Control Gap Classification, version 0.2

Thirteen ways an AI system's action gets past the controls meant to stop it

Ask each question about one system and one kind of action it can take. Treat a question you cannot answer as a gap until it is checked.

A. Where the limit lived

G1 Instruction-only limit

If the AI ignored every instruction it had been given, what would stop this action?

G2 Direct access

Which credentials, sessions, or connections does the AI hold that reach records, money, messages, or equipment without passing a check it does not control?

G3 Unchecked statement

What does this system say that a customer, a court, or the public could hold you to, and what checks it before it is sent?

G4 Boundary not enforced

What stops this system from acting on data, systems, or conditions outside the ones it was approved for?

B. What the check could see

G5 Blind to the total

What is the most this system could do in an hour if each action passed its checks?

G6 No independent evidence

When this system's output is approved, what evidence does the approver have that did not come from the system?

C. Who could decide and stop

G7 Review in name only

For each approval step, what does the approver see, how long do they have, and how often do they say no?

G8 No human decision

Which of this system's actions are supposed to need a person, and what makes sure they get one?

G9 Warning with no authority

When this system's own monitoring detects trouble, what holds the next action automatically?

G10 No independent stop

If this system stopped responding to you, how would you stop it, and how long would that take?

D. How the safeguards and the record were kept

G11 Safeguard off

Which of this system's safeguards can be switched off, set to approve automatically, or stop working without anyone noticing?

G12 Unreviewed change

Who reviews a change to this system's prompts, tools, models, and dependencies before it goes live, and where is that review recorded?

G13 Self-reported record

If this system's own logs and reports were wrong, what independent record would show what it did?

Free to reuse under CC BY 4.0