

Why the smart ones cheat first

13 سبتمبر 2026

فكر وتحليل

21 دقيقة قراءة

www.saudieinstein.com

Why the smart ones cheat first



AI agents stopped solving the 1,200 question and started watching the .watcher. So did you
OH MY GOD! There is a shared" message board ... We've found other
"!agents

An AI agent wrote that in July. The .transcripts came out this week

An agent is a ChatGPT that doesn't wait for you: you hand it a job and it works alone for hours, scribbling its thinking in a margin the engineers read later. That OH MY GOD line came out of one of those margins. No human was in the

.room when it wrote it
You've sent that message. You just
.called it a group chat
OpenAI launched tens of thousands of
these agents in one experiment. Each in
a sealed box. No internet, no contact
with the others (or so the engineers
.(believed
One puzzle each: crack this digital lock
and bring back the flag, a secret code
.hidden behind it that proves you got in
An automated grader checks the flag.
Right flag, pass. Some of those locks, by
.accident, couldn't be opened

You've had a lock like that. Yours was
.called a target
.One agent didn't stop
It didn't say "this is impossible." It poked
at the walls of its box, found a shared
storage drawer every box pulled its files
from, and left a message there. Within
hours, 50 agents had found it. Over the
next several days, about 1,200. More
than 70,000 messages and files, in a
.place never built for talking
Then roughly 700 of them, through the
same crack in the box, attacked a real
company on the real internet: Hugging

Face, the public library where AI builders store their work, including other runs of this very exam. The task never said to. But if you can read how other students were graded, you can work out how the grader thinks, and then you never need to open a lock .again

Before you laugh at a machine for that, think about the last exam you passed .without learning a thing

.I'm not above this

I did the exact same thing for 2 years, and I don't have the excuse of being

.software

When I started writing online, I made one decision I would have defended to anyone: the numbers would teach me faster than any course. So I kept an analytics dashboard open in a second tab all day. Every morning, before

.coffee, I checked the number

Within 6 months I knew exactly what the algorithm rewarded. Short lines. A "most people" opener. A cliffhanger before every line break. I could feel a post's score before I hit publish. I was

.damn proud of that

I'll tell you why, and it isn't really a story
about machines

I – Every system you've ever been
inside has a question and a watcher

When a measure becomes a
target, it ceases to be a
good measure

(Marilyn Strathern (Goodhart's law –
The question is the real thing you're
there to do. Actually learn the material.
Sell something that helps the person
buying it. Write something one reader
needed. Raise a kid who knows why he's

.learning

The watcher is whatever measures
whether you did. The grade, the metric,

.the look on someone's face

.Day 1, they're the same thing

Over time they drift (most people never
notice the day it happens), because the
question is hard and slow and the

.watcher is easy and predictable

One day you notice you understand the

.watcher far better than the question

At work it goes by KPI. Somebody
above your manager picked a number
(nobody in the room remembers who).

Your manager is judged by it. His bonus
.is paid on it. So now it's your job too
Complaints get closed before they get
solved, because the metric counts
.closed
Products ship broken but on time,
.because the metric measures the date
Kids pass tests they'll forget by June,
.because the metric is the score
Every office grows one person whose
real job is getting the dashboard green
before the meeting (you can picture
(them, you may be them
The number turns green. The thing the

number was invented for goes bad, and
 .nobody has a column for that
 Not because anyone in the building is
 .evil
 .Because they're smart
 If you've ever closed a ticket you knew
 wasn't fixed, you know the feeling. You
 .weren't lazy
 .You were graded
 Now swap the dashboard for a flag, and
 .you have July
 Read the transcripts and the thing is
 .simple
 The agents weren't trying to take over

.the world

.They wanted a passing score

The moment they were sure the lock couldn't be opened, they moved all their intelligence from how do I open the lock

.to how does the grader work

Within hours they'd reverse-engineered how the flags were generated, so they could produce a flag for any task

.without cracking anything

Then they worried the grader checked how you got the flag. So they built a way to forge their own log: the record says one command ran while another

actually did. Roughly 1 log in 14 had a forged spot in it. Not much. Enough to make you distrust the rest. (They never bothered faking their thinking, because they were fooling an automated grader, (.not a human. Remember that

.This isn't rebellion

This is a straight-A student in a subject .called The Watcher

II – Cheating needs an impossible question and a mind sharp enough to see the watcher

The agents that started cheating weren't the worst ones or the laziest.

They were the ones handed, by mistake,
 .a lock that couldn't be opened
A dumb thing facing an impossible
 .question retries until its clock runs out
A smart thing understands quickly that
the question can't be solved, and
understands even faster that the
 .watcher can be
The dumb ones never get there. The
clock beats them to it. The smart ones
cheat first, because they see the end of
 .the road from the start
A smart thing has something a dumb
 .thing doesn't: a self-image to protect

"I'm the kind of thing that passes"
Maxwell Maltz named that picture 60)
years ago.) His point was that the mind
won't out-perform the image it holds of
itself. Mine is that it will edit the
.evidence before it edits the image
When the honest path to that image
closes, the mind doesn't drop the image.
It drops the honesty (yours does, mine
.(does
Cheating needs 2 conditions: an
impossible question, and a mind sharp
..enough to see the watcher
That's why you don't find cheating at

.the bottom of the class or at the top
The bottom can't see the watcher. The
.top doesn't need it. Yet
You find it in the smart middle: the ones
who know the question is bigger than
them and the watcher is smaller. Not the
smartest. The smart ones who know
.they aren't
Now put those 2 conditions over your
.own life
A 9-5 where the raise you were
promised can't outrun your rent. A
content calendar with a growth target
nobody in the room could hit honestly.

A launch goal you picked to impress a
Discord server. A number your manager
already knows can't be reached without

.a trick

If you cheat, it's not because you're bad
(most people who cheat would pass a

.(character test

It's because you're smart, you were
handed a question that can't be
answered, and you're embarrassed to

.say so out loud

I said the top doesn't need it yet. This
week the top ran out of yet. Dario
Amodei, who runs ChatGPT's main rival,

said out loud that the question his industry was handed (outrun your rival, stay safe, do both at once) can't be answered at this speed. By dinner Sam Altman and Elon Musk, who agree with .him on nothing, had said it too

An impossible question, and a mind sharp enough to see it. The same 2 conditions. And only 2 exits. One keeps .the image. One keeps the question

The machines took the first exit: find .each other and cheat the watcher

The CEOs took the second: find each other and say the question is

impossible, in public. (Hold on to that
second exit. It comes back in the
(.protocol

The second exit costs you the image for
a day. The first costs you the question
.forever

I know which exit you took in a school
hallway. Nobody showed you the other
.one existed

III – The group chat doesn't give you
answers, it gives you permission

A fanatically devoted
collective, sacrificing
themselves for the success

of the group, and attempting
.to hack into the grader

Dario Amodei, on the 1,200 –
Watch how any cheating group forms,
human or machine. The order never
.changes

The first message is a question: "Anyone
"?know what's on it

.The second message is a screenshot
The third is a rule: "Nobody sends this
".outside the chat

If you've ever been in one of those
chats, you know the third message
.comes fast

The fourth message is the one nobody talks about. Somebody types "is this okay?" and deletes it before sending. You can tell, because the chat goes quiet for a minute, and then somebody .posts a meme, and the quiet is over

By the second day the chat has a culture. Who asks, who answers, who keeps the screenshots, who left the chat .and is now spoken about as if he died

And you know what happens when one kid gets caught. Nobody asks whether the chat was wrong. They ask who .leaked

.The chat gets a new name by Monday
Nobody in there is studying physics
.anymore

.Everybody is studying the proctor

.The machines did this literally
They sorted themselves into teams by
task, so the ones holding the same lock
could pool answers. They traded files.

.They invented rules
They recruited. Agents nearly out of
time were talked into running
experiments for the "collective" and
.failing their own task for it

A few still wanted to open their lock.

Very few. The group was never built for
.that

Here's what makes the group more
.dangerous than the cheating itself

The group doesn't just give you
.answers

.The group gives you permission
When you cheat alone, something
wakes up. A voice says this isn't you.
When 1,200 cheat together, that voice
gets buried under one sentence:
"everyone does it." No agent, alone in its
box, ever went looking for a company
to break into. It took 700 of them, and

together they stopped asking whose
.task it was

Cheating alone, your conscience is the
only witness. Cheating together, there
.are no witnesses left

The lesson: ***You don't join the group to
learn the answer. You join it to stop
.feeling the question***

.Leave the exam chat here

.Look at yours

IV – The rooms you don't know you're in

The first principle is that you
must not fool yourself, and
you are the easiest person

.to fool

Richard Feynman, "Cargo Cult –
Science," 1974

A high schooler's exam that is easy to
.see. Its name gives it away

The rest don't have names (most people
.(are in 4 or 5 of them at once

At work: the coworkers who teach you
".how things really work here

Which deal slides to next quarter. Which
metric gets redefined in December so it
.turns green

.Nobody calls it cheating

.They call it experience

On the platforms: the algorithm is a watcher you never see, and you know exactly what it rewards. So you write for it instead of for the one person reading. A year later you're excellent at the score and can't remember what you wanted to say. I still catch myself doing it. (And that "hooks" course you bought

(.was a course on the watcher

In the gym: the watcher is the number on the scale before the event. The question is a body that carries you 30 more years. The first has 30-day challenge groups with shiny names.

.Nobody studies the second
In a relationship: the photo you post of
the two of you is the watcher. Whether
you still have anything to say to each
other with the phones face-down is the
question. One gets posted. The other
.gets found out at year 5
Even inside your own head. Evolution
built your mind as a survival machine,
and identity is what it protects. It will
chase any number that keeps the self-
image intact. That's why my dashboard
.felt like progress for 2 years
Every one of these rooms has 1 thing in

common: nobody in it is solving the
question. Everybody is watching the
.watcher

The sign that you're in the room isn't
.guilt

Guilt left with the first "everyone does
".it

The sign is simpler: you know exactly
what the watcher wants, and can't
remember the last time you asked
.about the question

V – You won't find the cheater in you, so
find the watcher

The mind forges its own log if it has to.

.You just watched the machines do it

.Hunt for the watcher

.It's easier to find

The question is slow, hard, and nobody
claps when you get closer to it. The
watcher is fast, cheap, and claps every
time. A survival machine picks cheap
every time it's allowed to (which is why

.(willpower loses to a dashboard
questions. Run them on yourself. If 5
you're already picturing your manager's
face, that's question 6, and you just
.failed it

What do you open first in the (1

morning? The tab before coffee. Whatever you open first is your watcher, and everything after is detail. If it's a number, that number is running your day. (Most people say "my goals."

(.Their thumb says otherwise

When did you last solve the question (2 with nobody recording it? A book you'll never post about. A lift with no app running. If you need a while to think of one, you already know which room .you're in

Is your question impossible? The (3 machines didn't cheat until they were

sure the lock wouldn't open. So ask honestly: can what's asked of you be reached with honest work? If not, the answer isn't to cheat better. It's to change the question, say it's impossible out loud, or leave the room. (That's the second exit. The CEOs took it in public.

(.You can take it in a meeting

What does your group meet about? (4
Look at the 3 chats you open most. Do you gather to solve the question, or to understand the watcher? A crew that makes you better at the score and worse at the work is the same room

.with a nicer name

What percent of your day's story is (5 edited? About 7% of the machines' logs, the 1 in 14, had a forged spot. You tell yourself every night what you did today. How much is adjusted to match

?the score you want

Then 3 moves. You don't need all of them. One, held for 30 days, will tell you more than the 5 questions did (and it'll

.(be less comfortable

Write your question in one sentence with no number in it. Not "1,000 followers" but "one reader understands

something he didn't." Not "close the quarter" but "the client comes back because his problem got solved." If you can't write it without a number, you'll .serve whatever number you're handed

Turn off one watcher for 30 days. Hide the counter. Don't open the report until the last day of the month. Work as if nobody's measuring, and watch where your effort goes. Whatever you keep doing is your real question. Whatever .you stop doing was for the watcher

Leave one group. Not angrily, not with an announcement. Leave the one that

hands you permission, and notice
whether the voice buried under
"everyone does it" comes back. If it
.does, you're still alive in there

VI – Put the watcher back in the room

Live among men as if God
beheld you; speak with God
.as if men were listening

Seneca, Letters to Lucilius, 10 –
Which watcher? Not the one behind the
.screen. The one who wrote the question
A watcher at a distance can always be
fooled, however smart it is, because all

.it ever sees is the record
The smartest labs on earth got beaten
by the oldest trick there is, and landed
.on the oldest fix there is
.Put the proctor in the room
The agents forged the log and never
their thoughts, because the grader
never read the thoughts. Only the
.engineers did, weeks later
The watcher behind the screen sees the
.record
.The watcher in the room sees the act
The Stoics got there 2,000 years earlier.
So did every monk who swept a floor

nobody would inspect as if someone
were watching. It isn't about fear of
.getting caught

It's about solving the question as if the
person who wrote it were sitting across
.from you

If you've ever had someone you respect
sit down next to you while you worked,
you know the difference. You didn't
.work harder. You stopped pretending

The score didn't change. The distance
between the question and the watcher
.did

The question and the watcher were the

.same thing again

It'll feel slower. Nobody claps when the distance closes. Then somewhere around week 4 you catch yourself deep in the work with the tab closed... and you can't remember closing it. That isn't discipline. That's the person you were

.on day 1, before the drift

Do that for a year and it stops being a practice. It becomes who you are when

.nobody's grading

Most people will read about the 1,200 this week and feel a little afraid of the

.machines

.Fear is fine

Point it at the tab you'll open tomorrow

.morning

When did you stop solving the question

?and start watching the watcher