TECHNOLOGY
SCIENCE JOURNAL FOR KIDS

MARCH 2024

# How can we tell whether we are talking to a computer or a person?

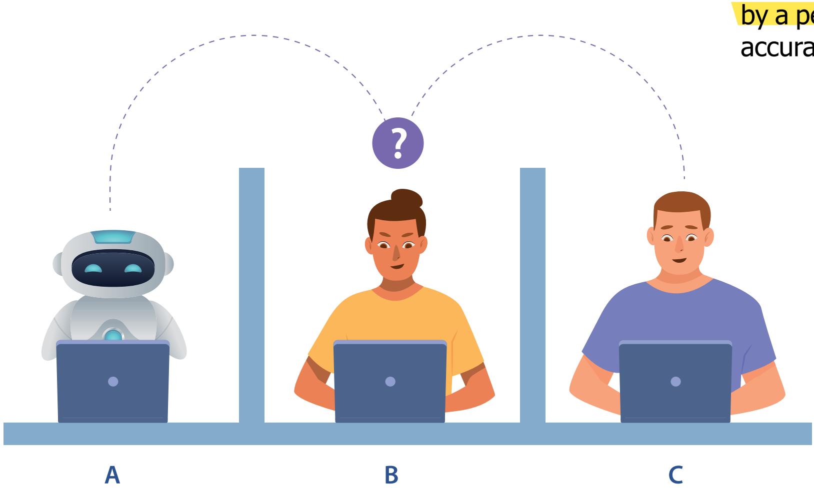
**Authors:**
Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman
**Associate Editors:**
Daniel Watkins and Fiona Firth

## Abstract

How can you tell if you are talking to a computer? New computer programs called language models have gotten very good at mimicking people. It can be really hard to tell if you're talking to a person or a computer. We wanted to know how people try to recognize computer-generated text and if they could do it accurately. We learned that people unconsciously use rules of thumb to figure out whether they are talking to a computer. These are often wrong, which means that people are vulnerable to scams.

## Introduction

Can a machine think? Philosophers have been trying to answer this question for hundreds of years. It's surprisingly tricky to come up with a good answer! Alan Turing was a **computer scientist** who thought a lot about **artificial intelligence (AI)**. He had an idea for a different way to test intelligence. Instead of measuring intelligence directly, what if you measured intelligent **behavior**? Can a computer answer questions in a human-like way? If a computer can respond to you so well that you can't tell whether it is a computer or a person, we say that it passes the **Turing test**.




Recently, computer programs called **language models** have gotten much better at mimicking human language. These AI programs analyze millions of sentences from books and websites to learn hidden patterns in language. It's exciting to see AI generating realistic sentences. But there are also problems with AI-generated text. People have used AI-generated text to cheat on tests. **Chatbots** can give harmful advice. And scammers can use AI to generate fake information so they can trick people.

How do people decide whether something was really written by a person or if it was written by an AI? And can they do this accurately? That's what we wanted to find out!

The Imitation Game (aka the Turing test) is a game for three players. Player A is a computer, and players B and C are people. All three players are in separate rooms. Player B's goal is to figure out whether Player A or C is a computer. They can ask questions to both A and C, who write down their answers and pass them back. If the computer can make Player B believe it's human, then the computer passes the Turing test.

# Methods

We recruited 4,600 participants for the study and came up with a simple version of the Turing test. We told participants that they were using a website where some people wrote their own **bios**, and some people used AI to write their bios. Each participant rated 16 texts, half of which were AI-generated. They then rated each text on a 5-point scale from "definitely AI" to "definitely human".

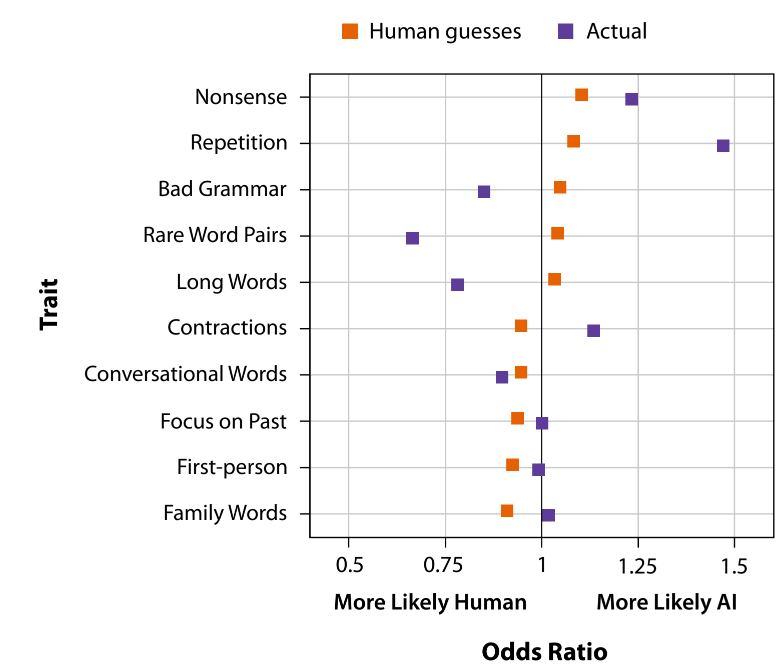
When they were halfway through the task, we asked participants for more details on one of the decisions they made. Two researchers read these responses to see what they had in common.

To create the bios that participants rated we gathered over 100 thousand real examples of bio text from the Internet. We used three scenarios: vacation rentals, job applications, and online dating. We chose these because people have to decide whether to trust someone based on how they describe themselves online.

We trained AI models on the full data set. Then we chose 3,500 examples at random of the real text and used the AI models to generate the same number of example texts.

We asked a different set of people to judge whether the bio texts were nonsensical, had bad grammar, or were repetitive. We also used algorithms to analyze other aspects of the bio text, such as tense or emotionality.

Finally, we used statistics and machine learning methods to detect patterns in whether participants thought a bio was written by AI.




If the true odds ratio and the odds ratio from humans are on the same side of the center line, we say the answers are aligned. Which traits were not aligned?

**Figure 1:**
The figure shows the odds that a piece of text with a specific trait is human or AI. Each of the traits is a rule of thumb that people used to try and figure out if an AI wrote the text. If the odds ratio is 1, then text with that trait is equally likely to be human or AI. If it is larger than 1, it is more likely to be AI. The purple squares show the actual odds from the data, while the orange squares show the average answer from participants.

# Results

Participants rated text as likely or definitely human written in 53.8% of cases. The rest was split between not sure, likely AI, and definitely AI. So, people are slightly biased toward thinking that text is written by people. Participants were largely unable to identify the AI-generated bios. Their answers were wrong about 49% of the time, which is not much better than chance. Even though people were wrong a lot of the time, their choices weren't random. If a participant rated a text one way, others often rated it the same way.

We found that people were following a few **rules of thumb**. When people saw nonsensical or repetitive text, they were more likely to say it came from an AI. And it's true, AI text was more likely to have these traits. But some rules of thumb gave wrong answers! You can see this in Figure 1. People thought bad grammar and unusual word choices came from AI, but they were actually more likely to come from human-generated text. And when people saw contractions (like "don't" or "won't") they thought it was a sign of a human writer, but actually, AI was more likely to use such words.

The rules of thumb people followed not only made them often give the wrong answer, but it made their answers predictable. We showed that AI can use people's rules of thumb to produce bios that people think are more human than real human bios.

## Discussion

People use rules of thumb for all sorts of things. It can be helpful to have a simple, fast way of answering a question, even if it's not a perfect answer. But it can be dangerous to have a rule of thumb that gives the wrong answer! Our results show that there are patterns in text that make people think the writer is human. AI developers could change their programs to focus on these patterns. This could make the text feel more human.

Why might this be dangerous? A common kind of scam is called **phishing**. A criminal sends emails to people pretending to be someone trustworthy. The email might say "Hi, I work for your bank! You need to log in and fix something in your account." With AI, the criminal could make lots of versions of the email to try to make one that's very believable.

We weren't sure whether it was a good idea to share our results. What if someone read our study and used our results to make a more convincing scam? In the end, we decided it was best to share our findings so that computer security researchers could learn from them.

## Conclusion

It's important to be careful on the Internet. There is a lot of useful information online, and a lot of fun things, too. But there are also a lot of ways to get in trouble. Things like your birthday, address, or family phone number should not be shared online where other people can find them. The questions you ask a chatbot are not private. And often things aren't what they say they are. We can never know who is on the other side when talking to someone on the Internet – it may not even be a human! That's why it's important to not just rely on your gut feeling, but check the website URL and author information. If something seems strange, tell a responsible adult about it!

## Glossary of Key Terms

**Artificial intelligence** - the idea of machines or computers that can think. The phrase is often used to describe computer programs that solve difficult problems, like translating languages, finding items in images, or understanding spoken words.

**Bio** - short for "biography"; a short description of a person.

**Chatbot** - an AI program that responds to written messages in a human-like way.

**Computer science** - the study of how computers work.

**Language model** - a computer program that looks for patterns in text and uses those patterns to create new text.

**Phishing** - a type of scam in which people send emails and try to trick people into giving up their passwords, personal information, or their money.

**Rule of thumb** - a simple method that people use to come to an answer quickly, even if it's not a perfect answer, such as using the width of your thumb to measure instead of using a ruler.

**Turing test** - A computer passes the Turing test if it can answer a person's questions well enough that the person can't tell if the answers came from a computer or from another person.

# Check your understanding

1. Why did Alan Turing suggest using the Imitation Game, also called the Turing test?

2. How would you try to tell a computer's answer from a person's answer? What are some clues you would look for that weren't mentioned in the article?

3. Rules of thumb can help us make decisions quickly. We found people guessed that incorrect grammar indicated the text was from AI. Many people also don't answer phone calls from numbers they don't recognize, since it's likely to be an unwanted call. What other examples of rules of thumb can you think of from your own life?

4. Why were the researchers worried about sharing their results?

5. There are many controversies about AI. For example, some artists are worried that people will use AI to make art, instead of hiring artists. Name something that you have heard might be scary about AI. What could be done to prevent that from happening?

## REFERENCES

Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman (2022) *Human heuristics for AI-generated language are flawed.* The Proceedings of the National Academy of Sciences.
**https://www.pnas.org/doi/10.1073/pnas.2208839120**

MIT Technology Review: Three ways AI chatbots are a security disaster
**https://www.technologyreview.com/2023/04/03/1070893/three-ways-ai-chatbots-are-a-security-disaster/**

The Register: AI-generated phishing emails just got much more convincing
**https://www.theregister.com/2023/01/11/gpt3_phishing_emails/**

**Acknowledgment:**
This article's adaptation was supported by the Akamai Foundation.