ENGCMP450

Public-Facing Research Essay

Dream On

Image: "Metamorphosis of Narcissus" by Salvador Dali, 1937

Earlier this year, OpenAI's ChatGPT made the rounds on social media and news outlets due to an incident where it went "off the rails" and spewed garbled streams-of-consciousness at its users. In one instance, a request for synonyms to the word "overgrown" resulted in a bizarre breakdown where it replied that a synonym for "overgrown" was… "overgrown", a fact it repeated over thirty times before crashing. A prompt for a picture of a dog sitting on a field yielded—cushioned and overlooking a vast field of grain amidst a gorgeous sunset backdrop—a fluffy cat. In response to one of its more dreamlike barrages declaring, "Fired of the photo-setting waves, nestling product muy deeply as though a nanna under an admin-color sombreret", one user asked, "*are you having a stroke?*"

These stories are examples of *AI hallucinations*, the propensity of certain AIs to proffer half-truths, misinformation, and in cases such as the one above, outright gibberish in response to user questions. AI hallucinations have grown larger in scope and visibility with the development of artificial intelligence and, by extension, large language models (LLMs). With the use of LLMs now becoming more widespread

and accessible to the average person, the issue has crossed the boundaries of the tech world and spilled over into the mainstream—so much so, in fact, that Dictionary.com's word of the year in 2023, "hallucinate", was deliberately chosen to highlight this phenomenon.

How exactly does a machine "hallucinate"? Hallucinations of the sort seen in ChatGPT's mental lapse emerge from the mathematical patterns in its training data. When an AI model generates text or images, it's essentially playing an elaborate probability game, predicting what should come next based on patterns it has observed. Sometimes, these predictions venture into territories that seem logical to the AI but have no basis in reality. To understand this process more deeply, imagine a vast web of interconnected words and concepts, each connected by threads of varying probabilistic strengths. When ChatGPT responds to a prompt, it traverses this web, making choices as to which path to follow at each turn. OpenAI's own postmortem of the February incident revealed that these choices involve "randomly sampling words based in part on probabilities" where the AI's "language" consists of numbers that map to tokens—essentially, pieces of words or concepts. When these mappings go awry, as they did during the incident, the result is a kind of mathematical vertigo where the AI loses its grip on the coherent patterns it usually follows.

AI hallucinations are, to be sure, an indictment of the tech industry, a humbling glimpse into the numbers and wiring lying within machinery purported to be accelerants of human development by esteemed tech houses the likes of Google and Meta. Championed as the next step in technological innovation, the cracks in artificial intelligence are exemplified by stories such as X's Grok AI falsely accusing famed NBA star Klay Thompson of "vandalizing multiple houses with bricks" during a dismal performance against the Sacramento Kings in April (although Thompson's career has stagnated somewhat since his move to the Dallas Mavericks at the time of this writing, it has not yet reached the point of throwing bricks through people's windows). Yet in these moments of mechanical delirium, we might recognize something oddly familiar; after all, machines aren't the only ones who sometimes see things that aren't there.

In the late 19th century, German psychiatrist Emil Kraepelin coined the term *dream speech* to describe the peculiar, fragmented, and often nonsensical utterances of people in the throes of dreams. Kraepelin noted how dream speech, though bizarre and seemingly illogical, often carries traces of coherent thought, emotional expression, or fragmented memories pieced together by the unconscious mind. In its own way, this "dream language" mirrors the logic (or rather, illogic) of hallucinations in artificial intelligence—a chaotic yet structured attempt to piece together fragments of its training data into a narrative that *feels* real, at least to the machine. Hallucinations generate outputs that, while incomprehensible to us, might be internally consistent with an artificial intelligence's probabilistic framework and logic system. Could it be, then that these moments of AI hallucination represent a primitive form of something similar? A rudimentary form of dreaming, or maybe an emergent property of a system designed to mimic human language and thought?

Where human dreams are shaped by our experiences, desires, and fears, AI "dreams" are shaped by the data they consume. A hallucinating language model is, in a sense, grasping for meaning in the spaces between its training data, filling the gaps with leaps of logic that, while flawed, evoke a kind of creativity. Through this perspective, AI hallucinations can be interpreted as a form of nascent imagination as opposed to pure malfunction. Just as Kraepelin saw dream speech as a window into the workings of the

unconscious mind, we might similarly view AI hallucinations as a glimpse into the mechanics of machine "thought", or as vestiges of humanlike cognition.

Central to this parallel between AI hallucinations and human dreams is our persistent tendency to humanize machines. Adorning robots with facial features or programming AI assistants to greet us with warm, human-like voices, are behaviors that, while unnecessary (or perhaps even irrational), reveal a deep-seated tendency to ascribe human traits, emotions, and even failings to our creations. Consider the user's baffled response to the hallucinating AI: "*Are you having a stroke?*". While the notion of a machine experiencing a medical emergency is patently absurd, this human impulse to frame the AI's behavior in medical terms reveals how naturally we reach for human analogies to make sense of mechanical aberrations. Such anthropomorphization isn't merely incidental—it underscores an essential truth about artificial intelligence: these machines are, in a very real sense, humanity's technological progeny. They are our creations, deployments of human ingenuity and, by extension, of ourselves.

Mechanical hallucinations aren't just *analogous* to human dreams, but are also themselves *essentially* human. If AI hallucinations are dreams, then they are *our* dreams, shaped by the vast amounts of humangenerated data used to train these models. They reflect our knowledge, our biases, and even our creativity, albeit filtered through a mechanical and alien lens. The surreal absurdity of an AI mistaking a cat for a dog or falsely accusing famed NBA star Klay Thompson of a rather serious misdemeanor is, at its core, an echo of human fallibility. Even the biblical house of Cornelius once spoke in garbled tongues.

Through its hallucinations, artificial intelligence holds up a mirror to humanity that simultaneously reflects and distorts what is familiar to us. In teaching machines to think like humans, we've inadvertently taught them to dream like us, to hallucinate like us. ChatGPT's raving dream-speech proclamations of "admin-color sombrerets", whilst objectively a failing stemming from a mathematical hiccup, also speaks to that strange, unknowable space between logic and imagination common to both machine and man. It is presumably the same space we draw from during brain-fogged moments where we spew "word salad"s or misremember facts from an old article about an NBA star, or in drunken stupors when we think quietly to ourselves that overgrown is overgrown is overgrown is overgrown.