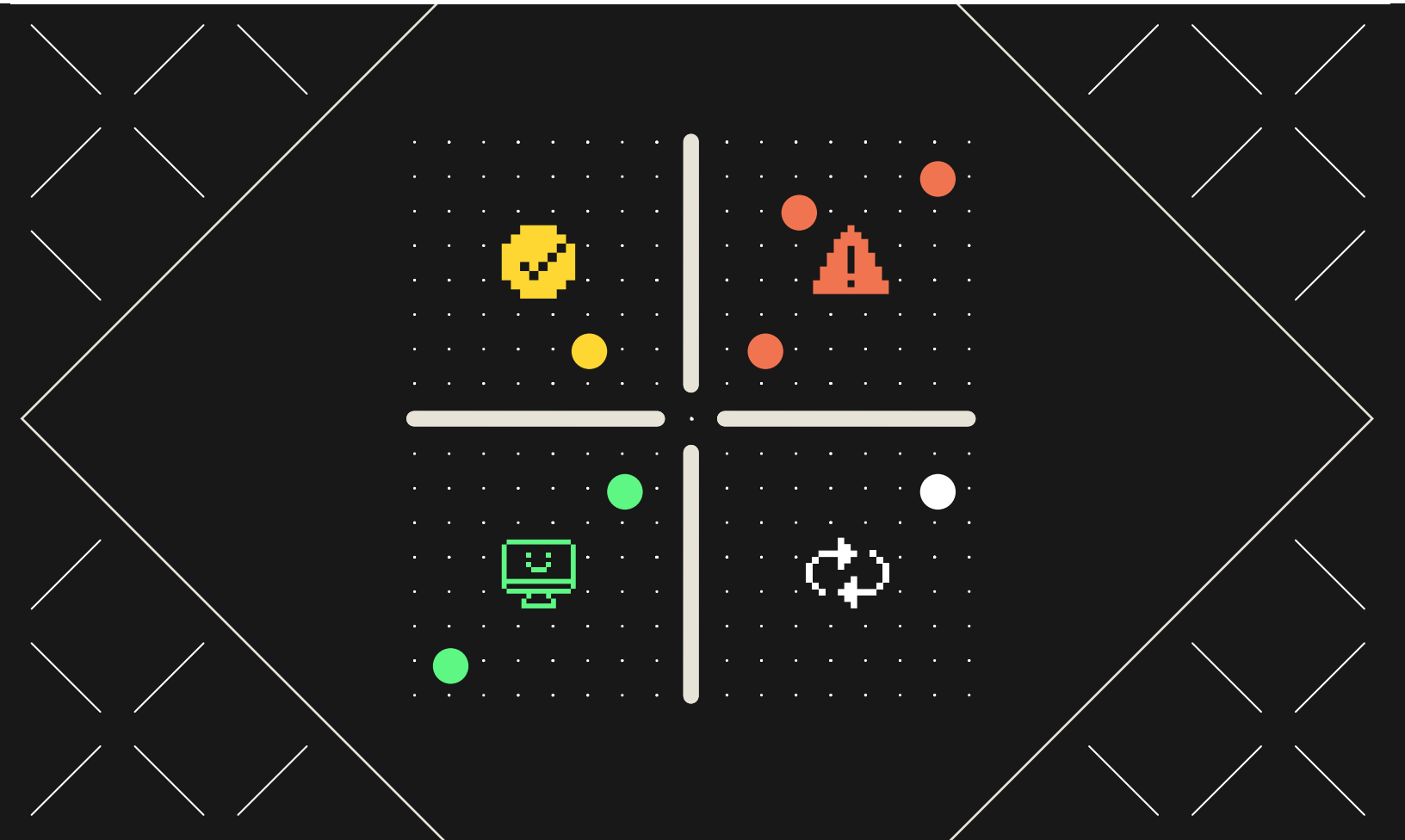
Trust vs. impact

A practitioner's framework for intentionally implementing AI and automation in the threat lifecycle



Contents

INTRODUCTION

The AI implementation problem isn't adoption. It's intent.

You can't talk to a security vendor today without drowning in AI hype. AI is everywhere, and "AI or die" is the expectation. But the real pressure for AI adoption isn't coming from vendors—it's coming from the top. Across the industry, CEOs are handing their CISOs the exact same mandate: "Use AI."

The truth is, we do need AI. Attackers are using AI too, to discover vulnerabilities and generate exploits, collapsing the weaponization window to mere hours. Defenders can't afford to ignore AI's use in keeping pace with the tempo that attackers have already adopted.

This creates a structural asymmetry: adversaries operate at machine speed, rendering human-speed defense a critical vulnerability. To survive this "AI vulnerability storm," building an AI-ready security program isn't just about operational efficiency anymore. It's an existential requirement to level the playing field.

We won't tell you AI is the answer to everything. Instead, we're sharing the framework we use internally at Expel, built from ten years of running AI and automation in production. Because every organization has a unique risk tolerance, team maturity level, and technology stack, you shouldn't consider this to be a rigid rulebook. Rather, this is a mental model to help you think critically about exactly where AI and automation belong in your security lifecycle, and where they don't.

But before you can map your workflows, you must rethink the executive mandate that got you here. It's time to trade the technology-first hype for an AI-intentional mindset.

Shift to an AI-intentional mindset

In our experience, security teams fall into two camps when it comes to AI adoption: those trying to be **AI-first**, and those striving to be **AI-intentional.**

The first camp has a mandate to be "AI-first," typically driven by understaffing, high alert volumes, or executive pressure. These teams deploy AI tactically across their security stack to automate tier 1 SOC duties—such as event correlation, triage, and disposition—to maximize operational throughput. However, teams are constantly balancing adding in or handing over a task to AI with risk. Being fast and wrong is still wrong. By prioritizing speed over holistic security, these teams accept a much higher-stakes risk profile than those taking a more deliberate, strategic approach to AI adoption.

This second group is what we consider “AI-intentional.” They work backward from the security outcome. These teams conduct honest capability assessments by outlining all the steps, tasks, and critical parts of the security program. They carefully weigh the risk-to-reward ratio of introducing AI, automation, or machine learning to each part of the program.

By doing so, they map exactly where AI models provide the highest efficacy, where deterministic automation is sufficient, and where human capabilities are needed. These are instances where judgment, creativity, and experience remain absolutely non-negotiable. They also establish a “you must be this tall” style threshold for those steps where they don’t feel AI is ready to be introduced. By doing so, they aren’t being a department of “no.” Rather, they are showing they have taken a measured and thoughtful approach to implementation.

You don’t need an AI-first mandate to survive the next generation of attacks. You need a framework to become AI-intentional. That’s exactly what this guide provides.

Our blueprint for intentional AI and automation

This framework helps you figure out exactly where AI and automation belong in your own SOC workflows, from alert to taking action. It’s built on ten years of knowledge running AI and automation within our SOC.

When we built Expel’s Managed Detection and Response (MDR) service, we understood that throwing more data, tools, and alerts at an analyst doesn’t make them faster. It actively prevents them from doing the one thing they are uniquely qualified to do—exercise critical judgment. To protect that human focus, we had to let machines do what machines do best: solve for speed and scale.

The framework is built on two variables: impact and trust. **Impact** measures what happens if the AI or automation gets it wrong. In other words, if the AI fails at the given task, how severe are the consequences? **Trust** measures how much confidence you have in AI or automation to handle a task correctly. But quantifying trust is far more complicated than most vendors will admit. We’ll get into that.

Our goal with the Trust vs. Impact Framework is to identify good candidates for AI and automation—not to prescribe what technical implementation to use within each process stage. Those decisions are best made within individual teams, as they are more familiar with their own environments and risk tolerances.

This framework isn’t theory. It’s what we learned the hard way—from testing, iteration, and real-world application.

Meet Ruxie

Ruxie is Expel's AI and automation engine. She's the unified system behind everything from investigation enrichment to AI-powered detection rule creation. Throughout this guide, we'll share examples of how Expel uses Ruxie to handle volume while our human analysts focus on the moments that matter most. She's been in production for ten years. The examples in this framework aren't hypothetical—they're real things we've implemented in Expel Workbench™ to enhance the work of Expel's SOC every day. Your use cases may vary, but the learnings still apply.

Not all AI is the same: Automation vs. AI vs. machine learning

These terms often get used interchangeably, however, they represent meaningfully different capabilities.

- **Automation:** Deterministic workflows that follow explicit rules. If X happens, do Y. Fast, reliable, and perfect for repetitive tasks that don't change based on context. Automation can leverage SOAR tooling or scripting with no AI required.
- **Machine learning:** AI models that identify patterns in data and make predictions based on training. Useful for classification, anomaly detection, and tasks where pattern recognition improves with data volume.
- **Generative AI:** Large language models (LLMs) that can reason, synthesize information, and generate content. These provide probabilistic content, not specific results. Valuable for summarization, analysis, and creating human-readable outputs from complex data.
- **Agentic AI:** AI systems that can autonomously plan and execute multi-step workflows, make decisions, and take actions with minimal human intervention. This is meaningfully different from automation. Agentic AI requires building and deploying actual AI agents, not just scripting deterministic rules. The infrastructure is different. The oversight requirements are different. The cost of getting it wrong is different.

MEET THE FRAMEWORK

The Trust vs. Impact Framework: A map for decisions you're already making

Integrating AI into a security operations center isn't an all-or-nothing proposition. We recommend iterative improvements by taking a strategic look at where it makes the most sense. Every decision about where to use AI and automation in your alert-to-action workflow requires you to understand the answers to two questions: how bad is it if this goes wrong, and how much do you trust the system to get it right?

Think about how you evaluate a new analyst. You don't hand a junior hire the same cases as your principal incident responder on day one. You assess the work's consequences and match it to the person's demonstrated capabilities. You expand their autonomy as they earn it. The same logic applies to AI and automation.

The framework involves mapping your workflow tasks on a 2×2 chart to understand the most applicable and effective places that AI and automation can improve your detection and response process. Let's discuss what this matrix is and how to use it.

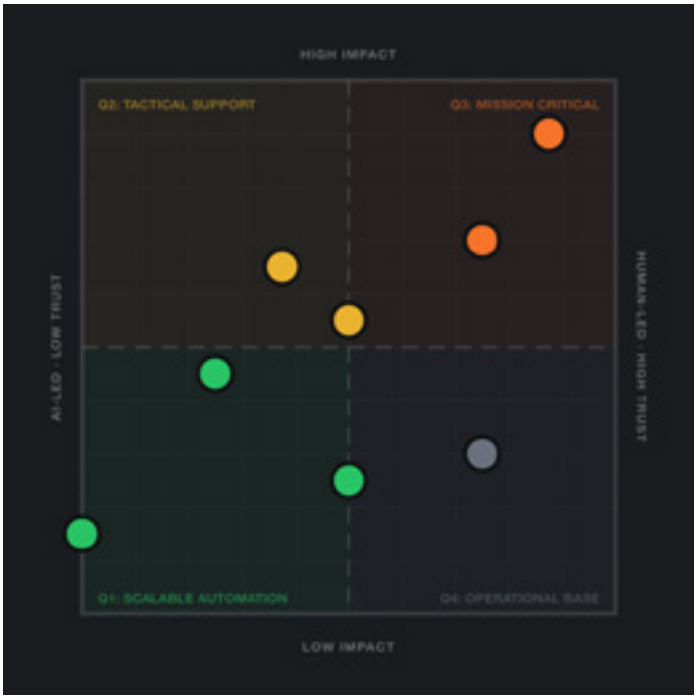


Figure 1: Tasks plotted against the Trust vs. Impact Framework, as featured in the Trust vs. Impact AI Assessment tool.

The Y-axis: Measuring impact, or the consequence of failure

The Y-axis, known as “impact,” measures the consequences of failure. If the task is performed incorrectly, or ignored completely, what’s the damage to the business?

On the high end of the impact spectrum, a failure would result in a major breach, significant financial loss, regulatory fines, or critical downtime. Examples of high impact actions could include isolating a domain controller or deciding to wipe a C-suite laptop.

On the opposite end, low-impact actions could include categorizing a phishing email subject line or closing duplicate tickets. A failure would result in noise, wasted analyst time, or minor administrative burden. Each can drain productivity, but they don’t pose a serious risk to the business.

The X-axis: Measuring trust, or need for human oversight

The X-axis, known as “trust,” measures the degree to which human intuition, nuance, and context is required to perform the task safely and accurately. Or, conversely, how mature and deterministic are current AI models and automation systems for this specific action?

On the higher end of the trust spectrum (human-led), are tasks requiring deep contextual understanding and nuanced decision-making, or they carry such high risk that a “human in the loop” is mandatory for accountability. In these circumstances, the AI model may be experimental or is prone to hallucinations for this specific task.

On the lower end (AI-led), these tasks are deterministic, repeatable, and computationally heavy. The outcomes are often binary (pass/fail), and current AI models have proven fidelity in this area.

Ruxie’s foundation: Ten years of doing the work

Ruxie started as automation workflows ten years ago, and they still run today without the AI processing costs. Additionally, we use AI capabilities where complexity requires it—where deterministic rules aren’t enough, where pattern recognition is needed to improve with data, and where synthesis and reasoning adds real value. Sometimes the old ways are the right ways. We build on top of them rather than replacing them.

USING THE FRAMEWORK

A practitioner's guide to placing your workflows on the matrix

This section walks you through three phases: establishing your human benchmark, overlaying AI capabilities, and building a trust progression plan.

Phase 1: Map what humans do today before thinking about AI

Before you place a single task on the matrix, you need an honest inventory of your current "alert-to-action" lifecycle based on how humans currently handle these tasks.

1. Deconstruct your lifecycle into discrete actions

Break your typical response workflow into specific, observable tasks. Every security operations center follows some variation of this lifecycle. The specific stages and terminology vary by organization, but the fundamental flow is consistent and may look something like this:

- **Alert generation:** Your security tools create alerts based on detection rules and vendor-provided indicators.
- **Initial enrichment:** Analysts (or AI) gather context. Who is this user, what does this behavior mean in this environment, and have we seen this before?
- **Triage:** Analysts assess whether the enriched alert represents a genuine security event or a false positive.
- **Investigation:** For genuine threats, analysts (and/or AI) gather deeper context, assess scope, and determine severity.
- **Containment decision:** Based on investigation findings, analysts decide how to respond.
- **Remediation:** Threats are neutralized through auto remediations or manual actions.
- **Communication and reporting:** Stakeholders are informed, the incident is documented, and detection logic is updated to catch similar threats faster next time.
- **Post-incident review:** Analysts review what happened, why, and what changes to process, detection, or tooling, would prevent recurrence.

2. Place each task on the matrix honestly

Place each task on the grid based on its inherent impact and the level of trust currently required to do it well. For example, you wouldn't ask a junior analyst to handle a complex DFIR engagement (high impact, high trust needed). Conversely, you don't want your principal incident responder performing initial alert triage (low impact, low trust needed).

Phase 2: Overlay AI and automation capabilities, but only where they've earned it

Now that you've mapped your current workflow, evaluate where AI and automation capabilities fit. This isn't about finding places to shoehorn in these capabilities. It's about identifying where AI and automation has demonstrated—or can demonstrate—the accuracy, transparency, and tunability to be trusted with a specific task.

1. Identify tasks that lead to low-value toil

Look at the bottom-left quadrant: low impact, lower trust required. These are your immediate candidates for AI and automation—the tasks that bore humans and can be handled by a machine.

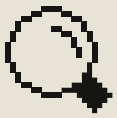
2. Identify tasks where AI can optimize for the human moment

Look at the top-right quadrant. These are high-stakes missions. AI and automation cannot be fully trusted here yet, but it can act as a hyper-efficient assistant to optimize the human effort.

Phase 3: Building a trust progression plan

This shouldn't be done as a one-time exercise. The Trust vs. Impact Framework is a living tool that evolves as AI capabilities mature, your team's expertise grows, and your organization's risk tolerance changes. Revisit it often.

Frontier models and AI products are maturing at a rapid pace. Tomorrow's applications of AI may be able to move the trust line higher. We recommend you set benchmarks for AI success the same way a human must meet criteria to get promoted at their job. You should consider that same "career ladder" for your AI/agents. Doing so will help with clarity across the organization, and help you articulate to vendors what you need to see in order to turn their AI features on.



REVIEWING THE RESULTS

Interpreting the four quadrants

Once you've mapped your tasks on the Trusts Impact Framework matrix, each quadrant defines your implementation strategy.

THE SCALABLE AUTOMATION ZONE

Quadrant 1: Low impact, low trust

This is where AI and automation provides the fastest return on investment by eliminating toil. Tasks are repetitive, the cost of failure is low, and models are mature enough to handle them reliably. That's a perfect workflow to automate.

Examples include initial alert triage, deduplication of events, parsing standard log formats, and basic threat intelligence lookups on IP addresses.

Ruxie in quadrant 1: Closing what humans shouldn't have to touch

Ruxie handles high-volume, high-confidence decisions autonomously, like alert deduplication, known-bad indicator lookups, initial enrichment, and auto-closing of benign alerts where the context makes the decision obvious. These workflows have run for ten years. They work because we defined exactly what 'obvious' means, and we've refined that definition continuously.

THE TACTICAL SUPPORT ZONE

Quadrant 2: High impact, low trust

This quadrant is rare. Usually, if a task is high impact, it inherently requires high trust. When tasks do land here, they represent critical steps that must be aggressively accelerated by AI and automation because they are severely bottlenecked by human speed, yet the output must be verified by a human before execution or sign-off.

The overarching goal shifts from manual toil to machine-speed execution with a human approval switch. For active threats, where adversaries move faster than manual defenders can react, AI use cases are autonomously investigating alerts and proposing a response playbook for a SOC analyst to run. Other use cases could be less critical, but

still important (like formal report writing), where AI can compress hours of log aggregation and retrace the attack path into seconds.

In both scenarios, the human's role is elevated to a critical approver: reviewing the AI's pre-computed proof or drafted narrative, verifying its accuracy, and authorizing the final execution. This dynamic is absolutely essential for actions like generating complex firewall blocks across multiple clouds, automatically quarantining endpoints, or summarizing findings and incident reports where accuracy is paramount.

> **Ruxie in quadrant 2: Ruxie proposes, you approve, then she moves**
>
> This quadrant is rare by design. When a task lands here, it's high enough impact to matter, and clear enough (or too risky) to automate. Ruxie generates the full action plan and surfaces it for human approval before executing. The speed is hers. The decision is made by experts. That's not a limitation. That's the appropriate division of labor when the stakes are high and the clock is running.

THE MISSION CRITICAL ZONE

Quadrant 3: High impact, high trust

In this quadrant, the stakes are simply too high for fully autonomous AI. This zone involves actions where an error could lead to severe operational downtime, massive financial loss, or significant reputational damage. Because these scenarios require deep organizational understanding and strategic judgment, AI cannot be allowed to operate independently.

Instead, this quadrant treats AI as a highly capable junior researcher supporting a senior expert. The AI does the heavy lifting of data processing: summarizing vast amounts of telemetry, correlating complex threat intelligence, and suggesting potential lines of inquiry based on its initial findings. It acts as a powerful force multiplier, presenting a complete case file to the human expert who can focus on the advanced analysis to make critical decisions.

Relying blindly on automation is dangerous here because AI inherently lacks the nuanced business context required to weigh the cascading consequences of a major action. Human operators must lead the engagement, verifying the AI's logic against their own expertise, institutional knowledge, and the specific, real-world constraints of their IT environment.

Ruxie in quadrant 3: The copilot, not the pilot

For high-stakes decisions like active incident response, containment of critical infrastructure, and communication with executive leadership, Ruxie operates in support mode. She synthesizes data, summarizes context, drafts communications for human review. The analyst makes the call. Every time.

THE OPERATIONAL BASE ZONE

Quadrant 4: Low impact, high trust

If a low-impact task is demanding significant human effort and a high degree of trust to verify, the problem is not a lack of AI. Instead, tasks landing in this zone are usually glaring indicators of bad telemetry, a fundamentally broken process, or a data source that cannot be trusted. The real question isn't how to use AI to automate this toil, but why we are tolerating the broken process or generating it in the first place. There are two rules of engagement for this zone:

1. Fix the process.

The mandate here is ensuring operational hygiene. If you rush to apply AI just to mask a bad workflow, you will end up institutionalizing technical debt rather than resolving it. Instead, you must step back and re-engineer the workflow from the ground up—or eliminate the useless task entirely—before ever attempting to introduce automation.

2. Make sure you have clean data to work with.

Even if a re-engineered process is a valid candidate for automation, such as using AI to auto-close benign alerts, its foundation must be reliable, normalized, and well-understood, otherwise it's going to add in more risk than it removes. AI inherently lacks a senior analyst's ability to navigate nuances and contextual blind spots in the telemetry. If you feed it flawed data, the AI will blindly execute its directive—clearing the queue—but potentially at the cost of making wrong calls.

Ruxie in quadrant 4: Fix the foundation first

Many of Ruxie's innovations initially started here and shifted across the trust axis once the underlying process or historical data made them ready to automate. For example, we introduced a fine-tuned, AI-assisted identity classification and triage capability to reduce the burden of noisy high-volume, low-severity identity alerts on our SOC team. Before we applied AI, our team did extensive work to ensure the inputs to the model were clean, reliable, and based on high quality log data and proper scope. By fixing the operational base first, we lowered the risk of misclassifying routine behavior and successfully shifted the workflow out of quadrant 4 and into quadrant 1, where scalable automation belongs.

REAL-WORLD EXAMPLES

What this looks like in practice: Ruxie across the threat lifecycle

Theory is useful. Examples are better. Here's how Expel applies the Trust vs. Impact Framework across the threat lifecycle, with specific capabilities Ruxie has in production today, and where they sit on the matrix.

Quadrant 1: Low impact, low trust

The scalable automation zone

This is where Ruxie handles volume at machine speed, with no per-action analyst review. The work is repetitive, the cost of any individual failure is low, and the logic is mature enough that autonomous execution is the right call.

Ruxie example: Identity alert classification

Identity alerts took our analysts 58% longer to work through than the average alert. So we trained Ruxie on a full year of Expel analyst decisions to classify incoming identity alerts the same way our best analysts would. Suspicious alerts get escalated to the top of the queue. Benign alerts—when Ruxie is 97%+ confident—get auto-closed with a full explanation of the decision. Our analysts spend time on real identity threats, not on verifying that a VP logged in from an airport lounge.

Ruxie example: Lead alert summaries

Before an analyst opens an alert, Ruxie has already pulled the logs, user history, and related activity into a plain-English summary. This ensures the human analyst is as efficient as possible, beginning their task with every detail at their fingertips. This guarantees well-informed decisions, without sacrificing speed.

Quadrant 2: High impact, low trust

The tactical support zone

This is where Ruxie does the heavy lifting on high-stakes work, but the output is reviewed before it takes effect. The speed is hers. The decision is the analyst's.

Ruxie example: Contain host auto remediation

When an analyst confirms lateral movement on an endpoint, Ruxie can sever the host's network connection in seconds. But she only acts on scenarios a customer has pre-approved in their Org Context, and only after an Expel analyst has validated the threat. Customers control the blast radius. Human analysts own the decision. Ruxie owns the speed.

Ruxie example: Draft or modify detections

When a novel vendor alert fires for the first time, Ruxie analyzes gaps in our existing coverage, decides whether to write a new rule or modify an existing one, and drafts the detection logic. Every rule then goes to an Expel detection engineer for review before it ever runs in production. Our engineers move from writers to editors, with Ruxie doing the YAML so they can spend their time on the logic.

Quadrant 3: High impact, high trust

The mission critical zone

The deliberate absence of Ruxie taking direct action in this quadrant is the point. In our most sensitive operations—active in-depth investigations, root cause analysis on novel attacks, high-stakes communications during a major incident—the analyst leads.

But Ruxie isn't absent. She's in support mode. The lead alert summaries, identity classifications, related-alert narratives, and user context she produced in quadrant 1 are already in the analyst's hands. The containment actions she can take in quadrant 2 are queued up, ready to execute the moment an analyst gives the word. The analyst doesn't spend the first hour of a critical incident gathering information. They spend it making judgment calls.

That's what the lower quadrants are really for: protecting the human capacity that this quadrant demands.

Quadrant 4: Low impact, high trust

The operational base zone

There's nothing to show here, and that's by design. If an Expel workflow ever lands in Q4—a task that's low-consequence but somehow still requires significant analyst effort—our instinct isn't to throw Ruxie at it. It's to re-engineer the workflow, fix the data source, or eliminate the task. Ruxie doesn't automate technical debt.

CONCLUSION

The goal was never autonomous AI, it was optimized humans

Most organizations are deploying AI reactively. And the result can become AI sprawl—lots of capabilities, minimal strategy, and analysts who are more overwhelmed than before.

Practitioners keep asking, *"Will AI replace human security analysts?"* But that's the wrong framing. The organizations winning right now aren't replacing their human security analysts with autonomous systems. They're replacing the work that was never worth an analyst's time in the first place: the tier 1 triage, the manual enrichment, the documentation, the boilerplate. What's left is the work that actually requires expertise. That's the work worth protecting.

The organizations that'll define the next decade of security operations aren't the ones with the most AI. They're the ones that use it most intentionally, with capabilities deployed precisely where they deliver value, with humans freed to focus on the work that actually requires human expertise. That's what the Trust vs. Impact Framework enables. Not AI for AI's sake. AI in service of better security outcomes.

Three principles that don't change as AI capabilities evolve

1. Match the capability to the requirement, not the hype cycle

Not every task deserves the same level of AI sophistication. Some need agentic reasoning. Some need simple automation. Some need humans. Resist the pressure to apply the most advanced available capability to every problem. It's expensive, it's fragile, and it usually makes things worse. The Trust vs. Impact Framework matrix helps you match capability to requirement instead of applying AI everywhere just because you can.

2. If automation solves it, use automation

Ruxie's oldest workflows are ten years old and they still run without AI processing costs. We only introduced AI capabilities where the problem genuinely required reasoning, synthesis, or pattern recognition that deterministic rules couldn't handle. If a scripted workflow solves the problem reliably, use that. Save AI for the complex stuff. Over-engineering is its own form of risk.

3. Optimize the end-to-end experience, not each step in isolation

By following the Trust vs. Impact Framework, you can apply automation, machine learning, and generative AI where each is appropriate across the full threat lifecycle. Don't optimize each step as its own isolated problem. Optimize the end-to-end experience for your analysts and your stakeholders.

Every decision in this framework comes back to the same principle: AI processes volume at machine speed. Humans apply expertise to the moments that genuinely require judgment, creativity, or accountability. The two capabilities aren't competing. They're complementary, but only if you deploy them deliberately.

Take action

Reading about AI strategy is one thing. Putting it to work is another.

The interactive Trust vs. Impact Framework matrix plots your own workflows to create a custom guide on where AI belongs in your SecOps.

START PLOTTING >