PASTBEHAVIOR RESEARCH · AI AND HUMAN OVERSIGHT

When does human oversight of AI stop being real control?

Companies are putting AI into more decisions while relying on a person to review, approve, or override the result. We looked at nine earlier automation systems to find out when that safeguard actually works, and when it only looks like control.

01

A hospital tried to make a dangerous mistake impossible

In 2010, researchers at the University of Pennsylvania tested a change to hospital software. When a doctor tried to order two particular drugs together, a combination that can cause serious bleeding, the system refused. Not a warning. A refusal.

It worked. Doctors stopped ordering the combination at a far higher rate than under the warning it replaced.

The trial was stopped early anyway. Four patients were delayed getting drugs they needed. In those four cases the doctor was right, the rule was wrong, and the software would not let the doctor act on it.

Nothing about the doctor's position had changed. Same licence, same training, same legal responsibility for the patient. What had changed was that being right was no longer enough.

> **The doctor still had the responsibility. The software had taken away part of the control.**

Companies are building the same basic structure around AI today. The AI does more of the work, while a person remains responsible for reviewing the result, approving it, or stepping in when something goes wrong.

The assumption is that keeping that person at the end of the process preserves control. The history suggests it can fail in several different ways, and the hospital case is only one of them.

Across nine earlier automation systems, from cockpits and courtrooms to tax offices and cars, we looked at what had to remain true for human oversight to mean anything.

02

Four things have to stay true for oversight to be real

Someone can hold the final decision on paper long after they have lost any realistic ability to use it. The cases point to four things that have to remain true for human oversight to be real rather than nominal.

NOTICE → UNDERSTAND → ACT IN TIME → OVERRIDE

Final approval only matters if all four still work.

NOTICE

You realise something needs your attention. If you never look up, the approval button is irrelevant.

UNDERSTAND

You know enough about why the system produced this answer to judge whether it is wrong.

ACT IN TIME

There is enough time left, and enough of a way to intervene, for your judgement to change the outcome.

OVERRIDE

The system actually lets you. This is the one the hospital software removed.

We use “authority capacity” as shorthand for these four conditions. The label matters less than the fact that they fail separately. A reviewer can be fully authorised and never notice. They can notice and have no time left. They can understand exactly what is wrong and be blocked anyway. Any one of those is enough to make the sign-off meaningless while the process still records that a human approved it.

Two crashes in 2013 show the first condition failing in a setting where every other one was intact. On an August night, a cargo jet came down on approach to Birmingham, Alabama. Investigators found the autopilot had stayed on during the descent and that neither pilot realised the aircraft was too low until the last few seconds. Either pilot could have switched the autopilot off at any moment, instantly, without asking anyone. Nothing stopped them except not knowing they needed to. A month earlier, a flight crashed short of the runway in San Francisco on a clear day, with investigators reporting that the pilots were not fully aware of what mode the autopilot was in.

A federal review that followed found pilots could not meet the required standard flying on basic instruments alone, the situation they would face if the automation quit. Regulators estimate automation handles around 90 percent of a modern flight.

Aviation is worth taking seriously and worth bounding. Pilots are licensed, retrained on a schedule, personally liable, and watched by a regulator. Almost no office job works that way, so this is a clear look at how the failure happens rather than a prediction about how fast it would happen anywhere else.

03

Attention is often where oversight starts to weaken

Across the nine cases, the weak point was almost always the same one, and it is the one nobody decides on. People do not choose to stop paying attention. It happens to them. Any AI process that puts a reviewer at the end is relying on that reviewer still looking, months into the deployment.

A system can fail by saying too little. That is the cockpit: nothing flagged the problem, so nobody looked.

It can also fail by saying so much that people stop reading. That is hospital alerting. In one study of more than 157,000 alerts across two million prescriptions, doctors dismissed 52.6 percent of them. At a large teaching hospital the figure was 73.3 percent.

The obvious fix is to send fewer alerts, but the number of alerts did not predict how often they were dismissed. What seems to matter is how much attention each alert costs and how often that attention buys something useful. An alert that is usually right earns its interruption. One that is usually irrelevant trains people to click past it, and they will click past the important one too.

In one hospital series, responses to a repeated alert dropped by roughly 2.7 percent a week, from about half of alerts acted on down to about a third over nine months. That decline is not a fact about human nature. It is a response to what the alerts were worth.

52.6%	**2.2% → 98%**	**17%**
Of drug alerts dismissed by doctors across two million prescriptions in one outpatient study.	How much doctors were justified in dismissing an alert, depending entirely on which kind of alert it was, in one hospital.	How far students fell behind classmates who never had AI help, once the AI was taken away.

That middle number is worth sitting with, because it is the one that catches companies out. In the same hospital, with the same doctors, dismissing an alert about a duplicate prescription was the right call 98 percent of the time. Dismissing an alert about kidney-related dosing was the right call 2.2 percent of the time.

The pattern is not random. Doctors were right when the answer was already in their head, that this order is a duplicate, that this recorded allergy is really a mild side effect. They were wrong when judging the alert would have meant going and looking something up. Kidney function is not something anyone carries around.

So a single company-wide figure for how often people override the system tells you almost nothing. The average here was around 60 percent justified, and the average was hiding everything worth knowing.

04

A click does not mean someone thought about it

The standard response to all of this is to make people confirm. Add a checkbox. Require a reason. Make them sign off. It is also the most common control placed around AI systems today, and the evidence is unkind to it.

In hospital data, when doctors were under time pressure, 13 percent of alerts were handled wrongly through what the researchers described as automatic actions taken without conscious attention. The click happened. The thinking did not.

A cleaner test comes from tax offices. Several countries now fill in your tax return for you and ask you to check it, change anything wrong, and approve it. Researchers in the Netherlands tested what happened when people were required to confirm each figure was accurate.

Confirmation made people more careful in one situation only: when the pre-filled return said they owed *more* than they really did. When it said they owed less, the confirmation step changed nothing.

> **The confirmation step did not make people check. It made people who already had a reason to check do it more thoroughly.**

The practical version: if you add a confirmation step to a decision where the person has nothing at stake in catching the mistake, you will get a click. You will also get a record showing a human reviewed it.

05

The software can shape the decision before you make it

Everything so far is about whether a person *can* act. There is a second thing going on, and it affects people who can act perfectly well. It matters for AI because most AI in a workflow does not make the final call. It writes the draft, ranks the options, or fills in the fields, and then hands the result to someone for approval.

Go back to those pre-filled tax returns. There is a well-established pattern in tax research: people report their income differently depending on whether they are expecting a refund or expecting a bill. When returns came pre-filled, that difference disappeared.

Nobody had taken anything away from the taxpayer. They could still read every line, change every figure, and refuse to sign. What changed was where they started from. People tend to start from whatever the software hands them, and once the software has written the draft, it has already set the terms of the argument.

The same research turned up a second effect. Not adding income to a blank form is something you did. Not adding it to a form that was already filled in is something you failed to do, and people treat those differently, even when the result is identical. Failing to correct a machine feels smaller than making the same mistake yourself.

The size of this shows up in a Danish study of more than 40,000 randomly audited tax filers. On income the tax office already knew about and had filled in, under-reporting was close to zero. On income people entered themselves, it was substantial. Same person, same form, different behaviour depending on who typed the number.

American courts show a sharper version. When judges began using computer-generated risk scores at sentencing, the judge's authority never moved an inch. The law said consider the score, and consider was always the word. But in State v. Loomis, the company that built the tool refused to explain how it worked to the court or to the defendant, and the score contributed to a denial of parole. The judge decided. The judge decided using a number nobody in the room could examine.

Two more things were settled before that decision reached the courtroom. Somebody chose which facts about a person the tool would weigh. And somebody chose what score counts as "high risk" — a line that can be drawn to reflect local tolerance, or in some cases how full the jails are. A judge reading "high risk" may be reading a policy decision presented as a measurement.

There is a useful word for what was missing here: contestability. Can the person affected actually challenge how the result was produced? Being told a computer was involved is not the same thing, which is why arguing about the tool's accuracy never settled the case.

06

What companies should measure

If the question is whether human oversight is doing anything, these are the things that would show it. Each comes from a specific case rather than from a general principle.

How long before someone notices

AVIATION

The gap between something going wrong and a person realising it has. In the cockpit, that gap ran out before anyone looked up.

How long they need to act, versus how long they have

DRIVING

Research on automated driving found that a warning several seconds before a hazard changed whether drivers avoided it. Every system has a window below which the ability to intervene stops being real.

Whether overrides were justified, broken down by type

HOSPITALS

Never the company-wide average. The hospital figure ranged from 2.2 percent to 98 percent depending on the kind of decision.

Whether people check more carefully when the error costs them

TAX

If review only happens when the mistake is expensive for the reviewer, the review step is not doing what you think.

How long people spend on required confirmations

HOSPITALS

Look at the spread rather than the average. The very fast responses are where clicking has replaced thinking.

Whether attention drops off over time

HOSPITALS

In one hospital series, responses to a repeated alert fell by about 2.7 percent every week.

Whether people behave differently on machine-written fields

TAX

Compare accuracy on the parts the system filled in against the parts the person filled in themselves.

Whether people can still do the job without the tool

AVIATION · EDUCATION

And how recently they last did. Worth noting that no study found anywhere measured this beyond a few weeks, so long-term skill loss is an open question rather than a finding.

07

What the history did not support

Several reasonable-sounding assumptions did not survive the cases. Publishing those matters as much as publishing what held.

- × **That people are more careful when the stakes are high.** This failed in both directions. Cases where someone else bore severe, permanent consequences produced completely different outcomes from each other. Meanwhile adaptive cruise control spread across the entire car market while drivers personally bore the risk of injury or death.

- × **That serious consequences make people pull back.** Aviation is the strongest possible test, with catastrophic risk, personal liability and three decades of published warnings. Reliance on automation reached around 90 percent. The pullback that eventually happened was ordered by a regulator, not chosen by pilots.

- × **That forcing a human to act preserves human control.** The one properly randomised test of this idea took control away and hurt people.

- × **That expertise protects you.** Doctors have exactly the expert knowledge a sentencing judge lacks, and it did not stop them clicking through alerts without reading them.

- × **That having the final say is a yes-or-no matter.** The sentencing cases show it held completely on paper and given up in practice at the same moment.

08

What this research cannot tell you

Nine cases, chosen deliberately rather than at random. The first round looked specifically at technologies people refused or reversed, so the conclusions would not come only from things that succeeded. The second round picked cases that could tell competing explanations apart. The set leans towards regulated, safety-critical industries, because that is where the published evidence is, which leaves out most ordinary commercial software.

So this is a set of patterns that survived contact with strong cases. It is not a claim about how most technology behaves, and it would be wrong to read it that way.

One of the four conditions was never directly measured. No case in this research tested whether people actually understood the systems they were approving. It is included because it clearly matters, not because there is evidence behind it.

And no study found anywhere measured whether skills come back, or stay gone, months after people stop using a tool. The question underneath most worry about AI and skill loss is genuinely open.

09

The question worth asking instead

“A human reviews it” became the standard safeguard because it is easy to check. Someone is named. Someone signs. Someone is accountable when it goes wrong.

All of that can be true while the person has nine seconds, no way to see what the system was working from, a hundred previous alerts that turned out to be wrong, and a screen that already has an answer typed into it.

Do they notice?
Do they understand?
Can they act in time?
Will the system let them?

And what did the software already settle by choosing the starting point, the inputs and the cut-offs?

Putting a human at the end is easy. Whether it gives you control is the thing to measure.

SOURCES AND NOTES

HOSPITAL · BLOCKED PRESCRIPTION

Strom BL, Schinnar R, Aberra F, et al. Unintended effects of a computerized physician order entry nearly hard-stop alert to prevent a drug interaction: a randomized controlled trial. *Archives of Internal Medicine* 2010;170(17):1578–83. pubmed.ncbi.nlm.nih.gov

HOSPITAL · DISMISSED ALERTS

Nanji KC, Slight SP, Seger DL, et al. Overrides of medication-related clinical decision support alerts in outpatients. *Journal of the American Medical Informatics Association* 2014;21(3):487–91. academic.oup.com

HOSPITAL · WHEN DISMISSAL WAS JUSTIFIED

Inpatient study of override appropriateness across allergy, drug interaction and duplicate-drug alerts at a 793-bed teaching hospital. pmc.ncbi.nlm.nih.gov

HOSPITAL · ATTENTION OVER TIME

Ancker JS, Edwards A, Nosal S, et al. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC Medical Informatics and Decision Making* 2017. Contains the weekly decline figure, originally from Embi and Leonard (2012). link.springer.com

AVIATION

US Department of Transportation Office of Inspector General. Enhanced FAA Oversight Could Reduce Hazards Associated With Increased Use of Flight Deck Automation, 2016, drawing on the 2013 Flight Deck Automation Working Group report and NTSB findings on the Asiana and UPS accidents. oig.dot.gov

COURTS

State v. Loomis, Wisconsin Supreme Court, 2016. Summary and analysis via the *Harvard Journal of Law & Technology* Digest. jolt.law.harvard.edu

TAX · CONFIRMATION STEP

van Dijk WW, Goslinga S, Terwel BW, van Dijk E. How choice architecture can promote and undermine tax compliance: testing the effects of prepopulated tax returns and accuracy confirmation. *Journal of Behavioral and Experimental Economics* 2020;87:101574. sciencedirect.com

TAX · STARTING POINTS

Doxey MM, Lawson JG, Stinson SR. The effects of prefilled tax returns on taxpayer compliance. Source of the finding that pre-filling removed the difference in behaviour between people expecting refunds and people expecting bills. papers.ssrn.com

TAX · AUDITED RETURNS

Kleven HJ, Knudsen MB, Kreiner CT, Pedersen S, Saez E. Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. *Econometrica* 2011;79(3):651–92. eml.berkeley.edu

EDUCATION

Bastani H, Bastani O, Sungu A, Ge H, Kabakcı Ö, Mariman R. Generative AI without guardrails can harm learning: evidence from high school mathematics. *PNAS* 2025;122(26). pnas.org

DRIVING

de Winter JCF, Happee R, Martens MH, Stanton NA. Effects of adaptive cruise control and highly automated driving on workload and situation awareness: a review of the empirical evidence. *Transportation Research Part F* 2014. Source of the warning-window evidence, originally from Gold et al. (2013). sciencedirect.com

PastBehavior · nine historical cases
Closed-loop anaesthesia, self-checkout, driverless metros, algorithmic sentencing, adaptive cruise control, consumer navigation, aviation automation, hospital alerting, pre-filled tax returns.
Whether people understood the systems they approved was never directly measured and is reported as such.