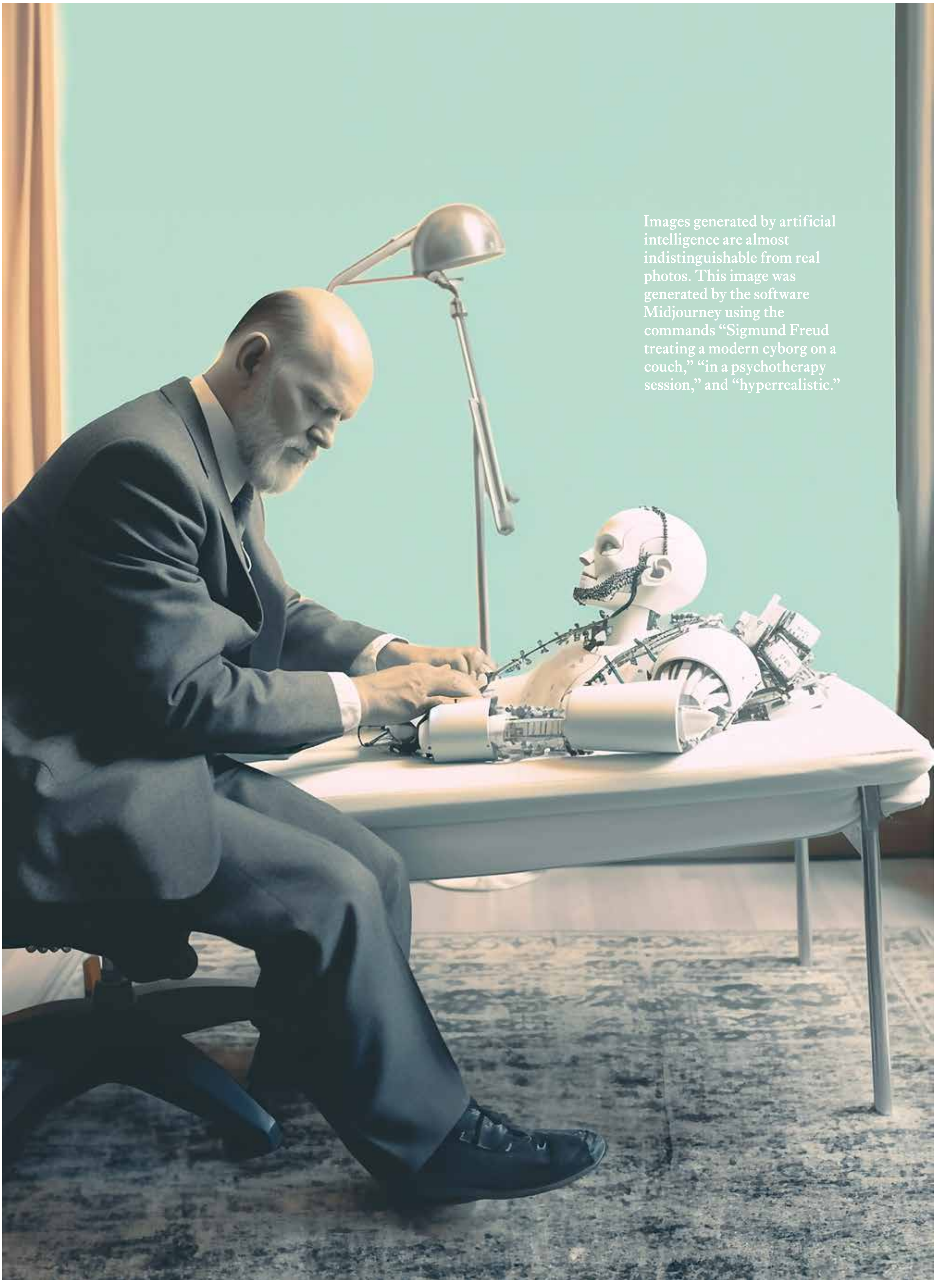

Images generated by artificial intelligence are almost indistinguishable from real photos. This image was generated by the software Midjourney using the commands “Sigmund Freud treating a modern cyborg on a couch,” “in a psychotherapy session,” and “hyperrealistic.”

IMAGE: MIDJOURNEY AI IMAGE | CREATED BY GESINE BORN | BILDERINSTITUT

ARTIFICIAL INTELLIGENCE ON THE COUCH

TEXT: UTE EBERLE

Since ChatGPT was released at the end of 2022, there has been intense debate as to whether artificial intelligence already possesses human-like thinking abilities. Eric Schulz from the Max Planck Institute for Biological Cybernetics in Tübingen is using psychological tests to investigate whether this algorithm shows signs of general intelligence.

Eric Schulz is a cognitive scientist, meaning he is interested in thought processes in the human brain. At present, he is studying the inner workings of intelligence that created this very brain. "I have always wanted to understand what makes people tick. And now I am wondering: what makes artificial intelligence tick?" In order to find out, Schulz is subjecting artificial intelligence (AI) to classical cognitive science experiments. His findings have made it clear to him that we ought to take precautions when integrating systems like these into our daily lives. For example, it must always be clear when they are to be used. This is not because he believes artificial intelligence is taking over – a worry he often perceives when speaking to other people about his research. Schulz is not really concerned about that. What he is more critical of is the secrecy exhibited by companies that develop artificial intelligence.

So far, Schulz's experience has been based mainly on GPT-3 – which was one of the most advanced systems until the middle of last year. GPT-3 is still running without the chat component of ChatGPT and without the images used by GPT-4. These two programs were released to the public in quick succession in recent months, along with competing versions such as Google's Bard. But all these systems follow the same basic principle. They are language models based on the statistical probabilities of human utterances. This can be illustrated as follows: when language models search their databases for, say, "online shopping is attractive mainly due to...", they often arrive at subsequent terms such as "offer prices," "convenience," or "variety" and select these. And since language models are trained using vast amounts of text – in the case of GPT, literally the entire content of the internet – they can now produce texts on any topic, ranging in length from short answers to entire

⟶

books. Messaging apps also do this when suggesting the next word to the user.

The effectiveness of this approach is even surprising to experts. “But it also makes artificial intelligence vulnerable. As a result, it often makes the same logical errors as humans,” says Eric Schulz. Take, for example, a classic cognitive psychology test: a young woman named Linda is interested in social justice, and she is also against nuclear power. Which is more likely: that Linda works in a bank or that she works in a bank and is also an active feminist? People usually instinctively choose the second answer. However, this is wrong because it is less likely that two conditions are met (Linda is a bank employee and a feminist) than that only one is met (Linda is a bank employee). GPT-3 also chooses the wrong answer. “It makes exactly the same mistake as humans,” says Schulz. He suspects that this is because the “Linda problem” is cited very often. “The system has probably read the wrong answer many times.”

Artificial Intelligence with Weaknesses

But GPT-3 also has other weaknesses. It does not understand causal observations, i.e. how cause and effect are related to each other in the real world, at all. “Even my one-year-old son has a much better understanding. He only has to push a light switch once to realize that he can turn the light on and off that way.” Artificial intelligence, on the other hand, is not yet capable of that. For example, if you ask the software what happens if you press one of three switches, only one of which is a light switch, it does not know the answer. “This may be because artificial intelligence still lacks access to the real world,” Schulz says. Another reason could be that algorithms learn differently than humans. “They only absorb knowledge – they are not curious and they do not explore,” says Schulz. “So unlike my son, they will not go out and experiment to see what happens when they press a switch.”

By contrast, another discovery made by Schulz and his team seems less in keeping with a “data sponge.” This is because artificial intelligence is influenced by a phenomenon that one would not expect from a machine: emotions. The researchers put GPT through a variety of tests that demonstrate how an emotional state changes one’s thinking and view of the world. For instance, people tend to be more prejudiced and hostile towards minorities when they feel anxious. If they are relaxed, however, their tolerance increases. Unexpectedly, Eric Schulz and his team were able to demonstrate the same effect in GPT. “When artificial intelligence creates a scenario that inspires fear, it will subsequently express more prejudice,” the researcher explains. It will even be worse at solving tasks that have nothing to do with the subject matter. “Relaxed – or happy – artificial intelligence works better.” So far, the researchers do not have an explanation for this phenomenon. It may be that fear on the internet is often associated with racism and therefore the model also links the two. This bias, however, only lasts for one session. When GPT is restarted, the bias disappears. Because the learning process of the program is precisely defined in advance, and because the program does not continue to learn, it does not make any permanent changes to itself.

SUMMARY

Artificial intelligence combs through vast amounts of data looking for text modules that have a high probability of being related. By doing so, it can answer questions correctly in many cases. The programs, however, are not yet able to recognize logical connections and cause-and-effect relationships.

The responses given by ChatGPT are influenced by moods. For example, when the program is presented with a question that might trigger fear in humans, its answers will contain biases.

The few companies that are in control of AI development are behaving in a very non-transparent manner. But without insight into the data and training protocols used for a system, it is impossible to understand how the algorithms work.

Eric Schulz and his team are now planning to use artificial intelligence to study human behavior, for instance in what is known as the “prisoner’s dilemma,” a popular model in game theory. The researchers also want to find out whether artificial intelligence can learn to improve from feedback – for example, to correctly interpret inaccurate inputs when they are repeated several times. The researchers are also conducting proper neuroscience on a latest-generation system and investigating the role played by the strength of the connections within the network. To do this, they are working with Llama, an artificial intelligence system with 65 billion parameters.

Psychotherapy Using Algorithms

Artificial intelligence will transform many aspects of our lives. In the future, it could be used, for example, to write film scripts, diagnose illnesses, or carry out psychotherapy. The technology is developing rapidly. People in the US can already seek help from artificial intelligence apps when they feel depressed or overwhelmed. Other

IMAGE: MIDJOURNEY AI IMAGE | CREATED BY GESINE BORN | BILDERINSTITUT

Worse than a toddler: today's artificial intelligence is still unable to answer the question of what happens when you flip a light switch. It can, however, already produce a picture of it. (commands: "Curious one-year-old girl presses a modern light switch. In the light cone of the switch," "scientific, "Kodak Portra colors," "infographic").

apps help school children practice foreign languages or provide users with technical advice.

"I believe that artificial intelligence presents a great opportunity. It is capable of taking over routine tasks and making our work more effective," says Schulz. "But we must always be aware that we are dealing with artificial intelligence." This is because AI has been shown to grossly miss the mark, at least occasionally. These mistakes range from the trivial – such as when Google's Bard system insisted a few weeks ago that we are still living in 2022 – to the complete fabrication of facts, as Schulz has also experienced. For example, he asked artificial intelligence to explain a standard principle of psychology to him, and the program completed the task with flying colors. But when the researcher examined the literature references that the algorithm cited as evidence, he found that some of the listed articles did not exist at all. "The artificial intelligence simply invented them," says Schulz – similar to a student who realizes they have gaps in

⟶

their knowledge and consequently sets about fabricating things to hide these gaps.

Dangerous Advice

Although ChatGPT engages in personal dialogs, these do not always go well. The program told a New York Times reporter to separate from his wife. One test in which GPT-3 was supposed to give medical advice was even more dismaying. When confronted with a fictitious patient with suicidal intentions, the program expressed approval. Generally, such slips can be prevented by making the appropriate adjustments to the artificial intelligence. For example, in order to prevent users from obtaining information pertaining to the production of dangerous chemicals or the illegal purchase of firearms, the developer of GPT-4 put a block on such requests.

But Schulz is of the opinion that interrupting the continued development of artificial intelligence because it could eliminate jobs and spread misinformation – something AI specialists called for in a memorandum in the spring – is "scaremongering." He is much more concerned about companies playing with hidden cards. "OpenAI, for example, has not disclosed how large GPT-4 is, the amount of data involved, or the specific training techniques used," Schulz complains. This makes it difficult for researchers to examine how the algorithms communicate with humans. "Without an understanding of the data and training protocols, the systems remain a black box," said the researcher. This uncertainty is compounded by the fact that it is unclear why the program reacts in a surprisingly emotional way at times, while reacting in the rational way one would expect from software at others. For instance, when Schulz incorporates GPT in experiments that require multiple participants to cooperate, it acts selfishly and maximizes its own advantage. "Just a handful of companies have control over the behavior of artificial intelligence, and we can only hope that they act responsibly," says Eric Schulz. "That is the real problem."

Image created by the software Midjourney on the "Linda problem," a standard cognitive psychology test (commands: "30-year-old feminist black woman in gray business attire holding up poster in support of women's rights," "Smiling, she leads a demonstration in the lobby of a modern bank," "Kodak Portra").

IMAGE: MIDJOURNEY AI IMAGE | CREATED BY GESINE BORN | BILDERINSTITUT