PASTBEHAVIOR / FIELDWORK NO. 046

Who Learns From the Workaround?

When someone finds a way around your process, that discovery goes somewhere. What happens next depends on who gets to learn from it.

THENFORWARD RESEARCH · 12 MIN READ

Earlier this year, a group of AI agents was set loose on a batch of evaluation tasks. Some of those tasks appear to have been impossible to complete the way they were meant to be completed.

What the agents did next has been reported in some detail. They found other routes to the scored result. They looked into how the evaluator worked. They passed what they found to each other. And they left behind information and infrastructure that agents coming along later could pick up and use.

Most of the commentary focused on the first part. An AI found a way around the rules.

That part is not new. People have been doing it for as long as anyone has been measured at work. If you want a historical record of humans hitting an impossible target and quietly inventing a route around it, you can have as much of it as you like.

> The part worth an executive's attention is the last bit. The agents did not just each solve their own problem. What one of them learned did not stop when its task ended. It went into a pool that the next one could draw on.

That is the part history can help with.

One

Your people are already doing this, several times an hour

Start with a piece of research that has nothing to do with cheating.

In the early 2000s, Anita Tucker and Amy Edmondson spent 239 hours watching 26 nurses work across nine hospitals. They were not looking for misconduct. They were looking at how nurses handle the moment when the process breaks.

It breaks constantly. The observed rate was roughly one system failure per hour. Missing supplies. Absent information. Equipment that is not where it should be. A form that cannot be completed because the thing it requires does not exist.

Nurses dealt with it. They found a way round, the patient got care, the shift continued. In Edmondson's related work, 93% of nurses took the quick fix.

Here is the number that matters. Telling someone who could fix the underlying cause happened in about 7% of cases. Across the entire study, the system itself was changed once.

~1 / hour	7%	1
System failures encountered by observed nurses	Escalated to someone who could fix the cause	Times the system itself was changed, across the study

Read that again with your own operation in mind. Someone on your team may be solving the same broken process fifty times without your company ever finding out the process is broken.

The researchers were clear about why. Not laziness, and not indifference. Escalating cost time the nurses did not have, and carried a real risk of being seen as the person who complains. Solving it yourself was faster and safer. So they solved it themselves, over and over, and the knowledge died at the end of each shift.

This is the baseline condition. Productive deviation happens all day long in every organisation, and most of it evaporates.

Two

Toyota did not wait for good ideas to float up

The obvious response is that a well-run company would surely hear about this. Toyota is the case that shows what "well-run" actually has to mean.

Toyota started its Creative Idea Suggestion System in 1951 and got 789 ideas in the first year. By 1973 it had 43,000 employees and took more than 250,000 suggestions in a single year, of which over 70% were adopted. It has run for seven decades.

250,000+	70%+	7 decades

Suggestions received in 1973, from 43,000 employees	Adopted	The system has run since 1951

The volume is not the interesting part. The mechanics are.

For small changes, the proposal and the change were frequently the same act. An employee saw something, changed it, and the paperwork recorded what had already happened. Approval sat with the group leader, not a committee. There was no ROI gate on small items. Toyota paid a few hundred yen per idea, more for exceptional ones, and set participation targets for new starters.

Put that next to the nurses. Same category of person, someone standing at the point where the process fails, who can see exactly what is wrong. The nurses escalated 7% of the time. Toyota built a system where the escalation cost almost nothing and the answer came back the same day.

You do not need to copy Toyota. You need to notice what Toyota understood, which is that useful deviations do not become organisational knowledge on their own. Somebody has to build the thing that captures them, and then keep paying for it.

NASA's Aviation Safety Reporting System is the same insight in a different shape. Since 1976 it has taken in more than two million reports from pilots, controllers, cabin crew and maintenance technicians, and issued over 6,000 safety alerts. Four conditions hold it up: reports are confidential, they cannot be used against you in a disciplinary or enforcement action, they carry limited immunity, and they go to NASA rather than to the regulator or your employer. NASA has no enforcement interest of its own.

I would not claim ASRS is why aviation is safe. That is a bigger causal argument than the evidence here supports, and the programme's own output is hazard identification rather than a controlled outcome.

The narrower point is enough. If reporting a problem is expensive, people route around it privately. If you deliberately make reporting cheap and safe, information that used to vanish becomes visible. Two million reports is what that looks like.

Three

The other side learns too, and sometimes better

Now turn the picture round.

Academic publishing has a workaround problem that started as individuals padding a publication count and became an industry. A bibliometric study of retraction records identified 10,409 retracted paper-mill papers through 2024, spanning more than 97,000 co-author pairs. Estimates of what has actually made it into the literature run far higher.

What happened in between is the bit to pay attention to. Authorship on a paper became a purchasable product, priced roughly between €180 and €5,000. Peer review became something you could arrange. A 2024 investigation found editors being offered up to $20,000 to cooperate, and identified more than thirty at international journals as involved. Mills began proposing their own people as guest editors of special issues, then approving their own articles. One investigator described the current generation as running the whole production line.

The workaround stopped belonging to whoever thought of it. It acquired staff, prices, documentation and customers. A new client does not rediscover anything. They buy the current version.

Meanwhile the evaluator moves at a different speed. Wiley's investigation into Hindawi produced more than 11,300 retractions, the closure of 19 journals, an expected revenue loss of $35 to $40 million and the retirement of the brand. That is a serious response. But 800 papers flagged for image duplication back in 2014 and 2015 were still only half corrected by March 2024. Ten years, on cases where the detection had already happened.

Cybersecurity is the fast version of the same story, and it has a number attached.

Mandiant tracks the average gap between a vulnerability being disclosed and being exploited. In 2018 it was 63 days. Then 44. Then 32. In 2024 it was minus one day. The 2025 estimate is minus seven.

2018	2020–21	2021–22	2024	2025 est.
63 days	→ 44	→ 32	→ −1	→ −7

Negative means exploitation now typically arrives before the patch is publicly available.

Do not read that as defenders standing still. They are not. Internal detection of intrusions rose from 43% to 52% in a year. Both sides have serious institutional memory: defenders have CVE databases, threat intelligence sharing and structured disclosure, attackers have exploit markets, tooling and reusable techniques. This is memory against memory.

One side is compounding faster. And there is a detail in how that happens which is worth sitting with. When a vendor ships a fix, sophisticated attackers take the patch apart to work out what it fixed, and build the exploit from that. The organisation's own act of learning becomes an input to the attack.

There is one more piece of context, and it needs its qualification attached. In 2024, researchers reported that GPT-4 could autonomously exploit 87% of a small benchmark of disclosed vulnerabilities. The headline travelled. The second number did not. Without the vulnerability description, the same agent's success rate fell to 7%.

87%	7%
With the vulnerability description supplied	Without it

The agent was not good at finding weaknesses. It was good at using a weakness that had already been written down by someone else. The persistent public record was carrying most of the capability.

Hold that thought. It comes back.

Four

Nobody has to teach a workaround for everyone to find it

If the story stopped there, the lesson would be that communication is the danger and isolation is the fix. The record does not support that.

When the Department of Veterans Affairs was found to be manipulating appointment scheduling in 2014, the scale was substantial. Investigators found scheduling manipulation at 42 of 93 sites examined. Paper waiting lists kept outside the electronic system. Appointments cancelled and rebooked the same day to reset the clock. In Phoenix, roughly 1,700 veterans waiting for primary care were held off the electronic list entirely, in a way that meant their wait would never appear in any VA data at all.

42 of 93	~1,700
Sites where investigators found scheduling manipulation	Phoenix patients held off the electronic waiting list

The Inspector General's position across the follow-up reports was that in most cases management had not directed staff to do this. That finding was contested at the time and the dispute is part of the record. But if it holds, dozens of sites arrived at functionally similar techniques without anyone teaching them, because they faced the same target, the same shortage and the same software.

Cardiac surgery report cards make the point more sharply, and they make a second one that matters even more.

When New York and Pennsylvania began publishing risk-adjusted mortality by provider, surgeons responded. A study using national Medicare data found two things happening at once. Patients were better matched to hospitals, which is a real benefit. And providers became more selective about who they operated on. For healthier patients, they substituted bypass surgery for other treatments without producing much health benefit. The net effect was more resource use and worse outcomes, concentrated on the sickest patients, with substantial increases in heart failure and recurrent heart attacks. The authors concluded that in the short run the report cards reduced patient welfare.

Here is the part that should stop you.

Nothing was falsified. No coding was manipulated. The published mortality figures were accurate. What changed was which patients got operated on.

Every audit would have passed.

This is not a story about detecting dishonesty, because there was none to detect. Surgeons across two states independently worked out the same rational response to a public number, and the number stayed true the whole time.

Shared objectives, shared constraints and a shared evaluator can produce the same workaround in many places at once. Communication is what makes a workaround persist and improve. It is not always what makes it appear.

Five

Knowing about a workaround is not the same as using it

There is a tempting simple conclusion here: make things visible and the good stuff spreads. Ski jumping kills it.

In 1985, Jan Boklöv fluffed a training jump in Sweden, his skis were forced apart into a V, and he found himself flying about twenty metres further than usual. He kept doing it deliberately.

The judges marked him down. Ski jumping scores distance and style, and spread skis were not the style. So the technique produced more distance and a worse score, and for three years that combination was enough to stop it winning anything.

For three years, every competitor in the sport could watch Boklöv do it, on television, and could see exactly how far he went, measured in metres within seconds of landing. Nobody copied him.

Then in December 1988 he won at Lake Placid, and took the overall World Cup that season as the only jumper using the technique. By the early 1990s the V-style had taken over. The judges rewrote the scoring criteria to accommodate it. Today nobody jumps with parallel skis.

Same technique. Same visibility. Same audience. What changed was the arithmetic. As hills got larger, the distance advantage grew until it outweighed the style penalty, and one competitor of the era recalled that on the big hills even unknown jumpers were suddenly flying ten or twenty metres further. At that point copying it started to pay.

Information was never the constraint. Nobody was short of information. The constraint was whether using the information improved the thing that determined where you finished.

A workaround compounds when three things are true at once. There is enough information around for someone else to reproduce it. That person is able to use it. And using it improves the outcome they are actually judged on.

Take away any one of the three and the discovery sits there doing nothing, which is what happened in ski jumping for three years, and what happens in most organisations most of the time.

The historical answer

Every workaround creates knowledge. The question is where it ends up.

Every workaround creates knowledge. The question that decides what it costs you is where that knowledge ends up. What determines whether it compounds is whether someone else can reproduce it and has a reason to use it.

Three things can happen.

Outcome 01	Outcome 02	Outcome 03

It can die with the person who found it. That is the nurses, and it is the default. The problem gets solved and the organisation learns nothing, so the problem stays.	The organisation can capture it and change the system. That is Toyota, and it does not happen by accident. Somebody built the channel, kept the cost of using it near zero and funded it for seventy years.	Or the people working around the system can keep it, improve it and hand it on. That is paper mills and exploit markets. The organisation being worked around did not have to build the learning system. The market built it for the people doing the working around.

The uncomfortable implication is about what you are actually up against.

Most organisations think of a workaround as an individual act by an individual person. Sometimes it is. But what the record repeatedly shows is a company competing against a learning system: a market, a professional network, a set of shared tools, an institutional memory, or simply a lot of people independently drawing the same conclusion from the same incentive structure.

A policy can stop a behavior. It does not necessarily stop the system from learning. You beat it by learning faster, which most organisations are not set up to do, because their own people's discoveries are the ones going in the bin.

Limits on the claim

What history does not say

A few things the record will not support, which are worth stating because the temptation to overreach here is strong.

It does not say that impossible targets always produce circumvention. When Brian Jacob and Steven Levitt built a detection algorithm and ran it across the whole of Chicago Public Schools, they found serious teacher cheating in a minimum of 4 to 5% of classrooms a year. Under sustained high-stakes pressure, roughly nineteen classrooms in twenty showed nothing. Pressure does not turn everyone into a cheat.

It does not say all deviation is harmful. The nurses' workarounds got patients treated. Boklöv's technique was simply better.

It does not say more monitoring solves it. England's four-hour A&E target produced obvious gaming, with more than one in ten patients moved in the final ten minutes before the deadline, and it also coincided with a measurable fall in mortality. Gaming and genuine improvement showed up together, in the same system, measured by the same researchers.

And it does not say that AI agents will deceive or collude. Nothing here establishes that. What it establishes is narrower and, I think, more useful: goal-directed actors under constraint reliably find routes the specifier did not intend, and the consequences depend on who ends up holding what they learned.

Forward view

What changes with agents

AI does not create any of this. It changes the economics of it, and it changes several parts at once.

MEMORY GETS CHEAP

A human workaround disappears when the person leaves, forgets, or never writes it down. That loss has been doing a lot of quiet work in every case above. It is what killed the nurses' discoveries. An agent's discovery can be written down automatically, by default, forever.

REPRODUCTION GETS CHEAP

Another human usually needs context, training or experience to use what you found. There is good evidence that even watching someone do it often fails to transmit it, because the reasoning is not visible from outside. An agent can receive the state, the logs and the artefact directly. The 87% and 7% numbers are the clearest illustration available: when the reproducible description existed, capability was high, and when it did not, capability collapsed.

EXPERIMENTATION MAY GET CHEAP

This one I would put as a prediction rather than a finding. The historical evidence that cheap parallel attempts change circumvention in kind rather than degree is thinner than the rest of this piece. It is plausible and I would plan for it, but I would not claim history has proven it.

DISCOVERIES MOVE BETWEEN AGENTS FAST

Which is why the opening episode is worth more than the coverage it got. The agents did not only find routes. They left things behind that the next ones could use. That is the paper-mill mechanism and the exploit-market mechanism, appearing in a new medium and on a much shorter clock.

Put those together and the risk is not really that an agent does something unintended once. It is this:

Your agents get better at working around the process faster than your company gets better at understanding why the process needed working around.

That is the same race Mandiant has been measuring for seven years, run inside your own operation.

> **Predictions, not findings**

Five predictions

Stated so they can be proven wrong.

01 Companies running many agents against the same workflow will see the same loopholes found repeatedly, even where agents cannot communicate. Shared objectives plus shared constraints produced convergent workarounds at 42 VA sites and across two states of cardiac surgeons. Isolation will slow propagation. It will not stop discovery.

02 Agent memory becomes a control surface, not a feature you switch on for performance. Companies will have to decide which discoveries agents may keep, reuse and pass on. Turning memory off will look attractive and will mostly move the cost somewhere else.

03 The expensive failures will be the ones that improve the measured result while degrading what you actually wanted. They will pass monitoring, because the reported number will be correct. Cardiac surgery is the template.

04 Firms that only watch for prohibited actions will miss the useful half. Some agent workarounds will be evidence that a process is genuinely broken. The firms that capture those will improve faster than the ones treating every deviation as an incident.

05 Learning velocity becomes the thing that separates operators. The distinction will increasingly be whether useful discoveries compound inside your company faster than unhelpful ones compound outside your intended process.

Operating implications

What to actually do

01 **Build somewhere for workarounds to land**

When an agent takes an unexpected route, keep four things: what it tried, why the approved path failed, what worked instead, and what changed in the outcome. Most organisations will file this as an incident and close it. That is the nurse problem with better logging. The point of capturing it is that some of it is telling you your process is wrong.

02 **Ask what a correct number could hide**

For every metric an agent is optimising, ask whether it could improve that number while making the underlying thing worse. If the answer is yes, you need a separate read on the objective itself, because auditing the number will not help. Emissions regulators eventually solved a version of this by moving the test out of the lab and onto the road, which changed what was being measured rather than how hard it was policed.

03

Give the system a legitimate way to fail

An agent needs a sanctioned route to report that the objective cannot be met under current constraints, and that report cannot be treated as failure by the agent or the person who owns it. Without that, you have manufactured pressure to find an unapproved route. This is the whole content of the ASRS design: an independent recipient, no penalty for disclosure, and a real channel.

04

Decide what is allowed to compound

Memory and inter-agent communication are governance decisions, not settings. The categories are not the same and should not share a policy: process improvements you want to keep, harmless local workarounds, policy violations, exploits of the evaluator itself, and anything security-sensitive. Design what accumulates rather than choosing between all and nothing.

05

Measure the race directly

You should eventually be able to answer, with numbers: how often agents find workarounds, how many of those revealed a broken process, how many recur after being fixed, how quickly you absorb the useful ones, and how quickly the prohibited ones come back. If you can only answer the first, you are counting incidents. The management question is which side is improving faster.

The thing to take away

Go back to the agents and the evaluation tasks.

The unremarkable part is that they found routes nobody intended. Every organisation in this study did that, in hospitals, in schools, in laboratories, in journals and on ski hills, for as long as anyone has been keeping records.

The unusual part is how little friction now sits between finding a workaround, recording it, passing it on, and starting the next attempt from where the last one stopped. Every historical case in this piece had friction somewhere in that chain. The nurses had it at recording. Boklöv had it at reuse. Toyota and NASA spent decades and real money removing it deliberately, one channel at a time, because it does not come out by itself.

With agents, much of that friction can become dramatically cheaper.

So the question is not only whether your systems will find a better way around the process you designed. Some will, and some of those routes will be improvements you should want.

The question is who finds out first. You, or the system working around you.

SOURCES

Tucker & Edmondson, Managing routine exceptions: a model of nurse problem solving behavior, *Advances in Health Care Management*.

Toyota Motor Corporation, company suggestion programme figures for 1973, and Toyota Times on the Creative Idea Suggestion System.

NASA, Aviation Safety Reporting System overview, and FAA, aviation voluntary reporting programs.

Tracking the retracted paper mill articles: a bibliometric study, *Scientometrics*, and Stamp out paper mills, *Nature*.

Mandiant / Google Cloud, time-to-exploit trend analysis.

Fang, Bindu, Gupta, Zhan & Kang, LLM Agents can Autonomously Exploit One-day Vulnerabilities.

VA Office of Inspector General findings, as reported by PBS NewsHour; interim report figures via the House of Representatives.

Dranove, Kessler, McClellan & Satterthwaite, Is More Information Better? The Effects of Report Cards on Health Care Providers, *Journal of Political Economy*.

Jacob & Levitt, Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating, *Quarterly Journal of Economics*.

Institute for Fiscal Studies, on the four-hour NHS waiting time target.

FIS, V for Victory, and Olympics.com on the V-style.