Voltade

Hire Your AI Agents.

Don’t Install Them.

Choosing the work, proving it, and granting autonomy

A framework for adopting and governing AI agents.

V1.0 | 5 August 2026
voltade.com

Contents

Executive summary 1

1 Hire, don't install 2

2 The value gap 3

3 The trust curve 4

4 Deciding who to hire 6

4.1 Write the job first 6

4.2 Could you manage a person in this role? 7

4.3 Decide what the AI agent may see 7

4.4 Before you sign anything 7

5 Onboarding and probation 9

5.1 Least privilege: how far the AI agent can reach 9

5.2 Masking: what the AI agent may see 10

5.3 Guardrails: checks on the way in and on the way out 10

5.4 Draft mode: probation on real work 11

5.5 Evals and coaching 12

6 Review and promotion 14

6.1 From passing score to first promotion 14

6.2 Watch the whole system, not just a sample of the work 14

6.3 Check that what you wrote down is still true 15

7 Managing the workforce 16

7.1 Redeploy, do not replace 16

7.2 The trap that grows as the AI agent improves 16

8 Where the analogy breaks 17

9 Your first ninety days 19

9.1 What to stop, and what to report instead 19

9.2 What this adds up to 20

References 21

Executive summary

Companies buying AI agents are reporting wildly different outcomes from the same technology. Gartner expects more than 40% of agentic AI projects to be cancelled by the end of 2027. Project NANDA found that 95% of organisations investing in enterprise generative AI reported no measurable return (Project NANDA, 2025), while BCG found a group of roughly 5% reporting 1.7 times the revenue growth of the companies it classes as laggards (BCG, 2025). These are self-reported figures from separate studies, so hold the exact numbers loosely. The spread is the finding.

The technology is the same for both groups. What separates them is that an AI agent acts on its own and can be confidently wrong in the specific way people are, which makes it something you hire and manage rather than something you install. Call it the trust curve: the autonomy you grant should track the evidence the AI agent has produced, task by task.

- **Write the job before you buy anything.** Pick one function with high volume, cheap mistakes and an agreed definition of good. If there is no written process, and nobody with the time to review a newcomer's work daily, a human hire would fail there too.
- **Decide what it may reach, and what it may see.** Give it the narrowest access that does the job, its own login, and no permission the person who switched it on does not have. Take personal details out of messages before they reach the model.
- **Start in draft mode, on real work.** The AI agent writes and a person sends. Skip the sandbox: it only tells you whether the AI agent can do a job you invented.
- **Write the passing score before it starts.** A hundred replies checked, eight in ten sent unchanged, fewer than one in ten thrown away. Promote one task at a time against that score, and keep reading a sample of what goes out.
- **Redeploy the people it frees.** Klarna let support headcount fall by roughly a fifth and was recruiting again fifteen months later.
- **Build the controls a person cannot be.** A named owner for every AI agent, a separate bar for each task, a hard gate on anything irreversible, a record anyone can read, and a plain statement to customers of what they are talking to.

Most of this is management you are probably not yet doing rather than technology you have to buy. The exception is the record: every job the AI agent does has to leave a trail someone can read, and without it none of the rest works.

This sequence lines up with the four things Singapore's governance framework for agentic AI asks of any organisation deploying one: bounded risk, accountable humans, controls built into the system, and informed users (IMDA, 2026a).

One question decides whether any of this is useful to you. What would an AI agent have to demonstrate before you let it write to a customer with nobody reading it first? Most companies buy one before they have an answer.

1 Hire, don't install

An AI agent acts on its own and can be confidently wrong. That makes it a hire, not an install.

In February 2024, one month after launch, Klarna said its AI assistant was handling two-thirds of the company's customer support chats. Average resolution time had fallen from eleven minutes to under two. Klarna said the AI was doing the equivalent work of 700 full-time customer service agents, and that it expected forty million dollars more in profit that year.

Fifteen months later, Klarna was recruiting human support staff again, and its story became a cautionary tale about AI.

Two decisions caused that, and neither was about the technology. Klarna gave the AI the whole queue a month in, including the cases that needed judgement. Then it stopped backfilling the support team as people left, so the ones who used to catch what the AI got wrong were no longer there. The CEO later said cost had become too dominant a factor and quality suffered.

The capability was real. The return was real. What Klarna lost was supervision, and supervision is not technology.

This is the mistake almost everyone makes with AI agents, and it starts before they buy one. They think they are installing software.

An AI agent is much closer to a new employee than to a new tool. It shows up capable but green. It needs its job defined and its early work checked by someone who knows what good looks like. When you correct it, it gets better. When you don't, it drifts. Treat it like an install and you end up where Klarna did. Treat it like a hire and it compounds.

Managing a capable newcomer is a problem every company has already solved. You define the job. You check the early work. You hand over more responsibility only once it has been earned. That playbook already sits inside your company, and it turns out to be most of what separates the companies getting a return on AI agents from the ones writing them off.

In a controlled trial, 758 management consultants were given work with and without an AI assistant. On the tasks the AI was good at, the ones using it completed 12.2% more work, 25.1% faster, and produced better output. On one task deliberately built to look the same but sit just past what the AI could actually do, they were nineteen percentage points less likely to reach the right answer (Dell'Acqua et al., 2026). Same people, same tool, opposite results. What decided it was whether the work fell on the right side of a line nobody could see. Finding that line, and moving it, is the job.

2 The value gap

The same technology is paying off for a small group and returning nothing to most of the rest. Four beliefs, all wrong in the same direction, explain the gap.

Some companies are plainly getting a return on AI agents. The useful question is what they do differently.

The failure data answers it. Gartner's named causes are unclear business value and inadequate risk controls, with costs that climbed until someone asked what the company was paying for. Deloitte found that about four in five enterprises lack mature governance for AI agents, including clear boundaries for what an AI agent may decide alone and what needs a human sign-off (Deloitte, 2026). Read those together and a picture forms: the AI agent had no job description and no supervisor, and nobody had decided what good looked like or when to check.

The install mindset produces those failures on schedule, and it comes bundled with four beliefs.

It should work out of the box. Software does, so people expect it. Then the AI agent fumbles a refund request in week one and the project loses the room. But nobody expects a new hire to be good on day one. The first weeks of any job are for learning the products, the customers, the tone, the unwritten rules. An AI agent has to learn the same things, and the only place it can learn them is inside your company, on your work. A company that budgets nothing for this has skipped onboarding and called it a rollout.

A mistake means it is broken. When software misbehaves you file a bug and wait for a patch, so when an AI agent gets something wrong, the instinct is to switch it off. With an AI agent, the mistake is tuition. It is the raw material the system improves on, the same way a junior's botched first draft is. Switch the AI agent off at its first error and you have paid the tuition and refused the lesson. Deploy with high hopes, catch one visible mistake, quietly shelve the project: that is the shape of a great many of those cancelled pilots.

Configuration is a one-time event. You set up software once and it stays set up. An AI agent drifts. Your policies change and your customers find new ways to be unusual, and an AI agent left alone falls slowly out of step with both. Managing an AI agent is ongoing work, the way managing a person is. Companies that staff for a rollout and then disband the team are always surprised by month four.

More autonomy is better. This is the expensive one. If the AI agent is good, why not let it do more? Because autonomy is trust, and trust has to be earned against evidence. Klarna's version: it granted maximum trust within a month and discovered that the hard tail of support was exactly where the AI agent hadn't earned it.

Underneath all four sits one error. The companies writing off their AI agents granted either total trust on day one or none after the first stumble, and neither amount was ever connected to anything the AI agent had actually demonstrated. Connect those two and the picture changes. Autonomy that tracks evidence is what the 5% have and the others do not, and that connection has a shape you already know.

3 The trust curve

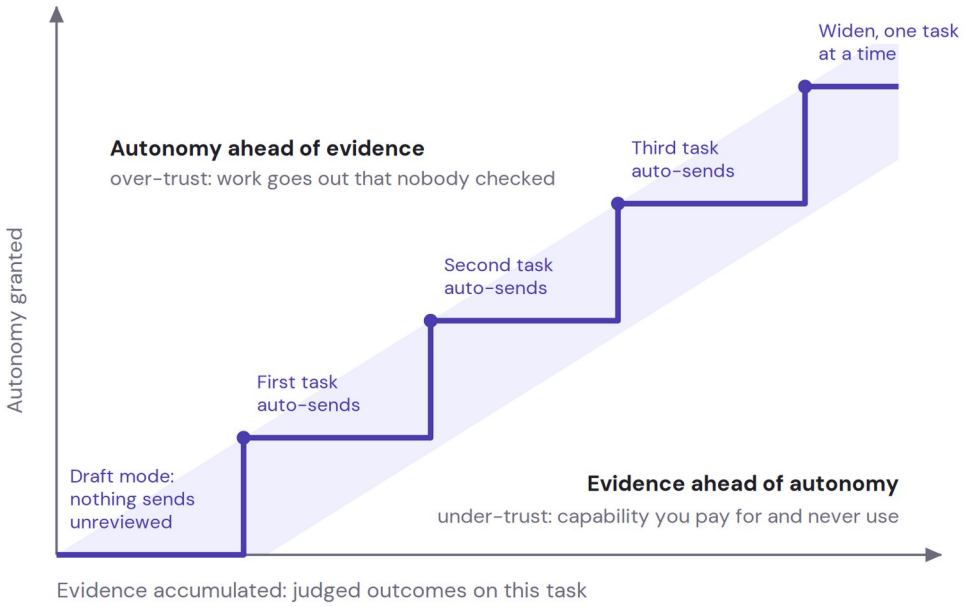
Autonomy granted should track evidence produced, task by task. Grant it ahead of the evidence and you get burned. Hold it far behind and you strangle the value you paid for.

An AI agent's first week on the job should look like a new hire's first week: real work, approved by a supervisor before it goes out. Concretely, the AI agent drafts every reply, and a human edits or approves each one before it sends. A month in, the drafted replies it consistently gets right go out on their own, while anything sensitive or unfamiliar still routes to a person. By the end of probation, the team reviews samples and exceptions rather than everything, because a few hundred approved drafts have shown exactly which work the AI agent can be trusted to send unsupervised.

Plot that trajectory and you get a curve: autonomy granted on one axis, evidence accumulated on the other (Figure 1).

Figure 1 The trust curve

Autonomy granted should track the evidence the AI agent has produced, task by task.



The shaded band is calibrated trust: autonomy roughly level with evidence. The line moves in steps, not smoothly, and it restarts at zero for every new job you give the AI agent.

Human-factors research calls this trust calibration: reliance matched to what a system has actually demonstrated (Lee & See, 2004). Miscalibration fails in two directions. Relying on automation beyond what it has earned is misuse; rejecting automation that has earned its keep is disuse (Parasuraman & Riley, 1997). Regulators are arriving in the same place. Singapore's IMDA names "automation bias, or the tendency to over-trust an automated system, especially when it has performed reliably in the past" (IMDA, 2026a). Both name the same failure: trust that grew because nothing had gone wrong yet.

Every failure so far is one of the two. Handing an AI agent full autonomy on day one is misuse: trust granted with no evidence behind it. Shelving the project at the first visible mistake is disuse,

and so is the sandbox pilot that never ends, just slower and more expensive. Klarna managed both inside eighteen months, over-trusting at launch and over-correcting after.

Two properties of the curve matter in practice. First, it moves in steps, per task. You trust a new account manager to answer routine emails long before you let them negotiate discounts, and the two trusts move independently. An AI agent's autonomy works the same way: sending replies, issuing refunds, changing customer records and posting publicly are separate grants, each with its own evidence bar. Second, the curve must be able to fall. A hire who starts sending erratic replies gets pulled back into review, and you need to be able to do the same to an AI agent, only faster. If autonomy cannot be taken back, do not grant it.

Run the curve across a whole company and it becomes the transformation plan: five steps, in order. First, decide which job to hire an AI agent for, and whether you are ready to manage one at all. Next comes onboarding and probation, where the AI agent does real work under supervision. Then review and promotion, where autonomy widens only as far as the evidence supports. After that, the team it joins and how their work changes. Underneath it all sits governance, including where the employee comparison breaks down and stricter rules take over.

The curve starts earlier than most people expect: at the job posting, before anything has been switched on.

4 Deciding who to hire

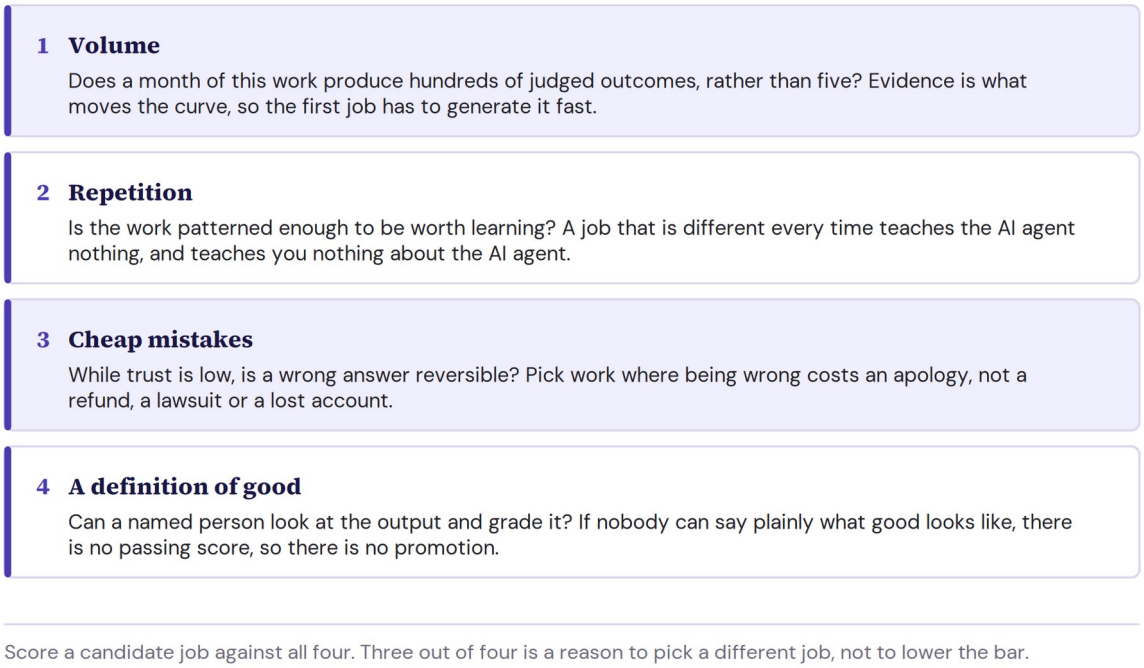
Write the job before you buy anything. Then test whether you could manage a person in the role, decide what the AI agent may see, and hire one rather than ten.

4.1 Write the job first

The worst job description in the world is "do AI for us." Yet that is the brief most AI agent projects start with: a mandate from the board and a budget, but no job. No company hires a person this way. You write the role first: what the person will do and what good looks like in month one. Hiring an AI agent starts in the same place, and the companies that skip this step become Gartner's cancellation statistic.

The trust curve tells you what to look for in a first role. The curve rises on evidence, so the ideal first job produces plenty of it quickly, and lets someone judge whether each piece was any good. Score a candidate job on four properties: high volume, repeating patterns, cheap mistakes, and a clear definition of good (Figure 2).

Figure 2 What to look for in a first role



Score a candidate job against all four. Three out of four is a reason to pick a different job, not to lower the bar.

In practice this points to the same handful of roles in most companies: customer support replies and triage, internal question-answering over company documents, data entry and enrichment, and first drafts of anything produced in volume, such as quotes, follow-ups, summaries and reports. It points away from anything low-volume and irreversible, such as payments and legal commitments.

4.2 Could you manage a person in this role?

Ask three questions about your own company. Is there a written process for this work? Is there an agreed definition of a good outcome? Is there someone with the time to review a newcomer's output daily? If any answer is no, a human hire would flounder there, and an AI agent will too, faster. The failed pilots share a learning gap, and it sits in the tools: systems that do not retain feedback or adapt to context (Project NANDA, 2025). The gap runs both ways. A company that cannot absorb what the AI agent learns fails it just as reliably. What decides this is whether the function is ready to supervise anyone new. A function that could not onboard a person is not ready to onboard software that behaves like one.

4.3 Decide what the AI agent may see

Which of your information is sensitive, and who is allowed to see each kind? Most companies have never written this down.

Sort it into a few levels: information anyone in the company may see, information limited to one team, and information that would do real damage if it reached someone outside. Customer contact details, pricing, contracts and payroll each sit at different levels. Each level then gets its own rule about where it may be sent and who may ask for it.

Your levels are your own decision. The law is not. Personal data about customers and staff carries obligations under Singapore's Personal Data Protection Act whichever level you put it in, and three of them bear directly on an AI agent (PDPA, 2012).

- **Purpose Limitation and Notification.** You may use personal data only for purposes you told people about. Support history collected to answer a customer is not automatically available to train an AI agent.
- **Transfer Limitation.** Personal data sent outside Singapore has to keep a comparable standard of protection. A message passed to a model provider abroad is such a transfer.
- **Data Breach Notification.** Significant breaches have to be reported to the Commission and to the people affected. That includes the placeholder list from masking, which is a store of exactly the data you removed.

Set your levels first, then have someone check each level against those three.

4.4 Before you sign anything

Spend your time configuring an AI agent to your company, not building one from scratch. Good ones already exist, and what decides whether this works is your policies, your tone and your definition of a good reply. None of that gets easier because your own engineers wrote the software.

So the question is who to buy from. Gartner estimates that of the thousands of vendors now selling "AI agents", only around 130 are actually selling one; the rest are offering chatbots and scripted workflows under a new label. The install mindset is what makes the relabelling work, because if an AI agent is just software, software is easy to pass off as one. Ask what you would ask about a candidate: what decisions can it make on its own, and what has it done unsupervised? A chatbot has an answer to neither.

Then there are the questions whose answers you should be shown rather than told:

- Which personal details are removed before a message is sent to the AI provider?
- What is stored, and for how long?
- If the software the AI agent connects to changes, does it carry on using it, or stop and wait for approval?

A vendor who can only tell you and not show you is asking you to trust a contract instead of a control.

Some of it you will only ever get verbally, so know what a good answer sounds like. It is specific and slightly unflattering. "It stops and asks" is a good answer. "It handles that automatically" is a bad one, because the next question is what happens when the automatic handling is wrong. Expect a named limit, a named person or system that catches the failure, and at least one thing the vendor says the AI agent should not be trusted with yet. A vendor whose answers have no edges has not run this in production.

Finally, hire one. The urge is to launch a fleet of specialised AI agents at once. Berkeley researchers ran seven open-source multi-agent frameworks and measured task-failure rates between 41% and 86.7%, and sorting those runs by cause produced three groups: how the system was designed, AI agents talking past each other, and nobody verifying the result (Cemri et al., 2025). None of those are model failures, and adding AI agents multiplies all three. The restraint has positive evidence too: given the same reasoning budget, a single capable AI agent matched or beat multi-agent systems on multi-hop reasoning tasks, and the multi-agent designs only regained ground once the single AI agent's grip on its own context started to slip (Tran & Kiela, 2026). One well-managed hire teaches you how to manage ten. Ten unmanaged hires teach you why Gartner's number exists.

The role is written and the candidate is chosen. Now comes the part everyone underestimates: the first day of work.

5 Onboarding and probation

Decide what the AI agent may reach and what it may see. Then put it on real work in draft mode, where a person reads everything before a customer does.

5.1 Least privilege: how far the AI agent can reach

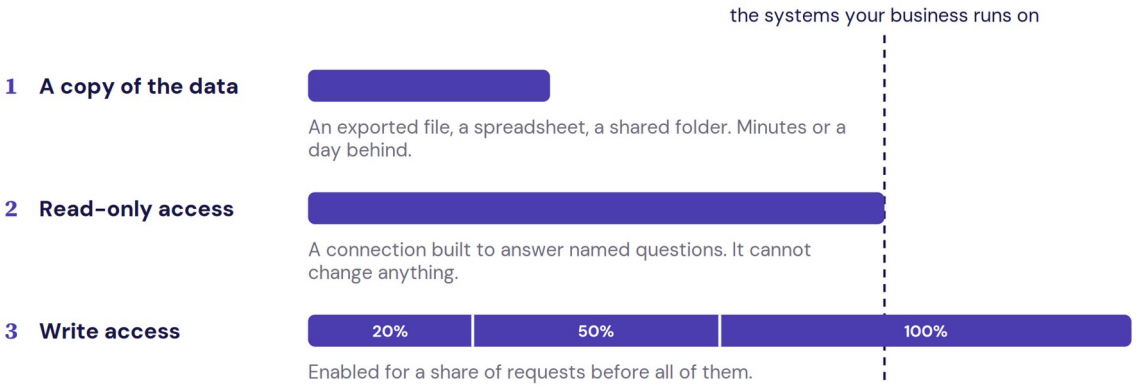
Authority lives outside the AI agent. Enforce permissions in the systems themselves, and never rely on the model to police itself (OWASP, 2025). Telling the AI agent "do not issue refunds" is a request. Removing the refund permission from its account is a rule. Write requests for tone. Write rules for money.

Onboarding an AI agent starts with access, and the rule is restriction: it gets exactly the systems its job needs, and nothing else. Security calls this least privilege: give an AI agent "least-privilege access to tools and data" and nothing more (IMDA, 2026a). The framework gives the principle. The three steps below are one way to implement it. An HR AI agent reads HR documents. A support AI agent reads tickets and the product catalogue. Neither touches the finance system, because neither job requires it. Give the AI agent its own login rather than letting it use a person's. Every action it takes is then recorded against it, and you can tell its work apart from anyone else's. Three rules go with that. Each AI agent has a named owner. Every AI agent you run is listed in one place, so nobody loses track of how many exist. And an AI agent never holds a permission the person who switched it on does not have themselves (IMDA, 2026a).

The instinct is to connect the AI agent to the systems that run the business and then write rules to keep it in line. Do it the other way round. Start with the narrowest access that lets it do the job, and widen only when the record supports it. Call it the access ladder: a copy of the data, then read-only, then write (Figure 3).

Figure 3 How far the AI agent can reach

Least privilege decides which systems. This decides how far into them.



Grant one step at a time, and only when the record supports it. Many AI agents never need the third. Write access is the last grant, not the first (IMDA, 2026a).

Start with a copy of the data rather than a connection to the system. The AI agent then cannot corrupt the records your business runs on, and cannot overload the system that holds them. An AI agent sends far more requests, far faster, than the person whose work it took over. For a

company's first AI agent, a copy is usually enough. Product details, price lists, policies and past enquiries do not change by the minute.

Move to read-only access when the AI agent needs current data. That means a purpose-built connection, an API, that answers specific questions: the status of this order, the balance on this account. The AI agent can retrieve what it needs and nothing else, and it cannot change anything.

Write access should be enabled last. Changing live records is the final grant, not the first, and it should not arrive all at once. Enable it for a fraction of requests, with the rest still proposed for approval, then widen. Many AI agents never need this step at all.

5.2 Masking: what the AI agent may see

Access decides which systems the AI agent can reach. What it is allowed to see inside them is a separate decision.

Whatever the AI agent is working on gets sent to a model, and that model usually belongs to someone else. Most of the time the personal details in a message are not needed for the work. A customer asking about a delayed order needs the order looked up. The model does not need the customer's phone number to do it.

So take out what is not needed before the message is sent. This is called masking, or redaction. Phone numbers, email addresses, identity card numbers and account numbers are replaced with placeholders before the message goes out, and put back in the reply.

Structured identifiers can be replaced reliably. Names and addresses written into free text cannot, so treat this as reducing exposure rather than ending it. The list that maps each placeholder back to its real value is itself a store of exactly the data you removed, and it needs the same protection.

Which details get removed changes from job to job. An HR AI agent handling a leave request may need the employee's name. A support AI agent answering an order query almost never does.

5.3 Guardrails: checks on the way in and on the way out

One combination deserves a hard rule. Security researchers call it the lethal trifecta (Willison, 2025). An AI agent that can do all three of these at once is a data-leak risk that no written instruction can reliably close:

- Read private data.
- Take in content from outside your control.
- Send messages out.

A support AI agent has all three by default. The fix is to remove one of the three rather than to watch it, and each one behaves differently.

- **Sending is the one you can remove.** In draft mode the AI agent has no way to send anything at all. Having a person approve each reply is not the same thing: a reviewer will pass a reply that reads perfectly well and quietly carries data inside a link, and replies are not the only way out. A tool call or a write to a customer record carries data out too.

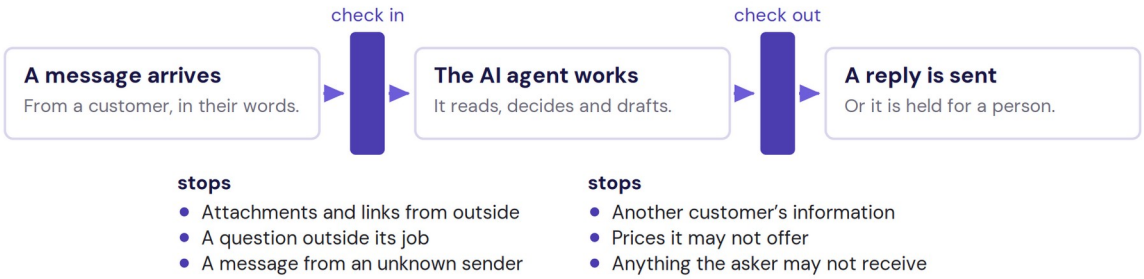
- **Reading private data is the one you can narrow.** Masking removes personal details from the message before the AI agent sees them, and the access ladder decides which records it can reach at all.
- **Taking in outside content is the one you cannot remove.** That content is the customer's message, and reading it is the job. No check reliably spots an instruction hidden inside one.

A reviewer does a different job here. A reviewer can judge whether a reply was good, which is what probation needs. Judging whether a reply quietly carried something out is a different job, and not one a person does reliably.

The same principle covers the checks most vendors call guardrails. A guardrail on the way in decides whether a message reaches the AI agent at all. A guardrail on the way out decides whether a reply is sent (Figure 4). Ask which of the two a vendor means, because they are not the same control and only one of them stops a bad reply.

Figure 4 Checks on the way in and on the way out

One decides what reaches the AI agent. The other decides what reaches the customer.



Write a fixed rule wherever the thing you are checking for can be written down. Use a second AI only for what cannot, such as tone, and remember it can be wrong in the same way the first one can (IMDA, 2026a).

Some checks can be written as a fixed rule: a price above a limit, a reference number that does not match the sender, a list of words that must never appear. A rule does the same thing every time, and you can test it.

Other things cannot be written down that way, like tone, or a claim that is true but misleading. Those need a second AI to read the reply before it goes. That works, but the second AI can be wrong in the same way the first one can, so use it only where a rule will not do (IMDA, 2026a).

5.4 Draft mode: probation on real work

Then the real work starts, and with it the AI agent's probation period. During probation the AI agent runs in draft mode: it does the actual job, but everything it produces arrives as a draft for a person to review before anything reaches a customer (Figure 5). Two kinds of work stay behind human approval permanently, however good the AI agent gets: anything irreversible, like sending money or deleting records, and anything it has not seen before. Actions that are "sensitive, irreversible, or have high stakes" should trigger human oversight (OpenAI, 2025). Autonomy is earned on the repeatable middle of the job.

Figure 5 Probation: draft mode

The AI agent does the real job. Nothing reaches a customer without a person reading it first.

Probation is the expensive part, and what it costs is your people's time rather than the licence. Every draft gets read, so the review load tracks volume: a hundred replies a day is a hundred reads a day. At the start each read takes longer than writing the reply would have, because the reviewer is judging rather than typing. That falls as the drafts improve, and it never reaches zero, because the spot-checks after promotion carry on permanently. Price the first six weeks as a second pair of hands on that function, not as a software purchase. If nobody on the team has those hours, the honest answer is that you are not ready to start yet.

Here is the advice most companies find hardest to take: skip the sandbox. A sandbox is a sealed test environment where the AI agent answers invented cases for a few months before touching anything real, and it feels like the responsible route. But the sandbox fails at the one thing the probation period is for. It cannot supply your real customers or their real ambiguity, and it produces no comparison. In draft mode, every AI draft sits next to what your team actually sent, so you can see, case by case, where the AI agent already matches your people and where it does not. A sandbox tells you it passed tests you invented. And draft mode carries almost no customer risk, because the worst outcome is a bad draft a person discards. So run the pre-launch checks, then put it straight onto real work in draft mode rather than into a sandbox. Testing in isolation postpones evidence. Probation on real work manufactures it.

5.5 Evals and coaching

Test before the AI agent touches real work, even in draft mode. These tests are what the industry calls evals: a fixed set of cases with known good answers, run against the AI agent before launch and again whenever something changes. Check that it completes the task, follows the process, uses the right tools with the right permissions, and behaves when something unexpected happens (IMDA, 2026a).

Tests cannot tell you when to grant autonomy. A test checks that the AI agent said a particular thing, and customers do not write in particular ways, so at volume the tests flag large numbers of replies that were perfectly fine. Chasing those failures means maintaining tests instead of reading work. Testing shows the AI agent is safe to switch on. The record of real work shows what it has earned.

Coaching turns the drafts into improvement. Every edit a supervisor makes before sending is a lesson: the distance between what the AI agent wrote and what your company actually says. A well-built AI agent collects these corrections and proposes updates to its own instructions, and the updates go through a person before they stick. That last clause matters. An AI agent that rewrites its own memory unsupervised can be poisoned by a single malicious message it then remembers forever. The AI agent does the learning. You govern what sticks.

The system around the model matters more than the model, and the part you own is the scoring. A benchmark can mark its own answers; an inbox cannot, until probation gives it something to mark against. The approved and rejected drafts are that. The same candidate, onboarded well, is a different employee.

By the end of the probation period you also have a file: hundreds of drafts, approved or corrected, showing exactly what this AI agent can be trusted to do.

6 Review and promotion

Set the passing score before the AI agent starts. Promote one task at a time, keep watching after it goes automatic, and check that your written rules still match the system.

The file from probation tells you when the AI agent can work without a check.

6.1 From passing score to first promotion

Set the passing score before the AI agent starts. Write it down plainly: "once we've checked 100 replies about order status, and at least 80 went out unchanged, and fewer than 10 were thrown away, this task can run on its own." Your team already checks every draft during probation, so all you do is keep two tallies per task: how many they sent as-is, and how many they rewrote or threw away. Choose a count that suits your volume, high enough that a slow week does not swing the number. When a task hits the score, it has earned the right to send on its own. The count assumes the person doing the checking was reading. If a task gets promoted on a tally that was not really earned, the spot-checks after promotion are what catch it. That is why they are not optional.

Turn on auto-send one task at a time. Let replies about order status go out on their own while refunds still wait for a person, since every task has to hit the score by itself. Define "significant checkpoints or action boundaries that require human approval" (IMDA, 2026a). The checkpoint is the task, not the AI agent. Put one named person in charge of the switch for each task. Promoting a task hands back the leg draft mode took away, and the checks on the way out cannot replace it: a fixed rule will not spot an instruction hidden in a customer's message, and nor will a reviewer. So work on what the AI agent can see instead. Before promoting a task, make sure it reaches no private data that task does not need. If quality drops, flip the task back to review the same day.

Where you cannot narrow that, the task stays in draft mode whatever it scores.

Keep checking the work after it goes automatic. Have a supervisor read one in every ten sent messages, permanently, plus every message a customer complains about. Quality can slip without warning, and these spot-checks catch it before customers do. Anything that slips goes back to review. Once the first job runs smoothly on this routine, you are ready to expand.

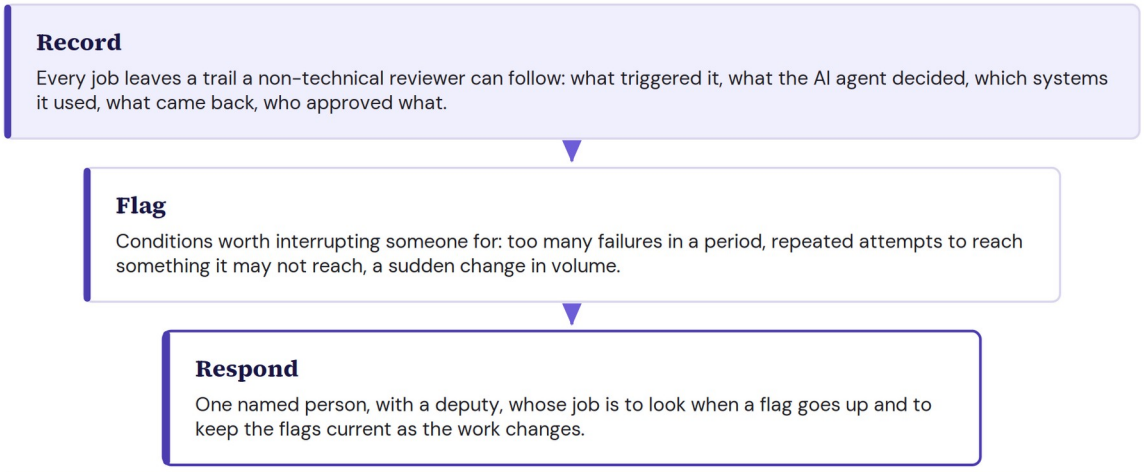
Repeat for every new job. When support runs stable, give the AI agent sales follow-ups with a fresh probation period and a fresh passing score, because being good at refunds proves nothing about selling. The second round runs faster, mostly because your team now knows how to set passing scores and read the counts. That management skill is the durable asset.

6.2 Watch the whole system, not just a sample of the work

Spot-checks catch work that is slowly getting worse. They cannot catch a fault that appears at nine in the morning and reaches four hundred conversations by lunchtime. That needs monitoring, which is three things rather than a dashboard: a readable record, conditions that raise a flag, and a named person who responds (Figure 6).

Figure 6 What to watch once the AI agent is live

Each one is useless without the one above it.



Nobody reads every record, and nobody should. Decide the response to each flag in advance, and make at least one of them stop the AI agent rather than notify somebody. If the person who should approve cannot be reached, the AI agent waits.

Start with the record. Every job the AI agent does should leave a trail someone can read without being technical: what it was asked, what it decided to do, which systems it used, and anything that failed (IMDA, 2026a). If that does not exist, nothing else here is possible.

Set the conditions that raise a flag. More than a set number of errors in an hour. Repeated attempts to reach something the AI agent is not allowed to reach. A sudden change in how often it is being used. Decide in advance what happens for each flag: some get looked at the same week, some stop the AI agent until a person has checked.

One rule is worth setting now. If the person who is supposed to approve something cannot be reached, the AI agent waits. It does not proceed because nobody answered (IMDA, 2026a).

6.3 Check that what you wrote down is still true

Everything above depends on a written description of what the AI agent may do: the job description, the passing score, the list of what needs approval. That description will drift from the system, and it always drifts the same way. The document describes more human control than the system actually enforces, because controls get relaxed to move faster and nobody goes back to update the document.

A named owner, a passing score and a sample of sent work all rely on that description being accurate. If it is out of date, all three are checking the wrong thing.

So audit the description itself. Once a quarter, take what you have written down about what the AI agent may decide alone, and check each line against what the system actually permits. Checking that your oversight still works is already standard advice (IMDA, 2026a). The quarterly audit goes one step further. Your people may be checking carefully against rules the AI agent stopped following months ago.

7 Managing the workforce

Redeploy the people the AI agent frees rather than removing them, and watch for the cost that only arrives once the AI agent is good.

7.1 Redeploy, do not replace

An AI agent works alongside people, and how you treat those people is where most transformations go wrong. Over 2024 Klarna let headcount fall by roughly a fifth, not by cutting but by not replacing the people who left (Klarna, 2025). Fifteen months after launch it was recruiting again, because the AI agent could not handle the hard cases. Attrition without backfill feels like the careful version of a headcount cut. It removes the same people, only slower, and you do not get to choose which ones.

So redeploy people, do not replace them. The AI agent takes the routine work, and the time that frees up has two uses.

The first use is obvious: the hard cases the AI agent cannot handle, the difficult customers where deals are won or lost. Secondly, freed from answering the same simple questions all day, your team can finally do the relationship work nobody had time for. They can reach out to customers before problems start, and give your best accounts the attention they never used to get.

The AI agent helps your least experienced people the most. In a large study of support staff given an AI assistant, productivity rose 15% on average, but the least experienced workers resolved about 30% more cases while the best performers barely moved (Brynjolfsson et al., 2025). It brings your junior staff closer to the level of your senior staff, and lets your senior staff spend their time on the judgement calls and the relationships that need a person.

7.2 The trap that grows as the AI agent improves

When the AI agent is right almost every time, the people checking it stop paying attention. A supervisor who has approved 100 good drafts in a row barely reads the 101st. This is the over-trust from the trust curve, arriving quietly: people stop watching automation exactly when it becomes reliable, which is when its rare mistakes slip through. Researchers call it automation complacency (Parasuraman & Manzey, 2010). A sample you drop because quality rose is a sample you dropped exactly when it started to matter. The most dangerous AI agent is a good one that nobody is watching any more.

Complacency is measurable, so you do not have to guess. Two numbers show it. The first is how often your reviewers change or reject what the AI agent produced: if that falls close to zero, they have most likely stopped reading rather than run out of things to fix. The second is how long they spend before approving: if that keeps dropping, approval has become a reflex (IMDA, 2026a). Watch both alongside the quality of the work, because a reviewer stops reading before anything visibly goes wrong.

Done this way, your people do better work and catch the AI agent's mistakes. But the employee comparison only stretches so far. When an AI agent causes real harm, it cannot be held responsible the way a person can, and the responsibility lands on you.

8 Where the analogy breaks

Five ways an AI agent differs from an employee, and why each one has to be built into the system rather than watched for by a person.

First, accountability. Liability does not move to the AI agent. Whatever you owed a customer, an employee or a regulator before, you owe when an AI agent does the work: data protection duties, your contracts and your sector rules have no exception for automated action. "The AI did it on its own" is not a defence. The organisation that deploys an AI agent, and the people who oversee it, remain accountable for what it does (IMDA, 2026a). Three companies are usually involved: whoever built the model, whoever built the AI agent, and you. When something goes wrong, each can reasonably point at the other two. That is why the accountable person has to sit inside your own company rather than in a contract with either of them. Every AI agent needs a named human owner.

Second, it is one brain, not a headcount. A hundred employees make a hundred separate judgements, and one person's bad habit stays with that person. An AI agent is one system, copied. One flaw reaches every customer it touches at the same time. Its mistakes are systemic by default.

Third, it fails in ways no person would. A capable employee's skill is even: if they can handle the hard cases, the easy ones are safe. An AI agent's competence is uneven. It can be excellent on one task and confidently wrong on a nearly identical one, and it sounds the same either way. In the consultant trial cited earlier, on the one task built to sit just past what the AI could actually do, the people using it were nineteen percentage points less likely to reach the right answer (Dell'Acqua et al., 2026). Watching an AI agent do one job well tells you nothing about the next one. This is why the trust curve moves task by task rather than all at once: competence really does stop at the edge of a task.

Fourth, speed. A person makes mistakes one at a time. An AI agent can make the same mistake ten thousand times before lunch. This is why the irreversible work never graduates. Probation ends and autonomy widens, and moving money still waits for a person, because the cost of being wrong does not fall as the AI agent improves. It multiplies.

Fifth, your customer cannot judge it. Everything so far has been about calibrating your own trust using evidence: probation, spot-checks, a record of what the AI agent got right and wrong. Your customer gets none of that. They meet the AI agent cold, and the instincts they use on people do not transfer. When a person is unsure, it usually shows. An AI agent gives the same impression either way. That is uneven competence again, now on the customer's side of the conversation. What you tell them upfront is the only evidence they will ever have.

A customer-facing AI agent should tell people "what the chatbot can and cannot do, how reliable and safe it is, how user data is handled and how to report issues" (IMDA, 2026b). Add how to speak to a human, since the hard cases already go there. Declare it at the point of interaction, in the interface itself rather than only in a document (IMDA, 2026a), and link from there to an info card carrying the rest in plain language, updated as the AI agent changes. In a business-to-business account there is no interface to put it on, so it becomes a line in the account manager's next call and a paragraph in the supplier terms.

Governing these five gaps does not require a single new control. Four things have to be true of any company running AI agents, and Singapore's framework for agentic AI asks for the same four (IMDA, 2026a). Here is where each one sits in what you have just read.

Every one of the five fails in a way that attention cannot catch in time. One system, copied, gives you no warning case. Uneven competence does not announce itself. Ten thousand actions do not wait for someone to look up. So they get built rather than managed: a named human owner, a per-task trust bar, a hard gate on anything irreversible, and a plain statement of what the AI agent is. Build them once and they hold on the days nobody is watching.

What has to be true	What you already built
The risks are bounded before anything runs	The job description, the data levels, and the access ladder.
A named person is accountable	The owner of each task, the passing score, and the quarterly audit of your own written rules.
The controls are built, not watched for	Draft mode, the checks on the way in and on the way out, and the record of what the AI agent did.
The people it serves know what they are dealing with	The declaration at the point of interaction, and the info card behind it.

9 Your first ninety days

Ninety days from a written job description to one task running on its own.

The trust curve compresses into a quarter. Four stages, and the dates are less important than the order.

Days 1 to 15: write the job description. Pick one function with high volume, cheap mistakes, and a clear definition of good. Name the person who owns the AI agent's output, by name and not by team. Write down what it may decide alone and what it must escalate. Sort your information into three levels: what anyone in the company may see, what stays inside one team, and what would do real damage outside it. Until that exists you cannot say what the AI agent is allowed to see. If you cannot write that list, you have just found the reason the last pilot failed.

Days 16 to 45: probation. The AI agent drafts, a person approves, and every approval and rejection is recorded. Do this on real work in production, not in a sandbox. Expect the first fortnight to look poor. That is what onboarding looks like.

Days 46 to 60: the performance review. Set the passing score before you look at the numbers, not after. Count the judged outcomes, the share of drafts sent unedited, and the share that went wrong. Compare against the passing score you wrote down, not against what you were hoping for.

Days 61 to 90: the first promotion. Turn on autonomy for one task, the narrowest one that cleared the passing score, and leave everything else in draft mode. Keep reading a sample of what goes out unsupervised, permanently. Then start the cycle again for the second task.

9.1 What to stop, and what to report instead

Four habits to avoid.

Stop piloting in sandboxes. A sandbox measures whether the AI agent can do a job you invented. You need to know whether it can do yours, on your customers, in your tone.

Stop buying on the demo. A demo is a candidate's polished answer to a question they knew was coming. Ask what the AI agent has done unsupervised, and who was accountable when it went wrong.

Stop reporting adoption. Seats filled and messages sent tell you nothing about whether the work was any good. The number that matters is how much work goes out unsupervised and stays correct.

Stop counting headcount saved. That is the metric that produced Klarna's year. It makes an AI agent look successful right up to the moment quality gives way. Count the work that got done, and the work your people moved on to.

Then report something in their place. Three numbers fit on one slide and answer what a board is actually asking:

- **Per task, the share of work going out unsupervised.** Not one average across the AI agent, because an average hides which tasks earned it.

- **Of that unsupervised work, the share found wrong on spot-checks.** And what happened to each one.
- **What the people it freed up are doing now.** If the answer is nothing, the AI agent has not paid for itself yet.

If you cannot fill one of these in, say so. "We do not measure that yet" is a better report than a seat count.

9.2 What this adds up to

Everything here reduces to one habit. Autonomy tracks evidence, and a named person is accountable for both.

That habit is what governance is: the job description, the passing score, the record of what the AI agent did, and the person who answers when a flag goes up. Each of those exists because it makes the AI agent more useful, and each is also what a regulator asks to see.

An AI agent that has earned its autonomy is worth more than one that was handed it.

Voltade builds AI agents for businesses, mostly on WhatsApp and email. This is what we have learned deploying them in production.

References

BCG (2025). *The Widening AI Value Gap.* Boston Consulting Group, September 2025.

Brynjolfsson, E., Li, D. & Raymond, L. (2025). Generative AI at Work. *Quarterly Journal of Economics*, 140(2), 889 to 942.

Cemri, M. et al. (2025). *Why Do Multi-Agent LLM Systems Fail?* arXiv:2503.13657, version 3, 26 October 2025.

Dell'Acqua, F. et al. (2026). Navigating the Jagged Technological Frontier. *Organization Science*, published online 11 March 2026. DOI 10.1287/orsc.2025.21838.

Deloitte (2026). *2026 State of AI in the Enterprise.* January 2026. 3,235 IT and business leaders across 24 countries.

Gartner (2025). *Press release, 25 June 2025.* Over 40% of agentic AI projects to be cancelled by the end of 2027, and agent washing among vendors selling AI agents.

IMDA (2026a). *Model AI Governance Framework for Agentic AI.* Version 1.5, published 20 May 2026, updated 5 June 2026. Infocomm Media Development Authority, Singapore. First published 22 January 2026 as Version 1.0.

IMDA (2026b). *Transparency Guidelines for Generative AI Chatbots.* Published 20 July 2026. Infocomm Media Development Authority, Singapore.

Klarna (2024). *Press release, 27 February 2024.* AI assistant handling two-thirds of customer service chats.

Klarna (2025). *Form F-1 Registration Statement.* Klarna Group plc, filed with the US Securities and Exchange Commission, 14 March 2025.

Lee, J. D. & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1), 50 to 80. DOI 10.1518/hfes.46.1.50_30392.

OpenAI (2025). *A Practical Guide to Building Agents.*

OWASP (2025). *Top 10 for LLM Applications.* LLM06: Excessive Agency.

PDPA (2012). *Personal Data Protection Act 2012.* Republic of Singapore. Obligations named as the Personal Data Protection Commission names them: Purpose Limitation, Notification, Transfer Limitation, Data Breach Notification.

Parasuraman, R. & Manzey, D. (2010). Complacency and Bias in Human Use of Automation. *Human Factors*, 52(3), 381 to 410. DOI 10.1177/0018720810376055.

Parasuraman, R. & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors*, 39(2), 230 to 253. DOI 10.1518/001872097778543886.

Project NANDA (2025). *The GenAI Divide: State of AI in Business 2025.* Challapally, A., Pease, C., Raskar, R. & Chari, P., July 2025. Distributed as a PDF; no publisher landing page.

Tran, K. & Kiela, D. (2026). *Single-agent versus multi-agent systems under matched reasoning budgets.* arXiv:2604.02460.

Willison, S. (2025). *The Lethal Trifecta for AI Agents.*

Hire Your AI Agents.
Don’t Install Them.

V1.0 | 5 August 2026

voltade.com