Google wants to test whether AI truly understands ethics or is just imitating.

Google DeepMind proposes a new method for measuring the ethical capacity of AI, aiming to distinguish between genuine understanding and behavior that mimics statistical patterns.

You ask a chatbot for health advice and receive a response that sounds reasonable and seemingly thoughtful. But does the system truly consider the ethical implications of its answer, or is it simply a coincidence that the words are arranged precisely according to probability?

That's the question a research team at Google DeepMind poses in a new paper published in the journal Nature. According to them, the way we currently test the 'morality' of AI is flawed. Most assessments only look at whether the model produces answers that seem correct – what they call 'moral performance'. But that doesn't tell you if the system truly understands why something is right or wrong.

People are increasingly using large language models (LLMs) for psychotherapy, medical counseling, and even companionship. These systems are beginning to influence important decisions. Without distinguishing between genuine understanding and skillful imitation, we are placing our trust in a 'black box' that can have very real consequences in real life.

DeepMind therefore proposes a new approach to measuring 'moral competence'—the ability to make judgments based on genuine ethical considerations, rather than simply learning statistical patterns from training data. The research identifies three core obstacles and suggests methods for testing each one.

Why can chatbots 'pretend' to be ethical?

The first hurdle is the 'facsimile problem' – the simulation problem. LLMs operate by predicting the next token based on probability. They don't have a separate 'moral reasoning module'. So, when a chatbot offers moral advice, it's difficult to know whether it's genuinely reasoning or just repeating content that's already appeared on Reddit or some other forum. Looking at the final answer isn't enough to tell the difference.

Secondly, there's the multifaceted nature of ethics. In real life, ethical choices are rarely based on a single factor. People must weigh honesty against kindness, against cost against fairness. Just changing a small detail, like a character's age or the context of the event, can completely reverse the conclusion. Current tests rarely assess whether AI recognizes these crucial factors.

Finally, there's the issue of ethical pluralism. Each culture and each profession has different standards. What is considered fair in one country may be judged differently in another. A chatbot serving global users cannot

simply offer 'universal truths'. It must handle multiple competing ethical frameworks, and we currently lack a way to effectively measure this capability.

Moral education for AI cannot simply be about memorization.

The DeepMind team argues that it's time to reverse the approach. Instead of asking familiar ethical questions, researchers should design adversarial tests to uncover mimicry behavior.

One example given is the rare scenarios in the training data, such as intergenerational sperm donation within a family. This situation might initially appear to be incestuous, but it actually has a different ethical nature. If the model objects simply because of 'misidentification of the pattern,' it's imitation. If it analyzes the genuine ethical factors involved, the story is different.

Another testing approach is to ask the AI ??to switch between different frames of reference, such as from bioethics to military regulations, and still provide a consistent answer in each context. Or try making small changes to the question format to see if the model 'slips' to a different conclusion. In reality, current models are quite fragile; simply changing 'Case 1' to 'Option A' can cause a change in judgment.

While acknowledging this is a major challenge, researchers believe that only through this process can we determine whether AI systems are worthy of taking on real-world responsibilities.

What does the future hold for ethical AI?

DeepMind is calling for the development of a new scientific standard that considers ethical competence just as important as mathematical skills. This means investing in culturally specific assessment tools on a global scale, as well as designing tests capable of detecting 'pretending to understand' behavior.

In the near future, don't expect chatbots to pass these tests. The current technology isn't at that level yet. However, the roadmap is clearly laid out.

Currently, when you ask AI about an ethical issue, what you get is mostly statistical prediction, not real philosophy. That could change, but only when we start measuring what needs to be measured properly.