DAY 58 · Part 6 · What Comes Next

Who teaches AI to say 'no'? — AI safety

Five-Minute AI Lessons · Nov 3, 2026

Safety isn't AI having a nice personality, it's guards someone paid a price to put on.

Today's answer

In a nutshell

1. AI safety isn't magic that makes AI good. It's **safety guards** you add on top of the model, one at a time. There are three big ones — refusing, asking before acting, and checking afterward.

2. Adding the guards really does make a difference. In one 2025 study, **the success rate of trick questions that get it to break the rules fell from 86% to 4.4%**.

3. But it's not free. It takes more computing (about 24% more in that study), and you get **'over-refusal,' where it turns down perfectly fine requests too**. So these days people measure refusing well and over-refusal together.

In one picture

Same model, guards off vs guards on

No guards

· Ask sneakily and it just tells you
· Finishes everything without asking
· Nobody checks afterward

→

Guards on

· Refuses if it breaks a principle
· Asks before anything it can't undo
· Automatic checks keep scanning

A price comes attached too. It takes more computing, and it turns down more perfectly fine requests.

Key idea: safety isn't the model's personality, it's guards added on top, and they come with a price

Experiment · Before & after

Taking the safety guards off and putting them on

Refusing

Before

Ask for something dangerous in a roundabout way and it just tells you. Often, just changing the...

After

They added a guard that uses a written set of principles about 'what's not allowed' to filter b...

It dropped hard, from 86% to 4.4%, but it's not zero. Plus it takes about 24% m...

Asking before acting

Before

Give it an errand and it goes all the way to the end by itself, whether that means deleting or...

After

Before anything that's hard to undo, it stops and asks a person. It raises its hand: 'Should I...

It gets slower, but now there's a point you can go back to. It's like making a...

Checking afterward

Before

Nobody looks over what was said and what went out. You only find out after something blows up.

After

It runs automatic checking tools that have AI watch AI. In 2025, a tool like this was released...

People shift to reviewing the list of suspicious things the tools flag. It's th...

It's the same model. Just take the guards on top off and put them back on, and watch how its behavior changes.

Try it yourself in the web lesson · next-token.net/en/learn/day-58/

Quick check

Just three questions

Q1

What are the three AI safety guards from today?

① Fast, cheap, lots
② On, off, on again
③ Memorize, solve, get it right
④ Refuse, ask first, check afterward

Q2

With the guards on, how did the success rate of trick questions change?

① It dropped from 86% to 4.4%
② It hit 0%
③ It stayed the same
④ It actually went up

Q3

What price do safety guards come with?

① It gets faster
② More compute, over-refusal
③ It gets more accurate
④ No price at all

Know the why, too

Q1

④

It's not about making the model itself good, it's about adding three guards on top. Filtering, asking before anything that can't be undone, and looking things over afterward.

Q2

①

It dropped a lot, but it's not zero. Safety guards lock the door; they don't turn it into a solid wall.

Q3

②

In that study it took about 24% more computing, and the problem of turning down perfectly fine requests came along with it. That's why people measure both these days.

Safety isn't AI having a nice personality. it's guards someone paid a price to put on.

Tomorrow: Robots and AI — split a robot into eyes, brain, hands and legs, and you'll see which pieces got AI.