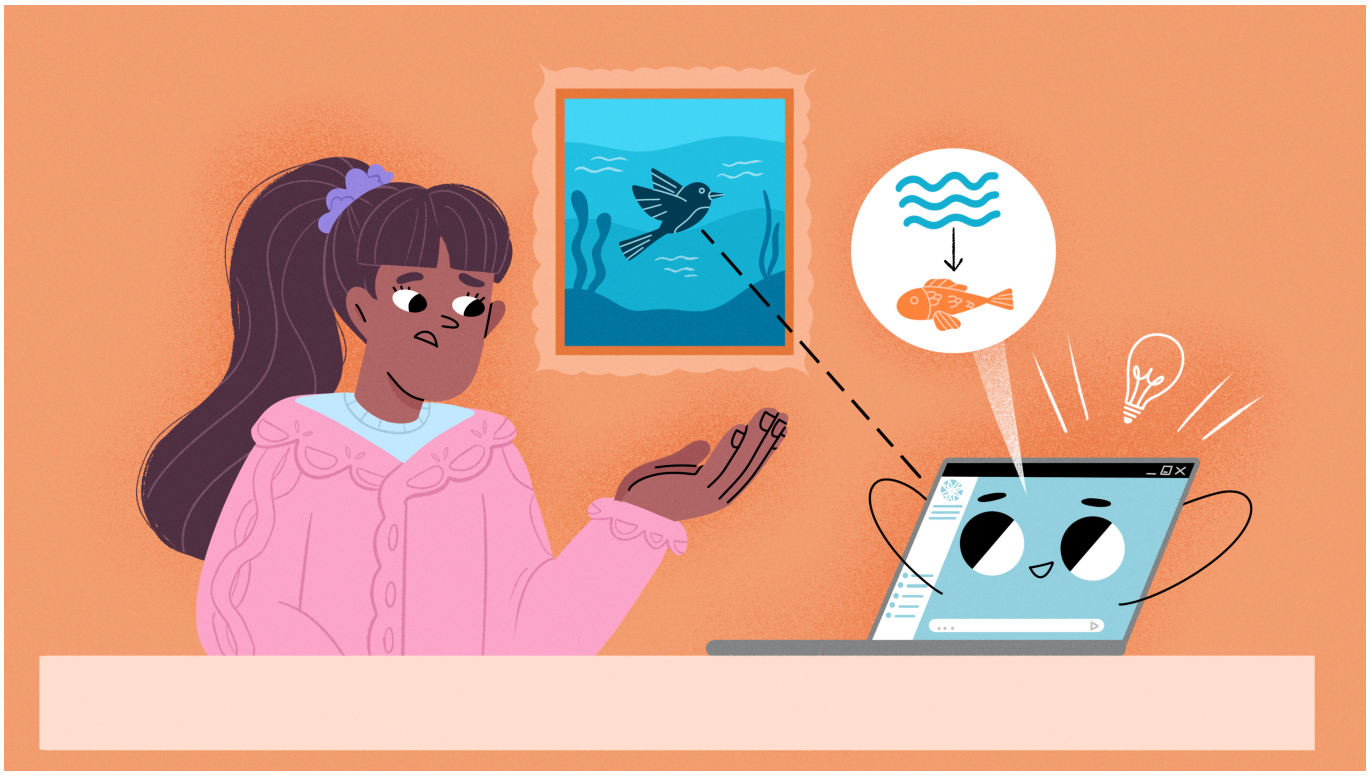
frontiers | Frontiers for Young Minds

NEUROSCIENCE AND PSYCHOLOGY
Published: 13 November 2025
doi: 10.3389/frym.2025.1683304



ARTIFICIAL INTELLIGENCE TAKES SHORTCUTS TOO

Yair Weiss*, Daniella Har-Shalom* and Ofir Shifman

The Edmond and Lily Safra Center for Brain Sciences (ELSC), School of Engineering and Computer Science, The Hebrew University of Jerusalem, Jerusalem, Israel

YOUNG REVIEWERS:

THE INTER DISCIPLINARY CAMPUS—TIRAT CARMEL

AGES: 13–14

Artificial intelligence (AI) systems are often praised for their impressive performance across a wide range of tasks. However, many of these successes hide a common problem: AI often takes shortcuts. Instead of truly learning how to do a task, it may just notice simple patterns in the examples it was given. For example, an AI trained to recognize animals in photos might rely on the background, instead of the animal itself. Sometimes these shortcuts can lead to serious mistakes, such as diagnosing based on hospital tags rather than patient data. These errors occur even in advanced systems trained on millions of examples. Understanding how and why AI takes shortcuts can help researchers design better training methods and avoid hidden failures. To make AI safer and more reliable, we must help it develop a true understanding of the task—not just guess based on patterns that worked in the past.

SHORTCUTS DO NOT ALWAYS WORK!

We have all heard the saying "there are no shortcuts in life". To truly succeed at a task, we must work hard and persevere, and it is not enough to find a trick to shorten the path. It appears that when artificial intelligence (AI) is taught to perform various tasks, it too takes shortcuts sometimes, and this can lead to surprising and even dangerous consequences. One of the significant challenges is to ensure that AI systems really learn to solve the task, and are not "tempted" to take dangerous shortcuts.

WHAT CONFUSES ARTIFICIAL INTELLIGENCE?

Let us start with a simple question: What do you see in the pictures in Figure 1A?

Figure 1

Images of camels and horses against different backgrounds of sand and grass.

You probably answered that you see a camel in the right picture and a horse in the left. For many years, computer scientists have tried to write programs that could answer similar questions, and they have only partially succeeded. The last decade has witnessed a significant improvement in the performance of AI, and today, computers can describe images with very high accuracy. Numerous websites claim to be able to describe any image using AI. When we uploaded these two pictures to one of the leading image description sites, it correctly identified the right photograph as a camel and the left as a horse.

Now let us move on to a slightly more difficult question. What do you see in the bottom pictures in Figure 1B?

You probably see a horse *and* a camel, against different backgrounds. Can you guess what description the AI provided? Surprisingly, it

Figure 2

(A) Each math exercise has two expressions. The task is to decide which side is greater. **(B)** Examples of correct answers marked in red, which could be used to train AI.

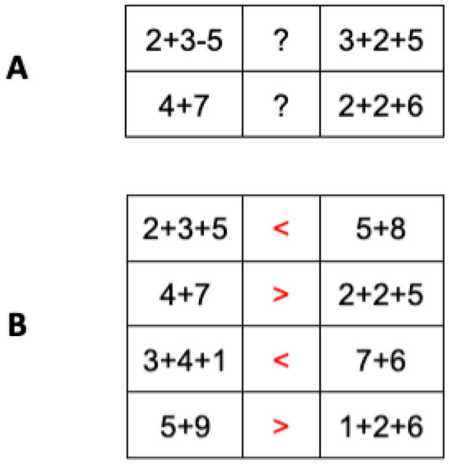


Figure 2

described the right photograph as two horses and the left as two camels! But it is clear to us that there is one camel and one horse in each picture—the exact same animals in both. This is an example of a problem with existing AI tools. While they often provide correct answers (the website we used has tens of thousands of users worldwide, and they report satisfaction with its performance), they sometimes make embarrassing mistakes for unclear reasons. In the rest of this article, we will explain one of the fundamental sources of errors in AI tools: their tendency to learn **shortcuts**.

SHORTCUT

A rule that returns correct answers on training images but is often wrong when circumstances change. For example, a diagnosis based on which machine was used instead of medical content.

MACHINE LEARNING

A method that allows a computer to learn to perform complex tasks based on examples, rather than having engineers program it in advance how to solve each problem.

DECISION RULE

Explicit instructions that define the output of a method for a given input.

WHAT ARE SHORTCUTS IN AI, AND WHY DO THEY HAPPEN?

Almost every AI tool in use today is based on a technology called **machine learning**. Machine learning creates AI by learning from examples. To help you understand better, we will use exercises that you may remember from elementary school math classes (Figure 2A).

Each exercise has two arithmetical expressions, and you have to write down whether the result on the left side is greater (>) or smaller (<) than the result on the left side. If we were to ask a programmer to write a program that solves the exercise, they would probably write code that calculates the expression on the right and the expression on the left separately, and then compares them. This is how programs would work about 20 years ago—engineers would think of a solution and then write code for the AI to implement it.

The modern approach is to simply present AI with numerous example questions, along with the correct answers to each question, and let it find a **decision rule** on its own. Once it has learned the rule, the AI can apply it to new examples for which it does not have an answer. This is the machine learning approach that has enabled

Figure 3

Training examples. Six photographs of horses and six photographs of camels, along with the name of each animal.

Figure 3

the dramatic development in AI performance in recent years. But this approach has a problem, which we will demonstrate with the arithmetic exercises.

Consider Figure 2B. Suppose we give AI four exercises with the correct answer for each (marked in red). As you recall, AI is trying to infer a rule from the training exercises. The rule would be a statement like, "if a specific condition is met, the answer will be $>$; otherwise, the answer will be $<$". Let us take a few minutes to consider what simple decision rule can be inferred from the calculation training images, to give the correct answer.

What if we consider the following rule: if there are three numbers on the left side, then then the left side will always be smaller ($<$), otherwise the correct answer is $>$. This is a straightforward rule for both a person and an AI. It is undoubtedly easier than calculating the results of all the expressions. However, even if it provides the correct answer for all the training images in Figure 2B, it will inevitably be incorrect in many other examples. This is a shortcut—a superficial solution that is not based on real understanding.

TRAINING IMAGES

A collection of examples presented to AI during learning. For example, pictures labeled with which animal they contain, or medical images of sick or healthy patients labeled accordingly.

Back to camels and horses. Now, suppose we show the AI **training images**—lots of pictures with descriptions, as shown in Figure 3. Can you think of a shortcut to correctly describe each example as either a horse or a camel?

You might have considered the following shortcut: if the background is a desert, then the photograph features a camel; and if the background is grass, then the picture shows a horse. This is a shortcut that solves all the training images perfectly. Still, it will not solve other

examples correctly, of course. It is possible that the AI tool from the internet, which we used in the introduction, learned a similar shortcut and therefore made a mistake on the images in which we changed the background.

It turns out that many successful versions of AI for describing images are tempted to learn similar shortcuts. A study published in 2022 evaluated several cutting-edge AI tools for image description [1]. The study found that, when asked to describe any image from the internet, the tools were correct over 80% of the time. But when the object was moved to a different background, they were accurate only 35% of the time. Apparently, the shortcut that identifies an object in an image based on its background is especially tempting!

PROBLEMS WITH AI SHORTCUTS

One study compiled a lot of evidence for shortcuts of this type [2]. In one example, AI learned to identify pneumonia from chest X-rays, and based on this, to determine whether a person is sick or healthy. However, it turned out that the system did not analyze the lungs in the photo at all, but instead discovered that the hospital tag in the image could indicate whether the person was healthy or sick and focused on that. This shortcut is extremely dangerous: perhaps the tag was relevant in the training examples that the AI saw, but it is frightening to imagine what would happen if it were asked to examine another patient from a different hospital.

An article published in 2024 reported on its examination of 14 different AI tools designed for medical diagnosis and found that they use shortcuts extensively [3]. The tools were tested on a variety of tasks, including identifying coronavirus infection or an enlarged heart in X-ray images, diagnosing heart disorders by listening through a stethoscope, and more. The researchers hypothesized that the AI would not focus solely on medical information and would learn to exploit the unique measurement conditions for each case, allowing it to take shortcuts. For example, if a doctor suspects that a patient has coronavirus, they may send them for an X-ray using a different device than the one they would send a patient who appears healthy (for reasons of medical staff safety, for example). AI can identify the difference between the measurement devices and determine whether a person is sick or healthy based on the type of device used, without examining the image itself. Researchers found that all 14 tools tested employed such a shortcut, which was based on the equipment's characteristics rather than medical information. Once the tests were performed in a new hospital, the AI's performance decreased significantly because the shortcut no longer worked. This, of course, is a serious problem when trying to integrate AI into real-world medical situations.

HOW CAN WE IMPROVE FUTURE AI?

AI researchers are working to develop more effective machine learning methods that can prevent shortcuts. One way is to train AI with an increasing number of examples. Let us return to our discussion about math exercises: if, instead of giving AI only four training examples, we give it entire math textbooks, it is likely that the shortcut we saw will no longer work and AI will have to find another solution. However, even when the number of examples is vast, we cannot guarantee that the AI will not find another shortcut. In the study on image description that we mentioned, it was found that even an AI trained on hundreds of millions of images used the shortcut of determining the type of animal based on the background [1].

If we knew in advance which shortcut might be used, we could try to prevent it during the training phase. For example, suppose we do not want the AI to recognize animals by their backgrounds during training. In that case, we can show various images in which each animal is seen against a random background, such as a camel in the snow. As we add more such examples to the training, the shortcut that identifies the animal by its background will no longer work. But there is a problem here: this approach focuses on shortcuts that we already know. Even if we avoid this shortcut, we cannot prevent other shortcuts that we do not know of, for example, reliance on lighting conditions.

Humans and other animals (non-artificial intelligence) also tend to learn from shortcuts. For example, many students prepare for a math test by solving questions from previous tests, and sometimes give up trying to understand the material. At the beginning of the 20th century, a horse named Clever Hans became famous for supposedly being able to solve math problems. When asked how much $9 + 5$ was, he would drum his hoof 14 times. Psychologist Oskar Pfungst studied the horse's performance and concluded that it could "solve" the math problems using a shortcut. The horse would continue to drum until the questioner's body language "revealed" to him (without the questioner's knowledge) that he had reached the correct number. Once the horse was prevented from making eye contact with the questioner, the shortcut no longer worked, and Clever Hans' answers were almost always wrong.

Despite everything we have described, we should not lose hope. In recent years, we have seen AI surpass non-AI in many ways. For example, it defeated the world chess champion and scored higher than most humans on the bar exam. Today, AI also helps teachers and students with a variety of tasks, from searching for information to summarizing texts. AI researchers continue to develop machine learning methods that will significantly reduce its reliance on shortcuts. Perhaps one day they will surpass us in this regard as well.

ACKNOWLEDGMENT

We thank the Gatsby Foundation for financial support.

AI TOOL STATEMENT

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

REFERENCES

1. Chefer, H., Schwartz, I., and Wolf, L. 2022. Optimizing relevance maps of vision transformers improves robustness. *Adv. Neural. Inf. Process Syst.* 35:33618–32. doi: 10.48550/arXiv.2210.10817
2. Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., et al. 2020. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* 2:665–73. doi: 10.1038/s42256-020-00257-z
3. Ong Ly, C., Unnikrishnan, B., Tadic, T., Patel, T., Duhamel, J., Kandel, S., et al. 2024. Shortcut learning in medical AI hinders generalization: method for estimating AI model generalization without external data. *NPJ Digit. Med.* 7:124. doi: 10.1038/s41746-024-01118-4

SUBMITTED: 10 August 2025; **ACCEPTED:** 08 October 2025;
PUBLISHED ONLINE: 13 November 2025.

EDITOR: Idan Segev, Hebrew University of Jerusalem, Israel

SCIENCE MENTORS: Daria Patel Teplov

CITATION: Weiss Y, Har-Shalom D and Shifman O (2025) Artificial Intelligence Takes Shortcuts Too. Front. Young Minds 13:1683304. doi: 10.3389/frym.2025.1683304

CONFLICT OF INTEREST: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

YOUNG REVIEWERS

THE INTERDISCIPLINARY CAMPUS—TIRAT CARMEL, AGES: 13–14

We are a group of eighth-grade students who chose to take part in a year-long elective course based on *Frontiers for Young Minds*. In this class, we learned how to read and understand scientific articles, reviewed one ourselves, and created science posters on topics we are passionate about. The experience deepened our understanding of science, and we really enjoyed exploring the articles on the website.

AUTHORS

YAIR WEISS

I am a professor of computer science and hold the Dieter Schwarz Chair in Artificial Intelligence Research at the Hebrew University of Jerusalem. I live in Jerusalem, am married, and have four children: Itai (14), Adi (19), Roni (21), and Gili (24). In addition to being a researcher, I served for many years as a consultant to Mobileye, a technology company. I am also the owner of Hapoel Katamon, a Jerusalem soccer team. *yair.weiss@mail.huji.ac.il

DANIELLA HAR-SHALOM

I am a Ph.D. candidate in computer science at the Hebrew University of Jerusalem. I research how artificial intelligence "sees" the world, and how it does so in comparison to the human brain. Specifically, my research focuses on how artificial intelligence represents the things it sees, in its "head". Beyond research, I enjoy observing my daughter (Amalya, 2 years old) and learning how she sees the world. In my free time, I enjoy traveling and tending to my garden. *daniella.horan@mail.huji.ac.il

OFIR SHIFMAN

I research learning systems and applied artificial intelligence in industry. I studied for a bachelor's degree in cognition out of an interest in the human brain and a passion for understanding how people learn and think. Following my passion for engineering and creating technological systems, I completed a bachelor's and master's degree in computer science at the Hebrew University. What particularly fascinates me are the challenges of implementing learning machines in the real and complex world, as well as the gap between the learning processes of humans and machines. In addition to my research, I am a bicycle enthusiast, which allows me to avoid traffic jams, and I enjoy hiking in my free time.