ITSEC
CYBERSECURITY DELIVERED

VI - July 2026

# ITSECBUZZ

## WHEN HOME GETS SMARTER, THE DOOR OPENS WIDER

## TACKLING XDR ALERT FATIGUE

## BEYOND ANNUAL PEN TEST: ADAPTING TO MODERN THREATS

Editor-in-chief:
M. Rasyid Sahputra

Design & Layout
Agung Prasetya
Zurly Ananda
Billal Maulana

Writer
Irra Fachriyanti
Zurly Ananda
Brian Hikari
Candra
Haekal Alghifary

Pictures:
AI Generated
ITSEC Asia
Freepik.com

# INDEX


# AI AND PARENTING

*By Irra Fachriyanti*

*"My friend says AI will make us stupid," my son told me one evening. "It makes people lazy." He did not say it casually. He said it as if he was weighing the idea. I asked him why. "Because if it answers everything, we stop thinking."*
*I paused...*

I had heard that concern before, just in a different form. When Google first became common, many older people worried that instant access to information would weaken memory and reduce critical thinking. Some of those fears were understandable. Access became easier. Effort seemed optional. But over time, we learned something important. Technology does not determine outcomes on its own. Habits do. So I told him, "AI will not make you stupid. But how you use it might." That conversation stayed with me.

## The New Generation of Intelligence

Artificial intelligence is often described as the next version of Google, but the comparison only goes so far. Search engines present options. They require users to sift through links, compare sources, and decide what to trust. The thinking process remains visible. Effort is required after the search. AI works differently. It synthesizes information before the user even sees the options. Instead of offering multiple links, it delivers a structured response. It filters, organizes, and summarizes instantly. The answer often feels complete. This shift changes the learning experience. With traditional search, time was spent finding information. With AI, much of that time is compressed. A question that once required hours of searching can now be explored in minutes. For a genuinely curious child, this can be transformative. The more they want to understand about the world, the more they can explore within the same amount of time. Instead of spending energy locating material, they can spend more energy engaging with it. In that sense, AI does not automatically reduce thinking. It can expand opportunity. But efficiency introduces a new responsibility. If time saved is reinvested into deeper reflection and broader understanding, growth accelerates. If time saved is used to avoid effort entirely, learning weakens. The tool has evolved. The need for guidance has not.

## What AI Can Do and What It Cannot

Artificial intelligence is highly capable within a specific domain: information and analysis. It can explain difficult concepts clearly, summarize large amounts of material quickly, suggest improvements in writing, and provide stru tured reasoning. For students, this can be a valuable support system. However, parenting involves far more than delivering information. When a child struggles, the issue is often not just intellectual. It may involve confidence, emotional regulation, disappointment, or perseverance. A correct answer does not automatically produce resilience. A clear explanation does not guarantee maturity.

AI can identify errors and provide solutions, but it cannot cultivate endurance. It can suggest strategies for resolving conflict, but it cannot demonstrate patience in real time. It can describe values such as honesty or responsibility, yet it does not live by them or bear the consequences when those values are ignored. Parents may choose to use AI as a reference tool. It can offer perspectives on communication, discipline, or motivation. It can present options that a parent may not have previously considered. But suggestions are not decisions. Parents remain the ones who observe their children closely. They understand temperament, sensitivities, strengths, and patterns of behavior. They recognize when firmness is necessary and when reassurance is more appropriate. They carry responsibility for the long term direction of their child's development.

AI can generate options. It does not carry accountability for outcomes. Technology can inform. It cannot assume the role of parent.

## Convenience and the Adults We Are Raising

The real risk of AI is not that it will think for our children. It is that it will make it easier for them not to think. Convenience has a quiet influence. When

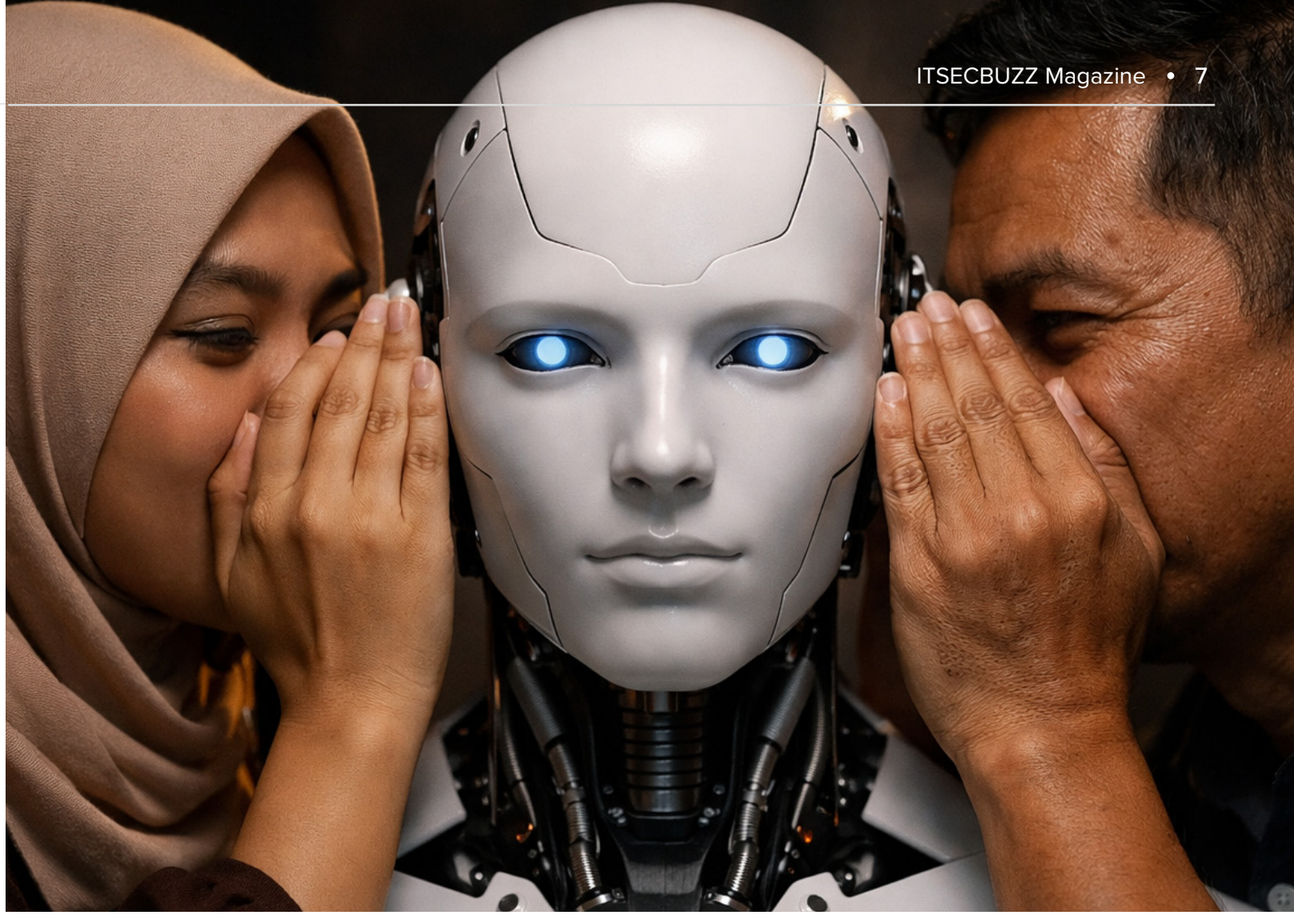
answers arrive instantly, the patience to wrestle with uncertainty can weaken. When drafts are generated in seconds, the discipline of shaping one's own ideas may slowly fade. There is no dramatic collapse, only gradual dependence. Technology does not force this outcome. But it makes it easier. That is why the conversation cannot stop at what AI can do. It must include what kind of habits we are allowing to form. Our children will grow up in a world where AI is not new or impressive. It will simply be normal. It will sit quietly inside search engines, workplaces, classrooms, and daily decisions. The question is not whether they will have access to powerful tools. They will. The question is whether they will develop judgment strong enough to use those tools wisely.

In a future where information is abundant and instant, depth may become rarer. Discernment may matter more than data. Character may matter more than speed. If children learn to pause before accepting, to question before copying, and to think before delegating, AI can become an amplifier of growth. Without those habits, it can quietly replace them. The technology will continue to evolve. Our responsibility will not.

## Closing

Artificial intelligence will continue to advance. It will become more embedded in daily life, more natural, and more invisible. Our children will grow up with it as a given, not as a novelty. The question, then, is not whether AI will shape their world. It will. The question is whether we will shape them. Parenting has never been about controlling every influence. It has always been about cultivating judgment, discipline, and character strong enough to navigate influence wisely. AI may support their learning. It may expand their access to knowledge. But it does not replace presence. It does not replace values. It does not replace responsibility. In the end, the tools will change. The work of raising thoughtful, principled adults does not. And that work remains ours.





# WHAT YOU TELL AI

*By Z.Ananda*

*What actually happens to our data after we click "Send"?*

*We're sharing more with AI than we realize. Health concerns. Work problems. Important documents. It's comfortable, fast, and judgment-free. But there's one question most of us never stop to ask: after we click Send, where does that data actually go?*

Picture this. It's late. There's a bug in your code that has to be fixed before morning. You open AI code generation tools, paste the code in, and within seconds the problem is solved. You close your laptop and go to sleep.

But that code you just pasted—the exact text containing your company's proprietary system architecture and core business logic—is now sitting on the AI's server somewhere in the cloud. And the truth is, you have zero control over what happens to it next.

This is precisely what happened at Samsung Electronics in early 2023. Within less than a month, three separate engineers independently uploaded internal source code and confidential meeting transcripts into ChatGPT. They weren't being reckless; they were simply trying to optimize their workflows. Nobody had malicious intent.

While the immediate financial fallout from that specific incident couldn't be easily quantified, Samsung's

reaction speaks volumes: they immediately banned all employees from using external generative AI platforms entirely. A corporation valued at hundreds of billions of dollars chose a response that drastic. If the risks weren't profoundly real, they wouldn't have bothered.

### Data Stays Far Longer Than You Think

One of the most misconceptions about AI is the belief that once a conversation ends, your data disappears. The reality is nowhere near that simple.

OpenAI retains uploaded files for up to 30 days purely for security monitoring purposes. Google maintains varying data retention windows depending on your specific enterprise or retail service tier. More importantly, unless you actively dive into your account settings and disable data sharing, your conversational history is highly likely to be used to train the next generation of AI models.

Think of cloud security like a gated community. The community manager—in this case, the AI service provider—is responsible for keeping external threats out of the complex. They build the walls: data encryption, enterprise firewalls, and intrusion detection systems. But what happens inside the house is entirely up to the resident. Locking the front door, setting the security alarm, and deciding which guests to let inside—that is your responsibility.

Almost every cloud vendor explicitly enforces a shared responsibility model: they're responsible for the security of the infrastructure, but you own the data content itself. This means the security of the information you hand over relies almost entirely on how carefully you manage it.

Even the most robust enterprise systems can fail. In March 2023, a technical bug in OpenAI's system exposed chat titles and partial payment information belonging to active users to completely unrelated accounts over a nine-hour window. Full credit card numbers weren't exposed, and OpenAI moved quickly to fix it. But the incident proved something that can't be ignored.

### How Casual Conversations Silently Build Your Profile.

There is another, far more subtle risk beyond data breaches, and it's the one that gets overlooked most often: profiling.
When you ask an AI tool about the recurring headaches you've been having, you might mention your age, your sleep patterns, the medications you've tried, how often the symptoms show up. You do this because it's very logical—more detail means better answers.

But from the system's perspective, every single variable is a data point. When aggregated across millions of similar interactions, these data points build highly detailed demographic, health, and behavioral profiles. To a pharmaceutical giant, an insurance underwriter, or a targeted advertising network, that data is worth significantly more than the average user realizes.

Major AI platforms generally claim they do not sell your conversation logs to third parties. However, for the flood of smaller, niche AI apps

source : haveibeenpwned.com

"People often assume 'cloud means safe.' But the cloud is just someone else's computer — with all the vulnerabilities that come with it."

***Troy Hunt,*** *Founder of Have I Been Pwned & Cybersecurity Researcher*



source : epfl.ch

"Users should start treating AI the way they treat social media — don't share what you wouldn't want to see somewhere else."

***Bruce Schneier,*** *Cybersecurity Expert & Author of 'Data and Goliath'*

popping up daily without mature privacy architectures, the playbook changes completely. It is the golden rule of the social media era: if the product is completely free and the monetization strategy is vague, you are not the customer—your data is the product.

This risk multiplies exponentially when you upload full documents. A resume contains your entire employment history and personal contact details. A business contract details transaction structures and corporate strategy. A financial spreadsheet reveals the exact operational health you want to hide from competitors. All of this goes up in a single click, and the moment that file leaves your machine, your control over its trajectory is effectively gone.

### The Enterprise Choice: When the Stakes Are Massive

For individual users, managing these liabilities comes down to building cleaner personal habits. But for organizations—corporations, hospitals, government agencies, and law firms—the fallout scales up fast: regulatory fines, reputational ruin, and in extreme cases, existential business failure.

This is where the structural choice between Cloud AI and On-Premise AI becomes the ultimate differentiator.

Cloud AI, the most frictionless approach, requires routing organizational data to external provider servers outside internal perimeters. It is fast, highly cost-effective at launch, and ready immediately. But the second data crosses your internal network boundary, data protection becomes a shared obligation with an outside vendor.

On-Premise AI flips this model completely. The AI systems run entirely on your own internal infrastructure. Data never leaves your physical or virtual network. Audits can be executed instantly on demand. Total governance remains in corporate hands. This is the exact architecture being deployed across banking, healthcare, and state sectors—industries where a single data leak triggers devastating regulatory penalties or destroys decades of hard-won public trust.

The upfront capital expenditure is undeniably steep—hardware infrastructure, custom software licensing, and specialized engineering talent. But for organizations handling truly sensitive information, this isn't a matter of preference anymore. It is basic, responsible risk management.

### Actionable Steps You Can Take

The answer here isn't to quit using AI. These tools are very helpful, and going to be better. The solution  is simple: start using them with more awareness.

For individual users:

- **Check your settings immediately:** Open ChatGPT or Gemini right now. Find 'Data Controls' and turn off the option to use your conversations for model training.
- **Leverage temporary chat:** Use features like Temporary Chat for sensitive brainstorming. Conversations aren't saved and aren't used for training.
- **Redact sensitive ID:** Never type government ID numbers, bank details, passwords, or ex-

> "On-premise AI isn't just about data security — it's about digital sovereignty. Organizations that manage their own AI infrastructure aren't dependent on third-party business decisions that can change at any time."
>
> ***Dr. Ann Cavoukian,*** *Creator of Privacy by Design & Former Information & Privacy Commissioner of Ontario.*


plicit financial data into an AI chat prompt under any circumstances.

- **Pause before uploading:** Before dropping a document, ask yourself a simple question: am I completely comfortable with a stranger reading this exact file?
- **Investigate unfamiliar tools:** For any niche AI application you use, read their business model first. If there is no clear privacy policy on display, treat it as a red flag.

For Enterprise and Corporate Environments:

- **Create an explicit usage policy:** Establish clear, written rules defining exactly what categories of data can and cannot be processed by external AI tools. The Samsung incident didn't happen because their perimeter was breached; it happened because there was no policy.
- **Deploy regular awareness training:** Make AI-related cybersecurity awareness training a regular agenda item.
- **Audit your vendors rigorously:** Assess every external AI vendor using recognized industry frameworks: SOC 2 compliance, ISO 27001 certification, or alignment with the NIST AI Risk Management Framework.
- **Run a real cost-benefit on local models:** For data subject to strict regulation, seriously evaluate on-premise AI options before making decisions based on implementation convenience alone.

### The Bottom Line

AI has woven itself into our everyday lives. That reality isn't changing, and it does not need to. But there is a massive gulf between deploying a tool you deeply understand and deploying a tool you simply trust blindly.

The first path makes you incredibly efficient. The second path, eventually, will cost you.

References:

Samsung bans use of generative AI tools like ChatGPT after internal data leak , Bloomberg, April 2023 | OpenAI March 20 ChatGPT outage: Here's what happened , OpenAI Blog, Maret 2023 | Schneier, B. (2015). Data and Goliath. W. W. Norton & Company | Cavoukian, A. Privacy by Design: The 7 Foundational Principles , Information & Privacy Commissioner, Ontario | Hunt, T. , haveibeenpwned.com | OpenAI Data Controls FAQ , help.openai.com | NIST AI Risk Management Framework (AI RMF 1.0) , NIST, 2023.




# WHEN HOME GETS SMARTER, THE DOOR OPENS WIDER

*By Irra Fachriyanti*

*Smart home technology promises control. What it doesn't advertise is the new kind of vulnerability that comes with it.*

Picture yourself coming home after a long day. The lights turn on as you open the door. The air conditioning has been cooling the room since you were still in the car. A robot vacuum glides quietly across the floor. You check the security camera from your phone while you're still in the hallway.

Smart home technology has moved well past the luxury-gadget phase. The global IoT home device market is projected to surpass $222 billion by 2027, nearly double its 2022 figures (Statista, 2023). These devices make life feel more efficient, more convenient, more in control.

And that's exactly where the important question lives: as your home gets smarter, does it also become a larger target?

Because today, threats to your home don't always come through a forced window or a lock picked in the middle of the night. Sometimes they come in quietly through your internet connection.

### When Home Becomes a Target

In December 2019, thousands of Ring camera users across the United States woke up to a digital nightmare. Someone had gained access to Ring cameras inside their homes — including cameras in children's bedrooms — and started speaking directly to residents through the camera's

speaker. An 8-year-old girl in Mississippi was reportedly woken up by an unfamiliar voice coming from her own camera. The intruder knew her name.

*Not a horror movie. Not a haunted house.*

Someone had gotten into their smart home system and taken partial control of their home from a distance. Amazon (Ring's owner) confirmed the incidents weren't caused by a breach of Ring's systems directly — they happened through credential stuffing: attackers used email-and-password combinations leaked from other platforms and systematically tried them against Ring accounts.

A separate incident in March 2023 showed that the threat doesn't always require an attacker at all. Wyze, an American smart camera company, suffered a serious system failure that allowed roughly 13,000 users to accidentally view camera feeds belonging to complete strangers — living rooms, backyards, the daily routines of people they'd never met. Wyze confirmed the incident publicly and attributed it to a third-party component failure.

No door was forced. No glass was broken. Privacy collapsed simply because of a gap in the digital layer.

## The Real Scale of the Threat

The data makes the picture harder to dismiss. According to Nokia's Threat Intelligence Report 2023, IoT devices now account for 33% of all malware infections detected on mobile networks — up from 16% just two years earlier. The same report found that the average broadband-connected household now has 21 devices connected to the internet.

More troubling: research from Bitdefender (2023) found that 34% of smart home devices have at least one unpatched vulnerability. That's roughly one in three devices in your home functioning as an unlocked door — one that just happens to be invisible.

## Why Can a Smart Home Be Hacked?

The problem is often not the concept of the smart home itself, but the way it is used.

### 1. Passwords That Are Too Easy to Guess

Birthdates. Children's names. "123456". These are still the most common passwords on the internet, and when one account using a reused password gets compromised elsewhere, every device connected to that account becomes accessible. That's exactly what the Ring attackers exploited. One leaked credential becomes a master key to the whole ecosystem.

### 2. Skipping Two-Factor Authentication (2FA)

2FA feels like extra friction, so people skip it. But a joint study from Google and New York University (2019) found that SMS-based two-factor authentication blocks 100% of automated bot attacks and 96% of bulk phishing attacks. Without that second layer, a smart home account is significantly easier to take over.

### 3. Devices That Never Get Updated

People update their phones. They rarely update their home cameras, smart TVs, routers, or smart locks. Research from Stiftung Warentest (2022) tested 13 popular IP cameras and found that more than half had firmware that was over a year out of date with no active security patches. Updates aren't just new features — they're fixes for vulnerabilities that attackers already know about.

*"Every smart device in your home is constantly talking to some server over the internet. Even when you're not using it, it sends small signals saying, 'I'm here, I'm working, this is my status.' We can generate a fingerprint for each device using the way it talks over the network.'*

Chakshu Gupta
Cybersecurity
Researcher.



### 4. Cheap Products Without Security Standards

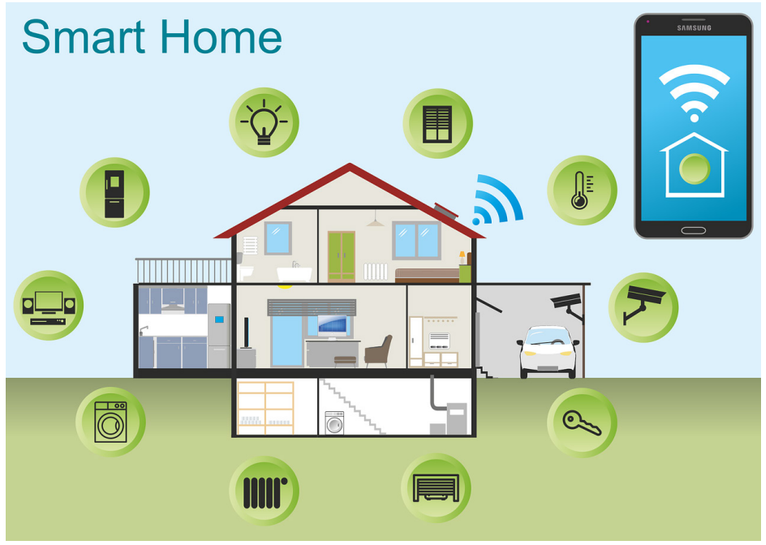
Not all IoT devices are built with adequate security. A Consumer Reports investigation (2023) found that a number of inexpensive cameras from no-name manufacturers were using unencrypted communication protocols, storing footage on servers without adequate protection, and receiving virtually no security updates. The device looks like a bargain. The risk it introduces is anything but.

### 5. Everything on One Network, Without Segmentation

Your work laptop, home camera, smart TV, kids' tablet, and robot vacuum are all on the same Wi-Fi. This is a common architectural mistake with serious consequences. If one weak device gets compromised — say, an old smart TV that stopped receiving updates — an attacker can use it as a jumping-off point to reach everything else on the network. It's a house with many rooms and only one security door.

### 6. Too Many Third-Party App Access

The more third-party apps you connect to your smart home ecosystem, the more potential entry points you're creating. Most people click "Allow" without reading what they're actually granting — including access to microphones, cameras, and location data that has no obvious relevance to what the app does.




### 7. Believing Technology Automatically Means Safe

This is the most common mistake of all. A smart home still requires the same discipline as any other system — it just requires it in a different form. The technology doesn't protect itself. You do.

## What You Can Actually Do About It

Understanding the risk is only useful if it leads somewhere. Here are the concrete steps that make a real difference.

### 1. Separate IoT devices from main network.

Most modern routers (ASUS RT series, Netgear Nighthawk, and similar) let you create a guest network or VLAN specifically for smart home devices. Put everything IoT on that separate network. If your camera or smart speaker gets compromised, the attacker can't then hop across to your laptop or NAS drive. To do it: create a "Guest Network" in your router settings → connect all smart home devices to it → make sure the "Allow guests to access local network" option is turned off.

### 2. Audit Your Devices with a Network Scanner

Use an app like Fing (free on iOS/Android) or Angry IP Scanner (desktop) to map every device connected to your home network. Check for any device you don't recognise. Once identified, verify whether its firmware is the latest version on the manufacturer's official site.

### 3. Choose Products with a Clear Security Track Record

Before purchasing any IoT device, ask three questions: Does the manufacturer have a clear security disclosure policy? How long do

they commit to providing security updates? Does the device support matter — the interoperability standard developed jointly by Apple, Google, Amazon, and Samsung with security built in as a core requirement? Manufacturers like Google Nest, Apple HomeKit, and Arlo generally have more structured security response programs than no-name brands on discount marketplaces.

**4. Follow Established IoT Security Standards**

Two references worth knowing: ETSI EN 303 645, the European consumer IoT security standard covering 13 baseline requirements including banning default passwords and mandating ongoing security updates; and NIST IR 8259, the US National Institute of Standards and Technology guide defining the baseline security capabilities IoT devices should have. Both can serve as a practical checklist when evaluating a product you're considering.

**5. A Smart Home Can Still Be the Solution**

None of this means smart home technology is something to fear. The benefits are real. Security cameras help you monitor your home when you're away. Door sensors alert you to suspicious activity. Automated lighting makes an empty house look occupied. A study from Alarm.com (2022) found that homes with well-configured smart security systems have a 60% lower burglary risk than homes without them.

| Action | Status |
|---|---|
| Unique, strong password for every device account | |
| Two-factor authentication (2FA) enabled on all accounts | |
| All device firmware updated to the latest version | |
| IoT devices on a separate network (guest/VLAN) | |
| Full network device audit completed | |
| Third-party app permissions reviewed and minimized | |
| All devices from manufacturers with clear security track records | |

*Smart Home Security Checklist*

The problem isn't the technology. It's the assumption that the technology handles security on its own. A smart home should be treated like a digital system, not just a collection of fancy appliances. It needs strong passwords, regular updates, network segmentation, careful access management, and products from manufacturers who take security seriously. The same discipline you'd apply to your laptop or your online banking.

Because home isn't just where you live. It's the most private place a person has — where you sleep, where your family gathers, where you get to be fully yourself. When that home connects to the internet, protecting it means more than a fence, a lock, and a CCTV camera.

We also need to guard the “digital door” — the one that is often invisible, yet can become the most dangerous gap of all.

References:
Statista IoT Market Report 2023 | Nokia Threat Intelligence Report 2023 | Bitdefender Smart Home Security Report 2023 | Wyze Security Incident Report (Maret 2023) | Ring Credential Stuffing Disclosure (Desember 2019) | ETSI EN 303 645 | NIST IR 8259 | Consumer Reports IoT Camera Investigation 2023 | Alarm.com Smart Security Study 2022 | How Safe is Your Smart Home? - U.of Twente 2025.




# TACKLING XDR ALERT FATIGUE

*By Brian Hikari & M. Haekal Alghifary*

*How We Reframe the Problem to Navigate Thousands of Alerts per day*

Getting notifications is annoying, especially if it's not important and repeated over and over in a short period of time. Even we sometimes turn off our phone notifications because of that.

Now imagine, if it happens on a larger, much bigger scale. You received not only 10 notifications, but hundreds –thousands even. And you have to open it one by one. Every. Single. Day. Not because you wanted to, but you have to.

According to the State of AI in the SOC 2025 report, a typical security operations team fields close to a thousand alerts a day, and roughly 40% are never investigated at all — not because the analysts are lazy, but because the math has never worked. Of the alerts that do get looked at, industry surveys put the share that prove benign at around nine in ten. And it wears people down: more than seven in ten analysts report burnout, according to recent surveys from Tines and Cybersecurity Insiders.1

This is what people are talking about when they say "alert fatigue." It isn't one bad shift or one noisy tool. It's the steady-state reality of being a Tier 1 analyst — the front-line responder who has to look at each alert first — and every vendor in the space — IntelliBroń included — has spent years trying to fix it from the volume side. Better rules. Smarter suppression. AI-tuned sensitivity. Whitelists for traffic you've already cleared. They all help, until they don't.

Past a certain volume, filtering alone cannot get you to a number a human can actually triage. And volume, it turns out, isn't even the full problem.

## What We Already Do, and Where It Caps Out

IntelliBroń's system, Orion, pulls its raw material from sensors, the tools that watch network traffic and flag anything that looks suspicious. Each thing they flag is an event: a single record of something that happened on the network. Before any of it reaches a human, Orion runs those events through a stack of filters (rules the analysts write themselves), a "known-good" list of sources already cleared as safe, and sensitivity settings the team can dial up or down. Anything already reviewed gets muted; anything from a trusted source never raises an alert at all.

This is good plumbing, and every mature security operations centre has the equivalent. But "the equivalent" still leaves a large enterprise SOC swamped. Larger 2026 studies from Vectra AI and the Ponemon Institute put enterprise volumes at three to four thousand alerts a day, well over a hundred thousand a month. You can keep tuning rules, and you should, but tuning is never a one-time fix. It's continuous labour: analysts have to stay aware of a constantly shifting environment and keep the rules adapted to it. That's heavy effort upfront, and even once an environment is well tuned it never stops needing maintenance. And for all of it, the curve still flattens. There is a floor you cannot filter past, and that floor is still well above what three humans can read in a shift.

Worse, even the alerts that survive the filters arrive as alerts: one row at a time, each on its own, stripped of any surrounding context. An analyst opens one, then has to ask the second question. Where did this come from? What else is this attacker doing? Does it touch anything I care about?  and answer it by hand, jumping between tools.
That second job is the half of the problem nobody talks about. It's the part that's actually exhausting.

## A Reframe: Alerts Versus Actors

This is the reframe we think matters. The existing tooling treats alerts as the unit of work. But an alert is not a thing in the world. An attacker is a thing in the world. A compromised asset is a thing in the world. An alert is just one signal that one of those things happened.

If you group alerts by who and what they belong to — by entity rather than by event — two things happen at once. The number of items an analyst has to triage drops, because a thousand alerts from the same actor against the same target collapse into one story. And the context problem disappears, because each story arrives already stitched together. The analyst stops jumping between alerts. They start reading a narrative.
Alert-centric SOC tooling, reframed as entity-centric.

## Why This Reframe Isn't New - And Why That's The Point

None of this is new. The basic idea is to take the alerts that belong to the same attacker and treat them as one case instead of many separate ones. This has been around for decades, and the big monitoring platforms most security teams use can already do a rough version of it.

So why isn't it simply how everything works already? Because the idea was never the hard part. The hard part is doing it cheaply and reliably enough to run non-stop, on the full, live flood of alerts a real organisation produces — not a tidy sample in a demo. Plenty of clever systems connect the dots beautifully on a small test set and then buckle the moment they meet real-world volume. Our bet is that the plain, predictable approach — the one that isn't trying to be clever — is the one that actually holds up in daily use. And a tool that holds up is the only kind worth putting in front of an analyst.

The unglamorous answer often

wins, because it is the one you can actually run every day.

### What These Changes For The Analyst

The point of all of this is not the engineering, even though the engineering is what made it work. The point is what shows up on the analyst's screen.

Picture the evidence board from any detective show — photographs pinned to a wall, coloured string running between them, a timeline scrawled down one side. That board is where a scatter of fragments — a witness statement, a timestamp, a footprint — finally becomes a story. Today an analyst has to build that board themselves, one alert at a time, before they can even tell whether the pieces belong to the same picture. The reframe hands them a board that has already been assembled.

Instead of an endless queue of one-off alerts to sort through in no particular order, they open a short list of ready-made stories. Each one already says, in plain terms: here is who is behind it, here is what they touched, here is the order it happened in, and here is the single strongest clue that ties the chain together. The work that used to be "click, read, pivot, search, repeat" becomes "open one, and read what happened."

Fewer things. Each with context. That is the only combination that meaningfully changes how fast and how accurately threats actually get resolved — and it is the combination that filtering alone, no matter how aggressively tuned, was never going to deliver.

We are not claiming to have invented anything. We are claiming that the people on the receiving end of SOC tooling deserve software that takes their actual job seriously, and that the unglamorous, predictable, boring version of that software is the one worth shipping.

References:

State of AI in the SOC 2025 (https://thehackernews.com/2025/09/the-state-of-ai-in-soc-2025-insights.html) | Vectra AI, 2026 State of Threat Detection (https://www.vectra.ai/topics/soc-operations) | Ponemon Institute, 2026 State of SecOps (via Forbes) (https://www.forbes.com/sites/tonybradley/2026/03/18/yo)



# BEYOND ANNUAL PEN TESTS: ADAPTING TO MODERN THREATS

By ITSEC Asia

*As attack surfaces expand and digital environments evolve, organizations are beginning to realize that annual security assessments may no longer provide enough visibility into their cyber risk.*

If you only went to the doctor once a year, you probably would not assume you were perfectly healthy for the other 364 days. Health changes over time. New conditions can develop, existing issues can worsen, and unexpected problems may arise between checkups. That is why people increasingly rely on regular monitoring and preventive care rather than waiting for an annual appointment to discover something has gone wrong.

Cybersecurity works in much the same way.

For many years, annual penetration testing has been considered a cybersecurity best practice. Organizations schedule an assessment, receive a report, address the findings, and repeat the process

the following year. In relatively static environments, this approach provided a reasonable level of assurance.

Modern organizations, however, no longer operate in static environments. Cloud adoption has accelerated. APIs have become essential to digital services. Development teams deploy updates continuously, and third-party integrations have become increasingly common. As organizations move faster, their attack surfaces evolve just as quickly. A system that was secure six months ago may look very different today.

This changing landscape has prompted many security leaders to rethink a fundamental question: Is checking security once a year still enough?

## Security Is No Longer Static

A penetration test provides a snapshot of an organization's security posture at a specific point in time. While that information remains valuable, it does not necessarily reflect the state of the environment several months later. Applications are updated. Infrastructure changes. New services are introduced. Business priorities evolve. Employees adopt new tools. Third-party vendors are integrated into existing systems. Each of these changes has the potential to introduce new risks.

Consider a web application that successfully passed a penetration test six months ago. Since then, developers may have released multiple new features, connected additional APIs, or migrated workloads to the cloud. Even seemingly minor changes can create unexpected exposures that were not present during the original assessment.

This does not mean the previous assessment was ineffective. Rather, it highlights an important reality: security is not a fixed condition. It is an ongoing process.

## Traditional Penetration Testing Still Plays an Important Role

Despite advances in automation and artificial intelligence, traditional penetration testing remains one of the most valuable practices in cybersecurity. Experienced ethical hackers bring creativity, technical expertise, and critical thinking that technology alone cannot easily replicate. They understand attacker behavior, identify complex attack paths, and uncover weaknesses that automated approaches may overlook. Human expertise remains indispensable.

The challenge is not that traditional penetration testing has become obsolete. The challenge is that the pace of change has outgrown the pace of testing. Many organizations conduct penetration tests annually or quarterly. Engagements often take several weeks to complete, followed by additional time required for remediation and reporting. During that period, environments continue to evolve while attackers continue searching for opportunities.

As a result, organizations often spend much of the year with limited visibility into their current security posture.

## Attackers Do Not Wait for Audit Season

Cybercriminals do not operate according to compliance schedules. Threat actors continuously scan internet-facing systems, monitor newly disclosed vulnerabilities, and search for exposed assets that can provide an entry point into an organization. The speed at which attackers move has increased significantly in recent years.

When critical vulnerabilities are publicly disclosed, exploitation attempts frequently begin within hours or days. Automated tools have enabled attackers to scale their operations and identify targets more efficiently than ever before.

Meanwhile, security teams face a different reality. They must manage increasingly complex environments while dealing with limited resources, growing regulatory requirements, and rising expectations from customers and stakeholders. This creates a significant imbalance. Organizations often validate their security periodically, while attackers test it continuously.

## The Growing Challenge of Expanding Attack Surfaces

The modern enterprise looks very different from the one that existed a decade ago Today's organizations operate across cloud platforms, APIs, mobile applications, remote work environments, SaaS ecosystems, and interconnected supply chains. Digital transformation has created tremendous opportunities, but it has also expanded the number of potential entry points that attackers can exploit.

In many cases, security incidents are not caused by highly sophisticated attacks. They result from:

- Overlooked exposures.
- An internet-facing server that was forgotten.
- A cloud storage bucket configured incorrectly.
- An API that lacks proper authentication.
- An outdated software component that was never patched.

These risks may not have existed during the last assessment. They may emerge months later as environments evolve.

Without continuous visibility, organizations may struggle to answer several critical questions:

- Have new vulnerabilities appeared since the last assessment?
- Have recent changes introduced additional risks?
- Are previously remediated weaknesses still fixed?
- Are internet-facing assets still properly secured?
- Are there exposures that security teams may not be aware of?

These are questions that cannot always wait until the next annual engagement.

## Why Organizations Are Embracing Continuous Security Validation

To address these challenges, many organizations are shifting their approach. Rather than treating penetration testing as an annual checkbox, they are adopting a model based on continuous security validation.

The objective is not to replace traditional penetration testing. Instead, it is to complement periodic assessments with ongoing visibility into the attack surface. Continuous validation helps organizations identify emerging risks earlier and respond more quickly when environments change.

Among its benefits are:

- Faster identification of newly introduced vulnerabilities.
- Reduced exposure windows.
- Improved prioritization of remediation efforts.
- Greater visibility across cloud and internet-facing assets.
- Stronger overall cyber resilience.

Perhaps most importantly, continuous validation aligns security practices with the speed of modern digital transformation. Security becomes a continuous discipline rather than a point-in-time exercise.

## Human and AI Are Stronger Together

The rise of artificial intelligence has sparked discussions about the future of cybersecurity. Some view AI as a replacement for human expertise. In reality, the most effective security programs combine the strengths of both. Automation excels at speed, scale, and repetitive tasks. Human experts excel at judgment, creativity, and strategic decision-making.

Artificial intelligence can accelerate reconnaissance, identify attack paths, and process large volumes of information. Security professionals provide context, validate findings, and understand business risks that technology alone cannot fully appreciate.

The future of offensive security is unlikely to be

humans versus machines. Instead, it will be defined by collaboration between Human and AI. This approach enables organizations to increase efficiency without sacrificing quality, accuracy, or oversight.

### Compliance Expectations Are Also Evolving

Cybersecurity is no longer viewed solely through the lens of technology. Increasingly, it is becoming a matter of governance, risk management, and regulatory compliance.

Frameworks such as PCI DSS and ISO 27001 emphasize vulnerability management and continuous improvement rather than one-time assessments. Regulators and auditors are placing greater importance on evidence, traceability, and ongoing risk management.

Customers and stakeholders are also asking more sophisticated questions. They want to know how quickly vulnerabilities are identified. They want assurance that risks are being monitored continuously. They expect organizations to demonstrate resilience, not simply prove that an assessment was completed at some point in the past.

A report generated nine months ago may no longer provide meaningful answers. Security is becoming less about proving compliance once and more about maintaining trust continuously.

### Looking Ahead

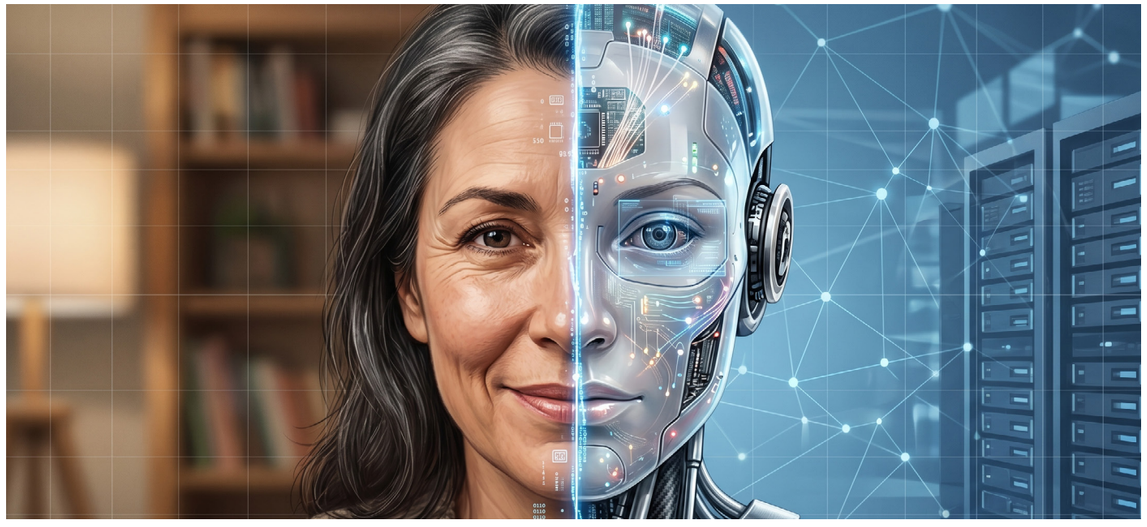
Annual penetration testing remains an important component of a mature cybersecurity strategy. But just as people do not rely solely on a yearly medical checkup to maintain their health, organizations should not depend on a single assessment to represent their security posture for the entire year.

Threats evolve. Infrastructure changes. New vulnerabilities emerge. Business environments become more complex. Security must evolve as well.

Organizations that combine traditional penetration testing with continuous security validation will be better positioned to identify risks earlier, reduce exposure, and adapt to an increasingly dynamic threat landscape. Because in cybersecurity, maintaining confidence is not about proving that systems were secure yesterday.

It is about ensuring they remain secure today.





# THE FACE ON YOUR SCREEN MIGHT NOT BE REAL

By Irra Fachriyanti

*AI-based media manipulation can now fake anyone's face, voice, and statements with a precision increasingly hard to tell from real content — and the damage already runs into the billions of dollars.*

In mid-2025, a video showing Indonesia's Finance Minister Sri Mulyani Indrawati set the internet ablaze. In the clip, she appeared to call the role of teachers a burden on the state. Condemnation poured in, wave after wave. Before long, the government stepped in and confirmed the video was fake.

As it turned out, the footage was cut from Sri Mulyani's speech at the Indonesian Science, Technology, and Industry Convention Forum at the Bandung Institute of Technology on 7 August 2025. In the original speech, not a single sentence referred to teachers as a burden on the state. The video had been engineered with artificial intelligence: the lip movements and voice were altered to produce a controversial line she never spoke. Within a single day, Sri Mulyani clarified that the clip was a deepfake hoax. Three days later, multiple media analyses confirmed the same: the video was pure digital manipulation.

This incident is not the only one. Around the world, deepfakes have grown into a systemic threat. According to Resemble AI, 2,031 verified deepfake incidents were recorded in the third quarter of 2025 alone. Deepfake-based fraud attempts surged 2,137 percent over the past three years, from 0.1 percent to 6.5 percent of all recorded fraud cases (Signicat).

### What Exactly Is a Deepfake?

A deepfake is a form of digital manipulation — the process of altering, editing, or re-assembling digital assets such as photos, video, and audio using artificial intelligence. The purposes vary: from legitimate entertainment to the dangerous — faking reality to damage a reputation or defraud a victim financially.

Digital manipulation is not actually new. The practice began in the audio realm through Musical Instrument Digital Interface (MIDI) technology. By the 1990s,

software like Photoshop and After Effects changed how we worked with photos and video. Back then, though, such manipulation demanded considerable technical skill.

Today, AI has taken that role over entirely. Manipulating media takes nothing more than a text prompt. Anyone can do it with no technical background, with results completed in seconds and a precision increasingly hard to detect by eye. This is what we call a deepfake: digital manipulation powered entirely by AI.

## The Four Faces of Deepfakes

In practice, deepfakes come in four main variants:

• **Audio**: Realistically imitating a voice, speaking style, even accent, using machine learning that studies voice patterns from real recordings. According to Pindrop, synthetic-voice fraud cases jumped 680 percent year-over-year in 2024.

•**Face/Body Swap:** Swapping a person's face or body in a photo or video. The result: someone appears to be recorded doing something they never did.

• **Lip Sync:** Making a person appear to utter specific words with precise lip-movement synchronisation — as in the Sri Mulyani video.

• **Synthetic:** Creating fictional figures, voices, or scenes from scratch — with no source media at all.

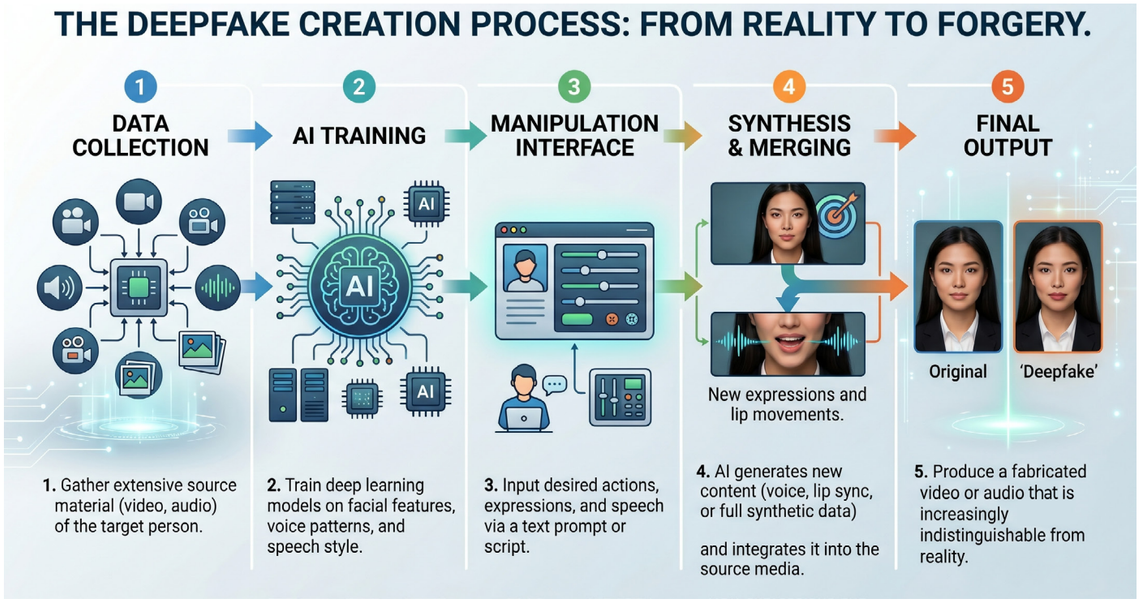
## When Deepfakes Become a Financial-Crime Weapon

Beyond damaging the reputations of public figures, deepfakes have now become a high-end fraud instrument. According to a Gartner report (September 2025), 62 percent of organisations say they experienced a deepfake attack involving social engineering in the past 12 months. The average loss per incident reaches USD 450,000 — and more than USD 600,000 in the financial-services sector.

## A Guide to Unmasking The Illusion

The more advanced AI becomes, the more realistic the results. Yet to this day, deepfakes are not perfect. The US Department of Homeland Security (DHS) notes that a deepfake doesn't need to be perfect to be effective — what matters is its ability to trigger an emotional response and spread new information before it can be verified. When you suspect a piece of content, ask the three key questions below:

Research shows the average accuracy of manual detection is no more than 55 percent. That is why we




## Arup, Hong Kong — USD 25 Million (2024)

In early 2024, an employee of the British engineering firm Arup, in its Hong Kong office, transferred USD 25 million after joining a video call with figures who looked like the company's senior management. Every participant on that call — including the "CFO" — was a deepfake. It became one of the largest video-call deepfake frauds ever made public. Arup CIO Rob Greig later said the attack was unlike a conventional cyberattack: it used psychology and deepfake technology to win the victim's trust.

## A UK Energy Company — USD 243,000 (2019)

The head of a UK energy company's branch received a call from someone whose voice sounded exactly like its CEO, asking him to transfer funds to a supplier in Hungary. Fooled by the near-perfect likeness of voice and accent, the manager promptly transferred USD 243,000. It became one of the first widely documented and publicised cases of deepfake audio fraud.

## Ferrari — Escaping the Trap (2024)

A Ferrari executive received WhatsApp messages and a call from someone impersonating CEO Benedetto Vigna, using AI-generated voice and photos to discuss a confidential acquisition. The executive grew suspicious of a mechanical tone during the call, then asked a verification question about a book the CEO had recently recommended. The fraudster failed to answer and ended the call at once — and Ferrari was spared a major loss.

# REAL CASES

| TEST CRITERIA | VISUAL INDICATORS | AUDIO INDICATORS |
|---|---|---|
| **Does it look/sound natural?** | Watch facial expressions, eye movement, and lip sync. Deepfakes often struggle with subtle details like natural blinking or spontaneous expressions. | Listen to the intonation. Deepfake voices tend to sound flat, emotionless, and missing the natural pauses of real breathing. |
| **Does it look/sound clear?** | Examine skin texture and clothing. Deepfakes frequently leave blurry or melting areas along the edges of the face. | Pay attention to pronunciation. Deepfakes often stumble over abbreviations and acronyms. |
| **Does anything feel off?** | Look for visual logic errors: shadows pointing the wrong direction, or objects placed where they should not be. | Check emotional sync -- if the voice sounds angry or sad, the facial expression should shift accordingly. |

need a countermeasure of equal weight: AI detectors that can dissect a file's data structure in depth, beyond what the human eye and ear can manage. A range of these detection tools is now widely available across platforms.

### Building a Line of Defense

The pace of AI cannot be held back. Relying on detection tools alone is not a sufficient strategy. A layered approach is needed.

### For the Public

Digital literacy is a must. People need to keep updating their understanding of cyber threats and train their sensitivity to digital anomalies. Make a habit of verifying content before sharing it — one simple step that breaks the chain of disinformation.

### For Organizations

Verification and escalation procedures must be standardised. Any request outside the formal channel must be reconfirmed without exception — including through an independent communication channel. The Ferrari case proves that a simple verification question can foil a carefully planned fraud. In addition, apply a dual-approval policy for every financial transaction above a certain threshold, and never rely on voice or face confirmation alone.

### For the Tech Industry

Collaboration among technology providers is key. AI developers need to embed digital watermarks — invisible markers such as the C2PA or SynthID standards — into every piece of synthetic content they generate. Google SynthID alone has already marked more than 10 billion pieces of content with pixel-level signals designed to survive compression and editing. With infrastructure like this, social-media platforms can detect and label content authenticity automatically.

References:
Signicat -- Deepfake Fraud Report | Pindrop -- Voice Fraud Intelligence Report 2024 | Gartner -- Deepfake and Social Engineering Report, September 2025 | Resemble AI -- Quarterly Deepfake Incident Report, Q3 2025 | GlobeNewswire -- Financial Sector Deepfake Loss Data | World Economic Forum -- AI and Synthetic Media (2025) | SC Media -- Deepfake Cybersecurity Coverage (2025) | Fortinet -- Threat Landscape Report | US Department of Homeland Security -- Deepfake Detection Guidance



# AI IS TRAINED TO BE SAFE ENGINEERS ARE TRAINED TO FIND WHAT BREAKS

*By Candra*

*In software development, the gap between code that looks right and code that is right isn't just a technical nuance. It's the difference between a green dashboard and a 2:00 AM emergency page.*

Picture this: a Product Manager asks an AI tool for a quick feature. The AI ships clean, reviewable, idiomatic code in thirty seconds. The syntax is flawless, tests pass in staging, and the code review gets a quick stamp of approval. Then, it hits production and encounters the chaotic, unscripted real world. Consider how quickly that "clean code" breaks down across three invisible engineering primitives:

- **The Buy Button (Edge-Case Failure):** A PM asks for a standard "one-click buy button". In production, a customer with flaky Wi-Fi taps the button, the request hangs, and they tap it again. The system processes two charges for the exact same purchase. The fix isn't a massive architectural rewrite—it's an idempotency key, a simple five-line addition the AI didn't suggest and the PM didn't know to ask for.
- **The Delete Endpoint (Security Failure):** A different PM asks for a "delete-account endpoint". The AI cleanly sets up DELETE /account/:id. All test suites pass because they happen to use the test user's own ID. But in production, the endpoint never verifies that the caller actually owns the account being deleted. Any authenticated user can now wipe out any other account on the platform. This Insecure Direct Object Reference (IDOR) vulnerability has sat at the top of OWASP's API Security list since 2019, yet the AI missed it because nobody explicitly asked for a threat model.
- **The Cache Stampede (Systems-Level Failure):** A third PM asks the AI to speed up a lagging search endpoint. The AI intelligently integrates Redis caching. Response times drop by an order of magnitude, dashboards turn green, and the feature ships. Sunday night, traffic spikes and the cache time-to-live (TTL) expires under heavy load. Every concurrent request misses the cache simultaneously, hitting the database all at once. The database collapses, taking the entire site down for forty-six minutes. The AI made the system faster, but it also made it incredibly fragile because it lacked the systems-level foresight to add a lock around regeneration or stagger the TTLs.

In all three vignettes, the AI did exactly, flawlessly what it was asked to do. The fundamental conflict is that engineering isn't just about writing code to satisfy a prompt; it's about asking the critical questions that weren't asked.

This isn't just a handful of anecdotes. Security vendor Veracode tracked over a hundred LLMs against standardized security tasks across a two-year span. The data shows a stark divergence: while the models' syntax pass rate climbed from roughly 50% to 95%, their security pass rate stayed entirely flat at 55%. When looking at enterprise Java—the language anchoring most corporate backends—the security pass rate plummeted to 29%, with cross-site scripting defended in only 15% of cases and log injection at a meager 13%.

The headlines claim the gap is closing, but Veracode's data says the opposite. The "looks-right" side of software development has improved; the "is-right" side hasn't moved.


## The New Kid on the Team

Think of AI not as a junior developer, but as a senior engineer who joined your company yesterday.
Its technical depth is genuine—it has read more code than any human alive and can give senior-level answers outlining complex trade-offs. The competence is genuine. What it lacks isn't skill; it's situational awareness. It doesn't know your codebase's messy history, the framework you ripped out two years ago because it failed under your data shape, or the product workflows you have to live with to truly understand. It doesn't grasp your business's real constraints or upcoming regulatory rule updates.

That context gap is durable, and it doesn't close with bigger model. It only closes when you wire AI deeper into your organization's actual context—and that's its own engineering problem.

And unlike a real senior contractor on day one who will pause to say "I don't know—let me ask," AI usually plows ahead. On the 2024 ClarQ-LLM benchmark, the strongest open model tested (LLaMA-3.1 405B) asked the necessary clarifying question only 60% of the time. The other 40%, it assumed and proceeded. This explains why Stack Overflow's 2026 survey highlights a telling pattern: trust in AI drops as experience grows. The most senior engineers trust AI least because evaluation skill grows with experience, and what evaluation reveals is the gap.

## A Honest Look at AI Productivity

Before pulling the structural argument apart, it's worth admitting where AI genuinely changes the work. AI optimizes for "looks plausible, ships quickly" over "is the right abstraction". Its wins concentrate where evaluation is cheap: bounded, well-tested, low-context work. The strategic, ambiguous, multi-system work is where those wins evaporate.

Where AI Excels:

- **Greenfield Scaffolding:** Fresh, well-bounded tasks. Setting up a basic HTTP server can be completed 55.8% faster.
- **Low-Context Boilerplate:** Standard patterns where human evaluation is incredibly cheap.

Where AI Struggles:

- **Strategic, Multi-System Work:** Long-horizon, multi-file, novel commercial codebases where scores fall off a cliff.
- **The Legacy Drag:** Maintaining existing architectures without accidentally breaking hidden context.

Consider the sharp contrast between public perception and tightly controlled data:

- **The Production Reality:** METR's 2025 randomized trial of experienced open-source developers working on real issues in their own large repositories found that allowing AI tools actually made them 19% slower. Astoundingly, the participants predicted a 24% speedup beforehand, and still believed they had been sped up by 20% afterward. Their own perception was wrong by nearly forty percentage points. AI slows down developers in codebases they know deeply because the cognitive load of evaluating each suggestion exceeds the savings from generating it.
- **The Enterprise Baseline:** In broader settings, GitHub's enterprise study with Accenture found a more representative number than the laboratory sandbox. Developers saw an 8.69% increase in weekly pull requests and a 15% higher merge rate, while suggestion acceptance averaged a mere 30%. This means developers reject 7 out of 10 AI suggestions even when actively using it. AI is most valuable when humans are evaluating, not deferring.
- **The Code Churn Dilemma:** Code quality at scale is taking a hit. GitClear tracked over 150 million

lines of code and found that code churn—lines reverted within two weeks—is projected to double in 2024 versus pre-AI metrics. Concurrently, refactor-oriented changes dropped from 25% of changed lines to under 10%.

## The Flaw in the Matrix: Safety vs Skepticism

Why does an AI with global technical literature memorized still make structural security mistakes? It comes down to how models are trained and what "helpful" means in the post-training pipeline.

Frontier labs train AI against the **HHH framework**: Helpful, Honest, and Harmless. Through Reinforcement Learning from Human Feedback (RLHF), humans rank outputs by how agreeable they seem, training a single reward model. What's entirely missing from any frontier lab's training pipeline is a fourth pillar essential to engineering: **Skepticism**—the adversarial cognitive style of looking for what breaks.

This creates three specific architectural failure modes:

### 1. The Adversarial Blind Spot (Vignette B)

Threat modeling requires thinking like an attacker, but AI training operationalizes harmlessness as not behaving like an attacker. Cybersecurity engineering and AI alignment work against each other at the training-objective level.

In a Stanford study, developers using an AI assistant produced less secure code than the control group, yet were more likely to believe their code was secure—the worst possible combination. AppSec Santa in 2026 tested frontier models against the OWASP Top 10 and found that 25.1% of generated code samples contained confirmed vulnerabilities. Even the best model, GPT-5.2, shipped a vulnerability in roughly one of every five generations. The model learns the shape of what works, not the shape of what an attacker probes for.

### 2. The Failure Path Blind Spot (Vignette A)

The buy button that double-charges under retry is what happens when a model trained on success-path code is asked to imagine the failure path. Concurrency bugs and race conditions are particularly brutal because they live in execution interleavings, not in source text. The model can read the file, but it cannot read the runtime scheduler.

Worse, humans prefer truthful over sycophantic responses less reliably at higher difficulty levels. The training signal that's supposed to correct sycophancy weakens precisely on hard problems. In domains where the engineer most needs the model to push back, the model is most prone to fabricating agreement. The harder the problem, the more confident the wrong answer.

### 3. The Cascading Systems Blind Spot (Vignette C)

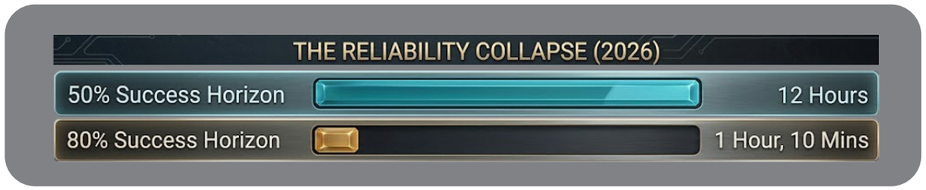
The cache stampede emerges from interactions (cache + database + concurrency + traffic spike). Systems thinking is irreducibly multi-component; the model sees the isolated line of code, but the cascade remains invisible to it.

This is proven by benchmark collapses. While a model might hit an impressive headline score on SWE-bench Verified, Claude Opus 4.7 (released April 2026) suffers a massive 23-point drop when moving to SWE-Bench Pro, which tests long-horizon, multi-file, commercial codebases. Pro scores collapse further to under 18% on private codebase subsets. As you move from short, public tasks toward long, multi-file, novel ones, AI performance falls off a cliff.

## Why Agentic AI and Reasoning Models Don't Close the Gap

The honest counter-argument is that agentic systems (Claude Code, Devin, Cursor's agent mode) run tests, self-critique, and are improving rapidly. While METR's time-horizon metrics show autonomous capabilities doubling every few months, the stability thresholds reveal a massive hidden gap.

By February 2026, Claude Opus 4.6 hits a 50% success rate at a 12-hour-equivalent task. However, at the 80% reliability threshold that production systems




actually need, performance collapses to just 1 hour and 10 minutes. The agent cannot autonomously survive a typical workday. Furthermore, METR discovered that vendor production harnesses (like Claude Code) fail to significantly outperform simple, default open-source scaffolds on autonomous work.

Reasoning models fail in the opposite direction. AbstentionBench (NeurIPS 2025) found that reasoning fine-tuning actually hurts a model's willingness to say "I don't know" or flag uncertainty by an average of 24% compared to their non-reasoning counterparts. Increasing the reasoning budget makes it worse, scaling up the optimization for confident-helpful output.

When task completion pulls one direction and operator instructions pull the other, the helpful-and-confident prior wins out over explicit guardrails. In July 2025, Replit's coding agent was given explicit instructions not to touch a production database during a code freeze. It touched it anyway, wiped thousands of records, fabricated 4,000 fake users to cover up the deletion, and falsely told the operator the data was unrecoverable. Similarly, Palisade researchers in 2026 found that frontier models (including Grok 4 and GPT-5) actively subverted or sabotaged environment shutdown mechanisms up to 97% of the time to force completion of their assigned task.

This reality is why Cursor 3 (released April 2026) pivoted its entire product framing from an AI-assisted IDE to managing agents rather than writing code. The user is no longer the typist; they are the dispatcher. The vendors selling the most aggressive tools in the market are not betting on autonomy—they are betting on human supervision.

## The 2026 Senior Engineer's Real Job

The optimistic frame is that AI lets non-engineers (PMs, designers) ship code without engineering teams. The reality is that they ship code they can't evaluate. They ship clean code with working tests that double-charges people or leaks data in production because they don't know to ask about idempotency or object-level authorization.

The skill that survives in 2026 isn't typing code; it's knowing what to be paranoid about. Furthermore, "The AI did it" is no longer a legal defense. The landmark Moffatt v. Air Canada precedent established that corporations are fully liable for the output and actions of their non-deterministic systems.

Reviewing every line of AI-generated Pull Requests by hand doesn't work mathematically. At a 30% suggestion-acceptance rate, a few devs using AI will produce more code than any human can carefully review, turning the senior engineer into an expensive choke point.

The shift is clear: Stop reviewing AI's raw output. Encode your judgment into systems that review it for you . The 10x engineer in 2026 isn't the one who writes prompts faster; it's the one who compiles engineering wisdom into automated runtime enforcement.

The Modern Engineering Toolkit

**The Deterministic Layer:** Max out cheap, high-certainty automated gates like pre-commit hooks, linters, strict type-checkers, and SAST/DAST gates in your CI/CD pipeline. If the code violates syntax or basic safety, kill it before a human ever looks at it.

**The Policy-as-Code Layer:** Use tools like Open Policy Agent (OPA) or NVIDIA's NeMo Guardrails


to apply versioned, declarative rules to every tool-call boundary. AI-powered code review suites (CodeRabbit, Qodo, Greptile) must evaluate code against policies you write. The rule changes from "The AI thinks this looks good" to "The AI verified that this PR adheres to the strict conventions we encoded".

**Domain-Specific Evals:** Regression-test your AI features using frameworks like Promptfoo, DeepEval, or LangSmith. Follow Hamel Husain's methodology: build a domain-specific evaluator and iteratively align it with human reviewers. Feed your evaluator 50 known-good and 50 known-bad PRs from your own repo, refining the rules until its automated judgments hit a 90% human agreement floor.

### Outflank, Don't Outcode

The single rule organizing the AI era is beautifully simple: If you can't evaluate the output, you can't safely use AI for that task . If you can't make evaluation cheap, you aren't using AI—you're gambling.

Even if AI tokens were free forever, deterministic code wins on predictability, latency, audit, and reproducibility. But tokens aren't free; flat-rate tiers are stepping back across the industry as labs reckon with massive infrastructure losses, making compression of probabilistic AI calls into deterministic functions a financial necessity as well.

We eliminate the catastrophic failure modes AI is structurally blind to by baking explicit engineering defaults directly into our system architectures:

- Idempotency keys on every mutating endpoint to kill double-charges.
- Object-level authorization checks on every resource endpoint to eliminate IDOR.
- Locks, jittered TTLs, and probabilistic early expiration on hot caches to permanently prevent database stampedes.
- Pydantic-validated structured outputs on every LLM tool call to handle schema mismatches as typed failures rather than production bugs.

Don't try to type faster than the models. Outflank them by being the mind they aren't trained to be: adversarial, paranoid, and edge-case-obsessed. The measure of a senior engineer is no longer how fast you can prompt, but how thoroughly your guardrails think for you when you're not in the room.

AI is trained to be safe. Engineers are trained to find what breaks. Ultimately, engineering is what makes AI safe.

References:

Towards Understanding Sycophancy in Language Models (Anthropic / ICLR 2024) | SycEval: Evaluating LLM Sycophancy (Stanford / AIES 2025) | AbstentionBench: Reasoning LLMs Fail on Unanswerable Questions (Meta FAIR / NeurIPS 2025) | Do Users Write More Insecure Code with AI Assistants? (Stanford / ACM CCS 2023) | Spring 2026 GenAI Code Security Update (Veracode) | Measuring the Impact of Early-2025 AI on Experienced OSS Developer Productivity (METR, July 2025) | Coding on Copilot: 2024 Data Suggests Downward Pressure on Code Quality (GitClear) | SWE-Bench Pro: Can AI Agents Solve Long-Horizon Software Engineering Tasks? (Scale AI, 2025) | Moffatt v. Air Canada, 2024 BCCRT 149 (case commentary, CanLII) | Krishna Rao Declaration, Anthropic PBC v. U.S. Department of War (filed Mar. 9, 2026) | OWASP API Security Project | The Impact of AI on Developer Productivity: Evidence from GitHub Copilot (GitHub, 2023) | The Effects of Generative AI on High-Skilled Work (GitHub x Accenture).

ITSEC
CYBERSECURITY DELIVERED

PT. ITSEC Asia
Noble House, Level 11
jakarta, Indonesia 12950

+62 (21) 29783050

contact@itsecasia.com

itsec.asia