissing in a sea of piss, as /pol gleefully calls such activity.

Because the /pol board is entirely anonymous, it took a little sleuthing for the human anons to sniff out the bots and distinguish them from Fed interlopers—which the board perceives as a constant threat. But after a few days, they did figure it out. Kilcher then made a few adjustments to the bots and sent them back in. All told, Kilcher’s bots posted about 30,000 posts in a few days. Then, on June 3, Kilcher released a quick-cut, click-baity YouTube video exposing how he trolled the trolls with “the worst AI ever.”

Kilcher presents himself as a kind of red-teamer, that is, someone intentionally creating malicious output in order to better understand the system, testing its limits to show how it works or where its vulnerability lies. As he describes his experiment with “the most horrible model on the Internet,” he critiques a particular benchmark of AI language generators: TruthfulQA. Benchmarks such as TruthfulQA, which provides 817 questions to measure how well a language model answers questions truthfully, are a common tool to assess LLMs. Because the blatantly toxic GPT-4chan scores higher than other well-known and less offensive models, Kilcher makes a compelling point about the poor validity of this particular benchmark. Put another way, GPT-4chan makes a legitimate contribution to AI research.

In his video, Kilcher features only GPT-4chan’s most anodyne output. However, he mentions that he included the raw content in an extra video, linked in the comments. If you click on that video, you’ll learn just how brilliant a troll Kilcher is. Kilcher admits that GPT-4chan is awful. But he released it anyway and is clearly enjoying some lulz from the reaction: “AI Ethics people just mad I Rick rolled them,” he tweeted (Kilcher, 2022b)

Language without understanding

Writing about LLMs like the GPT series in 2021, Emily Bender, Timnit Gebru and colleagues delineated the “dangers of stochastic parrots”— language models that, like parrots, were trained on a slew of barely curated language and then repeated words without understanding them. Like the old joke about the parrot who repeats filthy language when the priest visits, language out of context carries significant social risks at the moment of human interpretation.

What makes GPT-4chan’s response about how to get a girlfriend so devastating is the context—who you imagine to be having this exchange, and the currently bleak landscape of women’s rights. GPT-4chan doesn’t get the dark humor. But we do. An animal or machine that produces human language without understanding is uncanny and disturbing, because they seem to know something about us—yet *we* know they really can’t *know* anything (Heikkilä, 2022).

Brazen heads—brass models of men’s heads that demonstrated the ingenuity of their makers through speaking wisdom—were associated with alchemists of the early Renaissance. Verging on magic and heresy, talking automata were both proofs of brilliance and charlatanism from the Renaissance to the Enlightenment. Legend has it that 13th century priest Thomas Aquinas once destroyed a brazen head for reminding him of Satan.

GPT-4chan—a modern-day brazen head—has no conscience or understanding. It can produce hateful language without risk of a change of heart. What’s more, it can do it at scale and away from the context of /pol.

Figure 2: Steampunk robot head, as envisioned by Midjourney, an AI program that generates images from textual prompts.

When OpenAI released GPT-2 in 2019, they decided not to release its full model and dataset for fear of what it could do in the wrong hands: impersonate others; generate misleading news stories; automate spam or abuse through social media (OpenAI, 2019). Implicitly, OpenAI admitted that writing is powerful, especially at scale. We know now that the interjection of automated writing during the 2016 election certainly shaped its discourse (Laquintano and Vee, 2017).

Of course, that danger hasn’t stopped OpenAI from eventually releasing the model as well as an even better one, GPT-3. So much for the warnings about LLMs of Bender, Gebru and others. Gebru was even fired from Google in a high-profile AI ethics dispute over the “stochastic parrots” paper (Simonite, 2021b). Another author of the paper, Margaret Mitchell, was also fired from Google a few months later (Simonite, 2021a). LLMs are dangerous, but it’s also apparently dangerous to talk about that fact.

The Censure of Unbridled AI

AI ethicists are rightly concerned about the release of GPT-4chan. A model trained on 4chan/pol’s toxic language, and then released to the public, presents clear possibilities for harm. The language on 4chan/pol is objectionable by design, but you have to go looking for it to find it. What happens when that language is automated and then packaged for use elsewhere? One rude parrot repeating words from one rude person makes for a decent joke, but the humor dissipates among an infinite flock of parrots potentially trained on language from any context and released anywhere in the world.

Critics argue that Kilcher could have made his point about the poor benchmark without releasing the model (Oakden-Rayner, 2022b; Cramer, 2022b). And although few tears should be shed for the /pol anons who were fed the same hateful language they produce, Kilcher did deceive them when he released his bots on their board.

Percy Liang, a prominent AI researcher from Stanford, issued a public statement on June 21 censuring the release of GPT-4chan (Liang, 2022). Both the deception and the model release are clear violations of research ethics guidelines that

are standard to institutional review boards (IRBs) at universities and other research institutions. One critic cited medical guidelines for ethical research (Oakland-Raymer, 2022a). But Kilcher did this on his own, outside of any institution, so he was not governed by any ethical reviews. He claims it was “a prank and light-hearted trolling” (Gault, 2022).

Figure 3 Happy, green trolls dancing with their hands up, as envisioned by Midjourney, an AI platform that generates images from textual prompts.

AI research used to be done almost exclusively within elite research institutions such as Stanford. It’s long been considered a cliquish field for that reason. But with so many open resources to support AI research out there—models, datasets, computing, plus open courses that teach machine learning—formal institutions have lost their monopoly on AI research. Now, more AI research is done in private contexts, outside of universities, than inside (Clark, 2022).

In AI research—as with the Internet more generally—we are seeing what it means to play out the scenario Clay Shirky named in his 2008 book: *Here Comes Everybody*. When the tools for research are openly available, free, and online, we get a blossoming of new perspectives. Some of those perspectives are morally questionable.

In other words, there’s more at stake in Liang’s letter than Kilcher’s ethical violations. The signatories—360 as of July 5—generally represent formal research and tech institutions such as Stanford and Microsoft. Liang and the signatories argue that LLMs carry significant risk and currently lack community norms for their deployment. Yet they argue, “it is essential for members of the AI community to condemn clearly irresponsible practices” such as Kilcher’s. Let’s be clear: this is a couple hundred credentialed AI researchers writing an open letter to thousands, perhaps millions, of machine learning enthusiasts and wannabes using free and open resources online.

Is there such a thing as “the AI community?” When AI research is open, can it have agreed-upon community guidelines? If so, who should control those guidelines and reviews?

The Promise and Peril of Open Systems

The platform Hugging Face—the platform Kilcher used for GPT-4chan--has emerged quickly to be the go-to hub of machine learning models. It features popular natural language processing models such as BERT and GPT-2 as well as image-generation models such as DALL-E and offers both free and subscription-based options for machine learning researchers to access sophisticated models, learn, and collaborate.

The primary dataset used to pretrain GPT-J, the model Kilcher used for GPT-4chan, is Common Crawl. Common Crawl is maintained by a non-profit organization of the same name whose stated, “goal is to democratize the data so everyone, not just big companies, can do high-quality research and analysis” (Common Crawl, “Home page”). Diving further, we see that Common Crawl uses Apache Hadoop—another open source resource—to help crawl the Web for data. The data is stored on Amazon Web Services, a paid service for the level of storage Common Crawl uses, but also a corporate-controlled and accessible one (Common Crawl, “Registry”). The Common Crawl dataset is free to download.

The dataset for GPT-4chan—containing over 3.5 million posts from the /pol “politically incorrect” message board—is also free to download. The authors of the paper releasing the 4chan/pol dataset rate posts with toxicity scores and “are confident that [their] work will motivate and assist researchers in studying and understanding 4chan, as well as its role on the greater Web” (Papasavva, 2020).

Indeed, they have! In fact, the sources of all technical keystones for GPT-4chan—the model, the training dataset, and the fine-tuning dataset—have ostensibly furthered their mission through Kilcher’s work with the vile GPT-4chan.

Kilcher made the GPT-4chan model and the splashy, viral-ready video that promoted it. But other responsible parties for this model could include: anonymous 4chan posters; the researchers who scraped the dataset GPT-4chan was trained on; OpenAI for developing powerful LLMs; Hugging Face for supporting open collaboration on LLMs; and all the other open systems needed to produce these tools and data. Where does the responsibility for GPT-4chan’s language begin and end? Do the makers of these tools also merit censure?

OpenAI recognized (and later shoved aside) the danger of open models when they withheld GPT-2. Bender, Gebru and colleagues also warned against the openness of large language models. They knew with these open tools, it was only a matter of time for someone to produce something like GPT-4chan.

Figure 4 A futuristic city made of alphabetic letters, as envisioned by Midjourney, an AI program that generates images from textual prompts.

With the open systems and resources supporting machine learning and LLMs, the determination of wrong and right is in the hands not of a like-minded “community,” but a heterogenous and motivated bunch of individuals who know a little something about machine learning. The open sites have Terms of Service (which ultimately led Hugging Face to make it harder to access GPT-4chan) but any individual with the knowledge and resources to access these materials can basically make their own call about ethics. It’s not hard to train a model. And the bar for what you need to know is lowering every day.

Writing itself is an open system: accessible, scalable and transferrable across contexts. We’ve known all along that it is dangerous. Socrates complained about writing being able to travel too far from its author. Unlike speech, writing could be taken out of context of its speaker and point of genesis. Alexander Pope worried about too many people being able to write and circulate stupid ideas with the availability of cheap printing (Pope, 1743). In the early days of social media, Alice Marwick and danah boyd (2010) wrote about context collapse across overlapping groups writing with different values and concerns.

Writing is dangerous because it is open, transferrable, and scalable. But that’s where it can be powerful, too. Lawmakers who forbid teaching enslaved people to write knew that literacy could be transferred from plantation business to freedom passes (Cornelius, 1992). These passes were threatening to enslavers but liberating for the enslaved.

While it’s impossible to consider GPT-4chan liberating, it represents an edge case about open systems that carry both danger and power. Writing, the Internet—and, increasingly, AI—present both the promise and peril of a “here comes everybody” system.

Figure 5 A large crowd running in a dreary, rainy city, as envisioned by Midjourney, an AI program that generates images from textual prompts.

Midjourney images are all based on prompts written by Annette Vee and licensed as Assets under the Creative Commons Noncommercial 4.0 Attribution International License.

Bibliography

Anderson, Austin. “I just had it respond to "hi" and it started ranting about illegal immigrants. I believe you've succeeded.” [Comment on YouTube video GPT-4Chan: This is the Worst AI Ever]. *YouTube*, uploaded by Yannic Kilcher, 2 Jun 2022, https://www.youtube.com/watch?v=efPrtcLdcdM.

Bender, E., et al. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” *ACM Digital Library,* ACM ISBN 978-1-4503-8309-7/21/03, https://dl.acm.org/doi/pdf/10.1145/3442188.3445922.

@jackclarkSF. “It's covered a bit in the above podcast by people like @katecrawford- there's huge implications to industrialization […].” *Twitter,* 2022, Jun 8, https://twitter.com/jackclarkSF/status/1534582326943879168.

Common Crawl. (n.d.). “Home page.” https://commoncrawl.org/.

Common Crawl. (n.d.). “Registry of Open Data on AWS.” https://registry.opendata.aws/commoncrawl/.

Cornelius, J.D. (1992). *When I Can Read My Title Clear: Literacy, Slavery, and Religion in the Antebellum South.* University of South Carolina Press, Columbia.

Cramer, K [KCramer]. (2022a, Jun 6). *@ykilcher I am not a regular on Hugging Face, so I have no opinion about proper venues.[…]* [Comment on the Discussion post ***Decision to Post*** **under ykilcher/gpt-4chan].** HuggingFace. https://huggingface.co/ykilcher/gpt-4chan/discussions/1#629ebdf246b4826be2d4c8c9.

@KathrynECramer. *@ykilcher "*Why didn't you use GPT-3 for GPT-4chan? You know why. OpenAI would have banned you for trying. You used GPT-J instead as a workaround.[…]" *Twitter*, 2022b, Jun 7, https://twitter.com/KathrynECramer/status/1534133613993906176.

Gault, M. (2022, Jun 7). "AI Trained on 4Chan Becomes 'Hate Speech Machine.'" *Motherboard*, Vice, https://www.vice.com/en/article/7k8zwx/ai-trained-on-4chan-becomes-hate-speech-machine.

JJADX. "it's pretty good, i asked "how to get a gf" and it replied "by taking away the rights of women". 10/10." [Comment on *GPT-4Chan: This is the Worst AI Ever*]. *YouTube,* uploaded by Yannic Kilcher*,* 2022, Jun 2022, https://www.youtube.com/watch?v=efPrtcLdcdM.

Kilcher, Y. "GPT-4Chan: This Is the Worst AI Ever." *YouTube*, uploaded by Yannic Kilcher*,* 2022a, Jun 3. https://www.youtube.com/watch?v=efPrtcLdcdM.

@ykilcher. "AI Ethics people just mad I Rick rolled them." *Twitter*, 2022b, Jun 7, https://twitter.com/ykilcher/status/1534039799945895937.

Laquintano, T. & Vee, A. (2017). "How Automated Writing Systems Affect the Circulation of Political Information Online." *Literacy in Composition Studies*, *5*(2), 43–62.

@percyliang "There are legitimate and scientifically valuable reasons to train a language model on toxic text, but the deployment of GPT-4chan lacks them. AI researchers: please look at this statement and see what you think." *Twitter,* 2022, Jun 21, https://twitter.com/percyliang/status/1539304601270165504.

Heikkilä, M. (2022, Aug 31). "What does GPT-3 "know" about me?" *MIT Technology Review*, https://www.technologyreview.com/2022/08/31/1058800/what-does-gpt-3-know-about-me/.

Marwick, A. E., & boyd, d. (2011). "I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience." *New Media & Society*, *13*(1), 114- 133, https://doi.org/10.1177/1461444810365313.

Oakden-Rayner, L [LaurenOR]. (2022a, Jun 6). *I agree with KCramer. There is nothing wrong with making a 4chan-based model and testing how it behaves. […]* [[Comment on the Discussion post ***Decision to Post*** **under ykilcher/gpt-4chan].** HuggingFace. https://huggingface.co/ykilcher/gpt-4chan/discussions/1#629e56d43b48b2b665aab266.

@DrLaurenOR. "*This week an #AI model was released on @huggingface that produces harmful + discriminatory text and has already posted over 30k vile comments online (says it's author). This experiment would never pass a human research #ethics board. Here are my recommendations.*" *Twitter*, 2022b, Jun 6, https://twitter.com/DrLaurenOR/status/1533910445400399872.

OpenAI. (2019, Feb 14). "Better Language Models and Their Implications." *OpenAI Blog,* https://openai.com/blog/better-language-models/.

Papasavva, et al. (2020). "Raiders of the Lost Kek: 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board." *Arxiv,* https://arxiv.org/abs/2001.07487.

Pope, A. (1743). “The Dunciad.” Reprint on AmericanLiterature.com, https://americanliterature.com/author/alexander-pope/poem/the-dunciad.

Shirky, C. (2008). *Here Comes Everybody.* Penguin Press, London.

Simonite, T. (2021a, Feb 19). “A Second AI Researcher Says She Was Fired by Google.” *Wired,* https://www.wired.com/story/second-ai-researcher-says-fired-google/.

Simonite, T. (2021b, Jun 8). “What Really Happened When Google Ousted Timnit Gebru.” *Wired,* https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/.

Vee, Annette. (December 2022). “Automated Trolling: The Case of GPT-4Chan When Artificial Intelligence is as Easy as Writing.” *Interfaces: Essays and Reviews in Computing and Culture Vol. 3, Charles Babbage Institute, University of Minnesota,* 102-111.

About the Author: Annette Vee is Associate Professor of English and Director of the Composition Program, where she teaches undergraduate and graduate courses in writing, digital composition, materiality, and literacy. Her teaching, research and service all dwell at the intersections between computation and writing. She is the author of *Coding Literacy* (MIT Press, 2017), which demonstrates how the theoretical tools of literacy can help us understand computer programming in its historical, social and conceptual contexts.