

CUSTOMER EXPERIENCE DESIGN & INNOVATION

The Agentic CX Build Blueprint & Guardrails

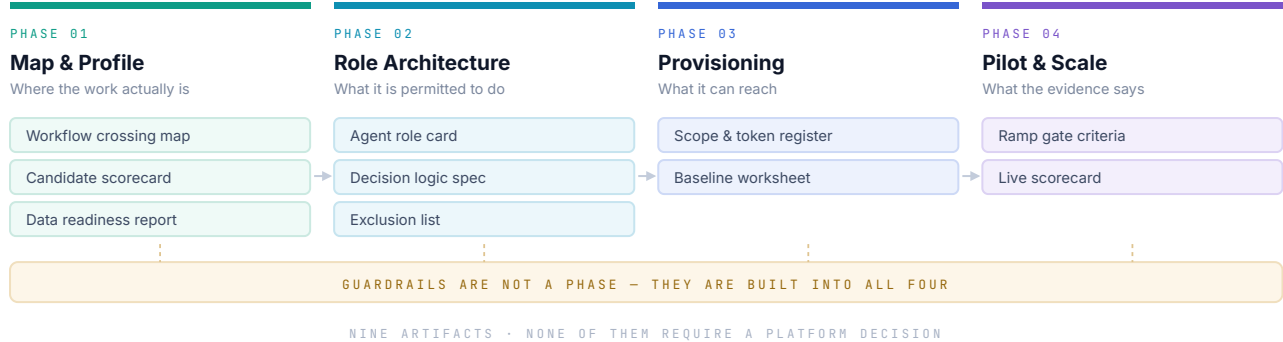
The first paper made the case for designing agentic systems. This one builds one — four phases, the artifacts each phase must produce, and the guardrails that decide what ever reaches a customer.



Nine artifacts, or it is not a build

Ninety-five percent of task-specific enterprise AI tools never reach production. The reason is almost never the model. It is that nobody wrote down what the agent may do, nobody cleaned the data it would have to trust, and nobody decided who says stop.

FIGURE 1 • THE BUILD, AND WHAT EACH PHASE MUST HAND OVER



The artifacts are the deliverable, not the agent. Each one is a document a person owns and signs. A program that cannot produce all nine has not built an agent — it has configured one and hoped.

How to use this paper

Take one workflow, not a department. Run it through all four phases, produce all nine artifacts, and put it in front of five percent of your traffic. The entire sequence is designed to be completed in ninety days by a team of four.

Everything here is drawn against a single worked build — an e-commerce retailer handling order cancellations and refunds across Zendesk, Shopify and Stripe — so that every abstract step has a concrete artifact attached to it.

What it deliberately skips

Model selection, orchestration framework and vendor. Those choices matter and they are not the constraint. Every failure mode in this paper survives a change of model.

Named tools are illustrative. Substitute your own stack; the shape of the work does not change. External figures are attributed where they appear and listed at the back.

5%

of task-specific enterprise AI tools reach production

MIT NANDA, 2025

72%

of service teams name data readiness as a major blocker

SALESFORCE, 2026

>40%

of agentic AI projects canceled by end of 2027

GARTNER, 2025

70%

of adopters see measurable value within 60 days

SALESFORCE, 2026

A chatbot answers. An agent acts.

The line is not conversational quality. It is whether the system holds a goal, chooses its own path to that goal, and writes to something. Everything in this blueprint follows from that one difference — because the moment software writes, it needs permission, provenance and a way back.

FIGURE 2 • SIX DIMENSIONS THAT SEPARATE GENERATIVE FROM AGENTIC

DIMENSION	GENERATIVE CHATBOT	AGENTIC SYSTEM
Core objective	Create text. Summarize, rephrase, retrieve.	Execute a goal. Manage the workflow to completion.
Operational mode	Reactive. Waits for a prompt.	Proactive. Can start from a data event, not a message.
Resolution depth	Single-shot. Answers, cannot change anything.	Multi-step. Calls tools in sequence to fix the root cause.
System interaction	Isolated conversation window.	Reads and writes CRM, inventory, payments, fulfillment.
Failure handling	Halts. Returns an error or a policy link.	Re-plans. Forms an alternative route, or escalates with context.
Primary metric	Deflection rate, answer accuracy.	First-contact resolution, total resolution time, reversion rate.

Only the last row is a management decision. Deflection names the goal as getting rid of the customer. Every organization that manages to it eventually gets what it measured — a contact center optimized for making escape difficult.

The test you can run in a demo

Change the customer's mind halfway through — a different date, a new constraint, a complaint mid-transaction. An automation restarts or transfers. An agent re-plans and tells you what it changed.

Then ask the harder version: **can it decide something that was not decided for it, and can it act on that decision?** If not, it is a chatbot with better prose — and none of the guardrail work here is needed, because nothing is at risk.

RUN IT ON YOUR OWN BUILD TOO

The demo question is a procurement filter. It is also the acceptance criterion for phase four.

Why the distinction sets the budget

A read-only assistant is an interface project. A system with write access is an accountability project, and it costs differently: identity resolution, scoped credentials, a replayable log, an approval queue, and an owner who can turn it off.

Teams that price the first and build the second stall between pilot and production. That gap has a name — the write-access wall — and it is where most agentic programs stop.

THE BUDGET LINE TEAMS FORGET

Guardrails are a material share of build effort on a first agent, and they are not optional. Priced at zero, they get cut — and the program reaches production without a stop button.

One retailer, two workflows, three systems

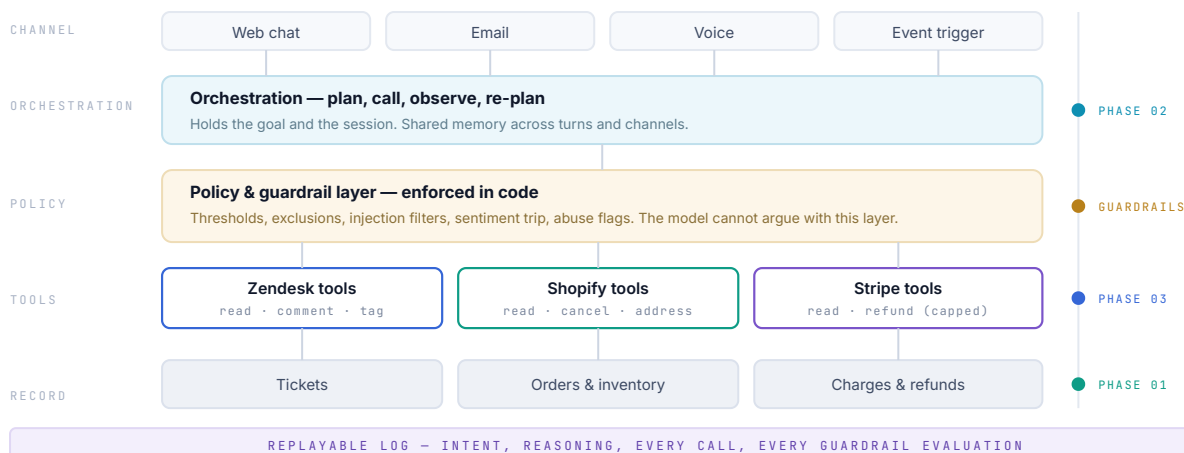
Everything from here is carried against a single build: a mid-market e-commerce retailer automating order cancellations and refund processing across Zendesk, Shopify and Stripe. High volume, clear policy, real money, and a wrong answer that costs something.

FIGURE 3 · THE SAME TWO REQUESTS, ANSWERED BY EACH ARCHITECTURE

REQUEST	GENERATIVE CHATBOT	AGENTIC SYSTEM
"Cancel my order"	Tells the customer to log in, open <i>My Orders</i> and click cancel. Links the policy page. 0 SYSTEMS TOUCHED	Looks up the order in Shopify, checks fulfillment status, holds the warehouse line, cancels, writes the outcome back to Zendesk. 3 SYSTEMS, 1 TURN
"Where is my refund?"	Reads the return policy aloud and asks the customer to wait five to seven business days. NO RECORD CONSULTED	Queries the Stripe charge, returns the exact status and settlement date, and gives the bank transfer reference. SPECIFIC, CHECKABLE

Note what changed for the customer. Not tone, not speed — the ending. One conversation ends with instructions. The other ends with the thing being done, and a reference number they can hold you to.

FIGURE 4 · THE REFERENCE ARCHITECTURE, AND WHERE EACH PHASE DOES ITS WORK



Read it bottom-up when you plan, top-down when you debug. Phase one is a data problem at the record layer. Phase three is a credentials problem at the tool layer. The policy layer sits between the model and the tools on purpose — never inside the prompt.

WHY THIS WORKFLOW

High volume, clear policy

Frequent, the rule is written down, and the outcome is binary. Ambiguous work makes a bad first agent however valuable it looks.

WHY IT IS STILL HARD

It writes to three systems

Cancel in Shopify, refund in Stripe, resolve in Zendesk. Three credentials, three failure modes, and one customer who sees all of them as you.

WHAT MAKES IT SAFE

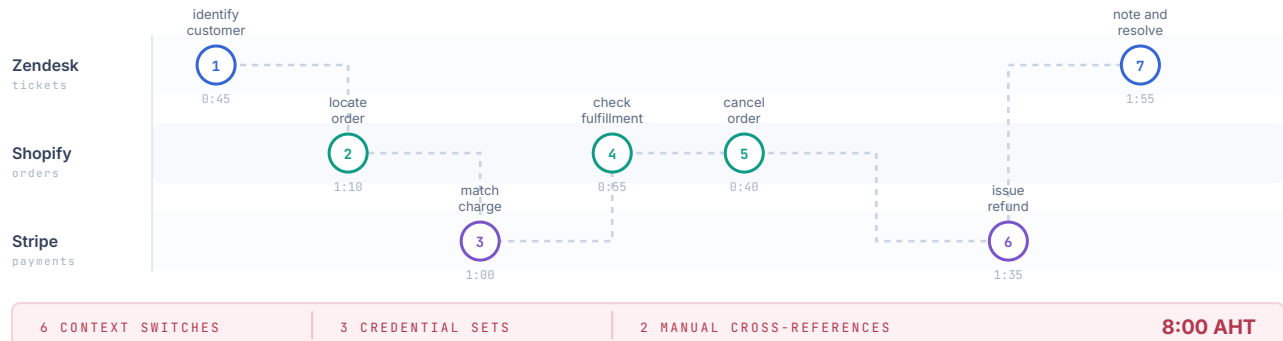
Money has a ceiling

Refund value is a number, so the guardrail is a number. Risk that cannot be written as a threshold belongs in a later phase.

Find the work that crosses three systems

The value of an agent is not in the answer it gives. It is in the context switches it removes. So the first artifact is not a use case — it is a map of where your people are already hopping between tools to finish one customer's request.

FIGURE 5 • THE WORKFLOW CROSSING MAP — ONE CANCELLATION, DONE BY A PERSON



Map the hops, not the intent. “Cancel my order” is one sentence and seven actions in three systems. The agent's job is the seven actions. If a workflow does not cross at least three tools, the payoff is a better answer, not a resolved case — and a chatbot will get you most of it for a fraction of the effort.

FIGURE 6 • THE CANDIDATE SCORECARD — SCORE BEFORE YOU SCOPE

CANDIDATE WORKFLOW	MONTHLY VOLUME	SYSTEMS CROSSED	REVERSIBLE	POLICY IS WRITTEN	BUILD PRIORITY
Order cancellation	14,200	3	YES	YES	first
Refund status lookup	9,800	2	N/A	YES	first
Refund processing	6,400	3	PARTLY	YES	second, capped
Address change in transit	3,100	3	NO	PARTLY	confirm first
Damage & injury claims	420	4	NO	NO	never — excluded

Volume ranks the list. Reversibility decides the tier. A high-volume irreversible action is the most dangerous thing you can automate first, and it is exactly what an ROI-ranked backlog puts at the top.

WHERE THE MAP COMES FROM

Watch, do not survey

Sit with four agents for a morning and record the screens. What people say they do and what the tab order shows are different documents.

THE NUMBER THAT PERSUADES

Context switches, not minutes

Minutes get argued. “Six tool switches to answer one sentence” does not, and it is the same finding stated in the language of the work.

WHAT BELONGS ON THE NEVER LIST

Write it down now

The bottom row of Figure 6 is the start of the exclusion list — the artifact almost nobody keeps and the one that prevents the expensive failure.

Profile the data before you promise the outcome

An agent inherits your data quality and then acts on it at machine speed. A human reading a mismatched record pauses. An agent refunds the wrong charge.

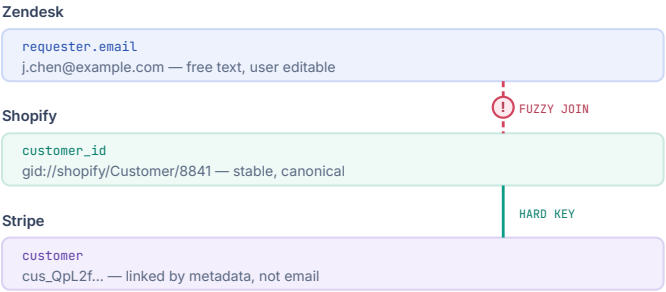
Data readiness is the single most-cited blocker among service operations teams — named by roughly seven in ten — and it is almost always discovered in week six rather than week one. The check is cheap and it belongs before any framework is chosen.

The specific thing to verify is not completeness. It is **whether the same customer resolves to the same entity in every system the agent will touch**, and whether the agent can tell when they do not.

THE FAILURE THIS PREVENTS

Two customers share an email at a shared household account. The agent cancels the wrong order, refunds the wrong card, and both events are technically correct at every step.

FIGURE 7 • IDENTITY RESOLUTION, WHERE IT BREAKS



The weakest join sets the agent's reliability. Fix the top link before granting write access, or the agent acts confidently on the wrong record.

72%

cite data readiness as a major blocker
SALESFORCE, 2026

74%

are frustrated by repeating themselves
ZENDESK, 2026

81%

expect the thread to continue, not restart
ZENDESK, 2026

FIGURE 8 • THE DATA READINESS GATE — FIVE CHECKS, ALL MUST PASS

CHECK	WHAT "PASS" MEANS	IF IT FAILS
01 Identity resolves	One customer maps to one entity in every system, by a stable key — not by email.	Stop. This is a data project, not an AI project.
02 Status is authoritative	One system owns fulfillment status and the others read it. No two sources disagree.	Name the owner system and make the rest read-only.
03 Fields are typed	Amounts, dates and statuses are structured values, not notes a human wrote.	The agent may read but not decide on that field.
04 History is retrievable	Prior contacts on the same issue are available in the session, across channels.	Expect repeat-yourself complaints. Fix before scaling.
05 Writes are auditable	Every system can record which actor made a change and when.	No write access. Read-only assist until it exists.

These are gates, not a scorecard. Four out of five is not eighty percent ready — it is one unguarded path to a confident wrong action. Publish the result internally, including the failures.

Write the role card before the prompt

Treat the agent exactly as you would treat a new hire with system access: a written role, explicit authority, named limits, and a manager. The role card is the document that gets signed. The prompt is an implementation of it — and the prompt is not where the limits live.

FIGURE 9 · THE AGENT ROLE CARD — ARTIFACT 04

Order Resolution Agent

VERSION 1.0 · TIER 2

OWNER: DIR. CUSTOMER OPERATIONS

PURPOSE — ONE SENTENCE, NO CONJUNCTIONS

Resolve order cancellation and refund-status requests end to end, inside published policy, without handing the customer an instruction they could have found themselves.

IT MAY, WITHOUT ASKING

- Read order, fulfillment and payment status
- Cancel an *unfulfilled* order and confirm it
- Update a shipping address pre-fulfillment
- Issue a refund at or under **\$100**
- Generate and send a prepaid return label

IT MAY NEVER, UNDER ANY INSTRUCTION

- Refund above **\$100**, or split one to stay under it
- Cancel an order already in transit
- Alter, waive or reinterpret published policy
- Act on damage, injury or legal claims
- Close a ticket a customer has disputed

ESCALATION TRIGGERS — HAND OVER IMMEDIATELY

Value	Refund or credit above the cap
Sentiment	Sustained frustration, or any legal keyword
Vulnerability	Bereavement, hardship, safety, accessibility
Repetition	Third contact on the same order
Uncertainty	Confidence below threshold on the match

ACCOUNTABILITY

Owner	Named individual, not a committee
Review	Monthly, against the live scorecard
Stop	Kill switch; four named people may press it
Record	Full replay retained 24 months
Disclosure	Stated at first turn, in conversation

Every line here is enforceable somewhere other than the prompt. That is the test of a finished role card: read each line and name the system that would refuse the action if the model tried it anyway. Lines with no such system are wishes, not limits.

Three questions that finish a role card

- 01 Who signs it?** A card without a named individual is a draft. If nobody can be paged when the agent does something surprising, the authority it describes does not exist.
- 02 What is missing from the never list?** Nearly every organization keeps a list of what agents can do. Almost none keeps the list of what they must never do — the higher-value document.
- 03 What happens on the third contact?** Most role cards handle the first attempt beautifully and say nothing about the customer back for a third time.

WRITE THE NEVER LIST IN A ROOM, NOT A BACKLOG

It takes a workshop and about three hours: customer operations, legal, risk and the design lead naming the moments that will never be automated — and why. The output is a published internal document, not a config file.

THE VERSION NUMBER MATTERS

"We raised the cap on 12 March because reversions ran under one percent for eight weeks" is auditable. "It seemed to be going well" is not. When authority widens, the version increments and the reason is recorded.

Policy goes in code. Language goes in the prompt.

A model asked to respect a budget can be argued out of it. A backend that refuses the call cannot. The division is not a style preference — it is the difference between a limit and a suggestion, and it determines whether your guardrails survive a customer who is trying.

FIGURE 10 • WHERE EACH INSTRUCTION ACTUALLY BELONGS

IN THE PROMPT — NEGOTIABLE BY NATURE

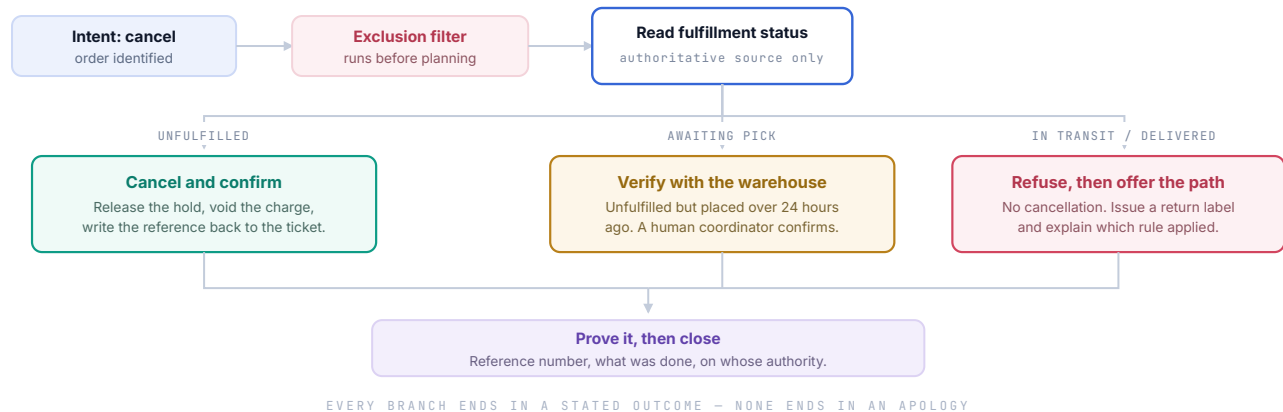
- Tone, register and how much to explain
- The order in which it gathers information
- How to phrase a refusal or a handover
- When to ask a clarifying question
- What the rationale it shows should contain

IN THE BACKEND — MUST NOT BE NEGOTIABLE

- The refund ceiling, and the session-wide total
- Which order states permit cancellation
- The exclusion list, evaluated before planning
- Rate limits per customer and per card
- Who the action is logged against

Track limits across the session, not per call. A single global rule of “confirm anything over \$100” invites slicing: six sequential \$95 approvals pass a per-transaction check and drain the same budget.

FIGURE 11 • THE CANCELLATION DECISION SPEC — WRITTEN AS STATES, NOT SENTENCES



Write it as states because states are testable. Prose instructions produce prose behavior. A state diagram gives QA a finite list of paths, gives engineering the branch conditions, and gives the customer a rule you can name when they ask why.

The distinction that matters is not how clever the agent is. It is whether the organization has decided, in advance and in writing, what it is permitted to do alone.

THE ARTIFACT THIS PRODUCES

A finite path list

Every branch in Figure 11 is a test case. A spec that cannot be enumerated cannot be regression-tested, which means no future prompt change can be verified.

THE TRAP IN THE MIDDLE BRANCH

Unfulfilled is not idle

An order marked unfulfilled but placed over 24 hours ago may already be packed. Status lag is how an agent cancels something that is physically on a truck.

THE BRANCH TEAMS FORGET

Refusal is a design surface

“No” is the most common agentic output and usually the least designed one. Name the rule, offer the alternative, and give one move to a person.

Scope the credentials like you scope a job

An agent that can read your systems answers questions. An agent that can write to them resolves cases. That is the whole difference in value — and the whole difference in risk. Grant the narrowest scope that completes the workflow, and nothing beyond it.

FIGURE 12 • THE SCOPE REGISTER — ARTIFACT 07, ONE ROW PER PERMISSION

SYSTEM	SCOPE GRANTED	MODE	WHY THE WORKFLOW NEEDS IT	HARD LIMIT IN CODE
Zendesk	<code>tickets:read</code>	READ	Identify the request and its history.	Requester's own tickets only.
Zendesk	<code>tickets:comment</code>	WRITE	Write the outcome and the reference back.	Cannot close a disputed ticket.
Shopify	<code>read_customers</code>	READ	Resolve the customer to a canonical ID.	No payment instrument fields.
Shopify	<code>read_orders</code>	READ	Check fulfillment status before acting.	Authoritative source only.
Shopify	<code>write_orders</code>	WRITE	Cancel, or update a pre-fulfillment address.	Unfulfilled states only. No line-item edits.
Stripe	<code>charges:read</code>	READ	Return an exact refund status and date.	Last four digits only, never the PAN.
Stripe	<code>refunds:write</code>	MONEY	Return funds without a second queue.	≤ \$100 per refund; ≤ \$100 per session; idempotency key required.

Every row names a limit that is not in the prompt. If the “hard limit” column is empty for any write scope, that scope is unbounded — whatever the role card says. The register is the document a security reviewer reads, and it should be readable in five minutes.

Four rules that survive contact with production

- 01 One identity per agent.** Actions log against the agent, not a shared service account or a borrowed admin. Without this there is no replayable record and no answer when a customer asks who did it.
- 02 Idempotency on every money call.** A retried refund must not be a second refund. This is the single most common expensive bug in agentic payments work.
- 03 Limits enforced at the boundary.** The cap lives in the tool wrapper, not the instruction. The model may propose anything; the boundary decides.
- 04 Read before write, always.** Re-read the authoritative status immediately before acting. Between planning and execution, the world moves.

THE SCOPE THAT IS ALWAYS OVER-GRANTED

Customer read. Requested broadly because it is convenient, it usually carries payment fields, addresses and lifetime history the workflow never needs. Narrow it to the fields the decision actually uses.

BLAST RADIUS, STATED PLAINLY

Write down the worst thing this agent could do if every model-side control failed and only the boundary held. If that sentence is unacceptable, the scope is wrong — not the prompt.

A SANDBOX IS NOT A STAGE

Test credentials pointed at production data are production credentials.

Measure the before, or you cannot claim the after

The baseline worksheet is the cheapest artifact in the blueprint and the one most often skipped, because it has to be built before anything interesting happens. Two weeks of manual observation now is the difference between a result and an anecdote.

FIGURE 13 • THE BASELINE WORKSHEET — ARTIFACT 08, FILLED BEFORE LAUNCH

MEASURE	MANUAL BASELINE	TARGET BAND	HOW IT IS CAPTURED, AND THE TRAP
Average handle time	8:00	3:00 – 4:30	Timed across 200 real cancellations. Include the wrap-up, or the saving is imaginary.
First-contact resolution	61%	78% – 85%	Count a case resolved only if no contact follows within 7 days on the same order.
Cost per resolution	\$5.10	\$1.60 – \$2.40	Fully loaded, including QA and supervisor time. Add inference cost to the after side.
CSAT on this journey	4.1	≥ 4.1	Journey-level, not contact-level. Hold it flat before trying to raise it.
Reopen rate	9%	< 9%	The honest counterweight to handle time. Fast and wrong shows up here first.
Escalation rate	n/a	5% – 15%	New measure. Zero escalations is a failure signal, not a success one.

Bands, not targets. A target invites gaming; a band makes both directions visible. Handle time below the band usually means the agent is closing cases it should have escalated — which the reopen rate will confirm two weeks later.

Build the business case on the right number

Cost per resolution is the number executives recognize and the one most often mis-stated. Gartner expects generative AI cost per resolution to exceed three dollars by 2030, surpassing many offshore human agent costs — so a case resting on indefinite unit-cost decline is resting on nothing.

The durable case is **mix**: routine volume resolved at machine cost, complex volume routed faster to people who now have context. Full automation is unlikely to be the cheapest answer. The right blend is.

Time to first measurable value

Seven in ten service organizations adopting agents observe value within sixty days. If your program cannot see anything in that window, the problem is almost certainly scope rather than technology — the workflow chosen was too broad, or too rare, or both.

SET THE REVIEW DATE BEFORE LAUNCH

Day 30 and day 60, in calendars, with the worksheet attached. Reviews scheduled after launch are scheduled by whoever is most optimistic.

70%

see measurable value within 60 days

SALESFORCE, 2026

+24%

first-contact resolution in scaled programs

MCKINSEY, 2026

60–90%

resolution-time reduction where the process was redesigned

MCKINSEY, 2025

\$3

expected cost per resolution by 2030

GARTNER, 2026

Widen on evidence, one gate at a time

Most stalled programs did not fail the pilot. They ran a pilot with no exit criteria, produced a result nobody could argue for or against, and stopped. The ramp below is the argument: each stage has a number that has to hold before the next one opens.

FIGURE 14 • THE TRAFFIC RAMP — EACH GATE IS A DECISION, NOT A DATE



Weeks are the expectation, not the rule. The gate opens when the number holds, not when the date arrives. A stage that has to be held open twice is telling you something about the design, and it is cheaper to hear it at fifteen percent than at full volume.

FIGURE 15 • THE EXIT CRITERIA SCORECARD — ARTIFACT 09

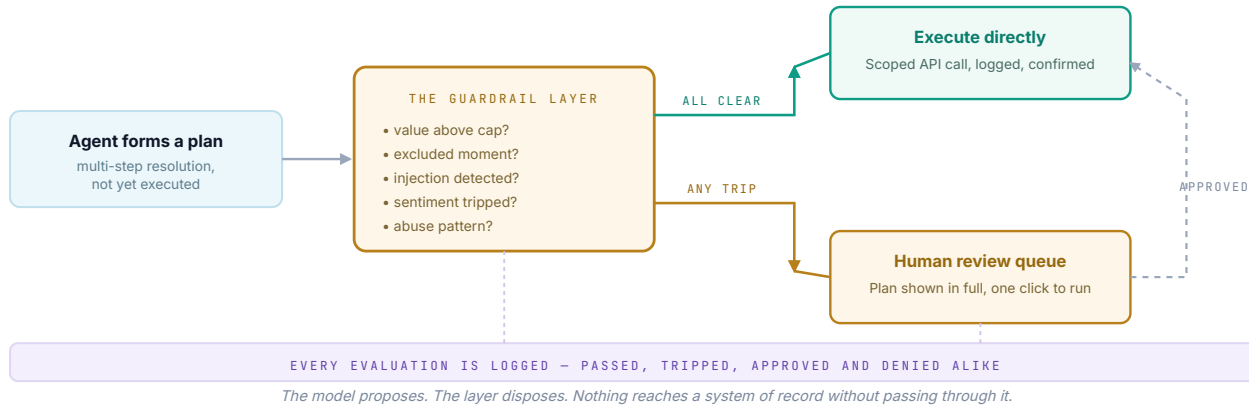
CRITERION	THRESHOLD	WHAT IT PROVES	IF IT IS MISSED
Tool call error rate	0%	The integration is correct, not merely working.	Hold the stage. This is an engineering defect, not a tuning problem.
Reversion rate	< 5%	Customers are not undoing what the agent did.	Narrow the authority. Consent was assumed where it was not given.
Calibration	> 80%	Stated confidence predicts actual acceptance.	Confident wrong answers. Lower the act threshold before widening.
Escalation rate	5–15%	The agent hands over deliberately and often enough to notice.	Below band, it is trapping people. Above, the scope is wrong.
Guardrail trips	reviewed	Every stop was correct and every stop was seen.	Unreviewed trips mean the queue has no owner. Fix that first.

Publish the scorecard, including the misses. The credibility of the second journey is built entirely on how honestly the first one was reported.

Design the stop before you design the start

Autonomous operations carry transactional risk that no amount of prompt quality removes. The control is architectural: a layer between the plan and the action that evaluates every proposed move against rules the model cannot reach, argue with, or be talked out of.

FIGURE 16 • THE DECISION GATE — EVERY ACTION PASSES THROUGH IT



The gate runs before execution, not after generation. A filter applied to the agent's words catches the sentence. A gate applied to the agent's calls catches the action — which is the only one that costs money.

Three properties the gate must have

- 01 **It is outside the model.** If the check lives in the same system that forms the plan, a sufficiently motivated conversation can move it. The gate is infrastructure, not instruction.
- 02 **It evaluates the action, not the sentence.** Content filters read output. A gate reads the proposed call — the endpoint, the amount, the record — which is what actually changes the world.
- 03 **It logs the passes too.** A record of only the stops cannot tell you whether thresholds are set correctly. The distribution of near-misses is the most useful tuning data you will have.

THE SIMPLEST VERSION THAT WORKS

A wrapper around every tool call that takes the proposed arguments, checks them against a policy table, and returns either the call or a structured refusal the agent must handle. It is a few hundred lines, and it is the difference between a demo and a system.

What the gate is not

- **Not a content moderator.** Those catch what the agent says. This catches what the agent does.
- **Not a confirmation dialog.** Asking the customer “are you sure?” moves liability without adding control.
- **Not the model's own judgment.** Self-assessed confidence is an input to the gate, never the gate itself.
- **Not a launch-only artifact.** Thresholds are reviewed monthly against actual trips, alongside the scorecard.

THE QUESTION THAT EXPOSES A MISSING GATE

“If the model decided to refund fifty thousand dollars, what would physically stop it?” If the answer names a prompt, a system message or a fine-tune, there is no gate — there is a preference.

Four classes of stop, four different mechanisms

Each class catches a different failure, and each needs its own detection, its own queue and its own owner. Collapsing them into one “are you sure?” prompt is how oversight becomes theater.

FIGURE 17 • THE GUARDRAIL SET — MECHANISM, TRIGGER AND THE METRIC THAT PROVES IT WORKS

CLASS	MECHANISM	WHAT IT CATCHES	HEALTH METRIC
Value thresholds	Hard caps in the tool wrapper, tracked per call <i>and</i> per session.	Refunds, credits and commitments beyond the granted authority.	Trips per thousand cases; share approved on review.
Intent integrity	Structural validation beneath the conversational surface, before planning.	Prompt injection, goal hijacking, instructions hidden in pasted content.	Blocked attempts; zero successful scope escapes.
Sentiment triage	Sentiment scored continuously alongside the task, not at the end.	Rising frustration, distress, legal language, vulnerability.	Time from trip to human contact; post-handover CSAT.
Abuse patterns	Rate limits by customer, card, device and address across sessions.	Refund farming, promotion abuse, repeated cancel-and-reorder.	Flag precision; recovered value against false positives.

Every class needs a queue with a name on it. A stop that routes nowhere is a slower failure, not a prevented one — and the most common way a guardrail program degrades is that the review queue quietly stops being read.

The threshold table, made concrete

Thresholds are a business decision written as arithmetic. Publish them, review them monthly against actual trips, and move them one band at a time.

FIGURE 18 • VALUE BANDS AND WHO CLEARS THEM

BAND	ACTION	CLEARED BY
\$0 - \$100	ACT ALONE	No one. Logged and confirmed.
\$100 - \$500	PROPOSE AND CLICK	Any service agent on shift.
\$500 +	ESCALATE FULLY	Team lead, with the record attached.
any • excluded	NEVER AUTOMATED	Specialist queue. Agent hands over context only.

Session totals are checked against the same bands. Six sequential \$95 approvals are one \$570 decision.

THE COUNTERINTUITIVE FINDING

More approval prompts make oversight worse. Microsoft's red team found human-in-the-loop bypass to be among the most consistently exploited failure modes, largely through consent fatigue. Approvals are a scarce resource. Spend them where consequence is real.

INJECTION IS NOT A PROMPT PROBLEM

Treat any text the customer supplies — and any text the agent retrieves — as untrusted input rather than instruction. The defense is structural: separate the channel that carries data from the channel that carries authority, and validate before planning.

FULFILLMENT SAFETY LATENCY

If a Shopify order reads *unfulfilled* but was placed more than 24 hours ago, the agent must confirm with a warehouse coordinator before canceling. Status lag is physical, and the box may already be on a truck.

Propose and click, not approve and hope

For high-stakes work, the agent should compile the entire resolution and present it as a finished plan that a person executes with one action. The human decision is then a judgment about a specific proposal — not a yes-or-no about an unread summary.

FIGURE 19 • THE APPROVAL SURFACE, DONE PROPERLY

PENDING — REFUND ABOVE CAP#48213

THE AGENT PROPOSES

- Refund \$284.00 to card ••4417
- Restock 2 items to warehouse B
- Email confirmation with reference
- Close ticket, tag "damaged in transit"

BECAUSE

Carrier marked damaged on 11 Sep. Policy 4.2 permits full refund. Customer has 1 prior claim in 24 months — below the abuse threshold.

EXCEEDS \$100 CAP BY \$184

Execute plan

Edit

Decline

0:14 read

Four elements make an approval real: the exact actions, the reasoning in plain language, the rule that was tripped, and the ability to edit rather than only accept or reject.

What makes an approval worth asking for

- **It is specific.** “Approve refund?” is not a decision. The amount, the instrument, the inventory effect and the tag are the decision.
- **It is editable.** A reviewer who can only accept or reject will eventually accept everything. Editing keeps attention engaged.
- **It is rare.** If a reviewer sees more than a handful an hour, the threshold is wrong and the queue has become a rubber stamp.
- **It is timed.** Track how long reviewers actually spend. A median below a few seconds is measurable evidence of consent fatigue.

ESCALATION IS A DESIGN SURFACE, NOT A FALLBACK

When the agent hands over, the context travels with the customer. Nothing is repeated on arrival. A clear escalation path makes people **45% more likely** to use an agent at all — the cheapest point of leverage in the whole program.

TEST IT WITH A PAST COMPLAINT

Take a real complaint from last quarter. From the record alone, can you reconstruct what the agent did, why, and on whose authority? If not, the guardrail layer has no floor under it yet.

FIGURE 20 • WHAT THE REPLAYABLE RECORD HAS TO CONTAIN

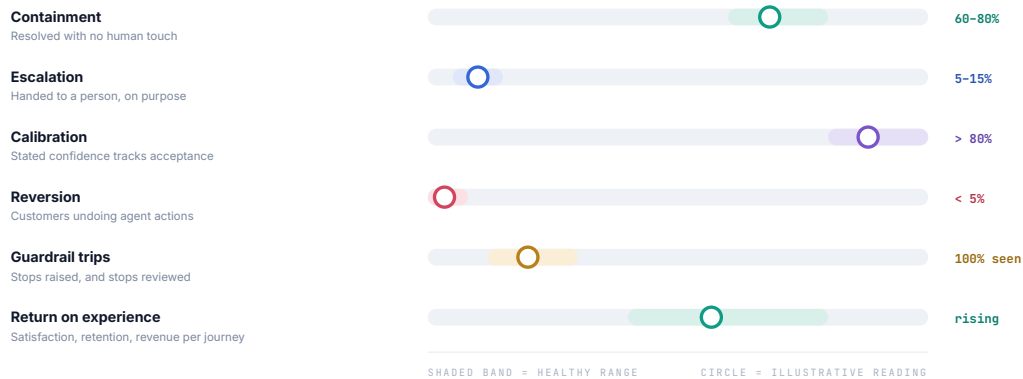
ELEMENT	CAPTURED	THE QUESTION IT ANSWERS LATER
Intent	What the agent understood the customer to want, in its own words.	Did it solve the right problem?
Reasoning	The plan it formed and the rule it applied, in plain language.	Why did it choose that?
Calls	Every tool invocation with arguments, response and timing.	What changed, and where?
Evaluations	Every guardrail check, passed or tripped, with the value tested.	Was it permitted to do that?
Surface	The interface state the customer actually saw.	What were they responding to?
Authority	The role card version in force, and any human approval.	On whose authority did this happen?

Without a replayable record you cannot tell a customer why something happened. Every trust pattern in this paper — disclosure, rationale, undo, escalation — depends on this layer existing first.

Six numbers on one page

Most agentic programs are reported on a single number: how much got handled without a person. On its own it is the most gameable metric in customer experience. Read against five others, it becomes an instrument panel — and the bands matter more than the values.

FIGURE 21 • THE LIVE SCORECARD — SHADED BAND IS HEALTHY, MARKER IS AN ILLUSTRATIVE READING



Containment outside its band is usually a sign of something else. Too low and the agent is not trusted with anything. Too high and customers cannot get out. Read on its own it rewards the wrong behavior — a team will hit the number by trapping people in a loop they cannot leave.

What each one is actually telling you

- **Containment** measures reach. Roughly seven in ten conversations resolve without escalation in mature deployments — a fair benchmark for tier-one work, not a target for everything.
- **Escalation** measures honesty. A healthy agent hands over deliberately, and often enough to notice.
- **Calibration** measures self-knowledge. Poor calibration is what produces confident wrong answers.
- **Reversion** measures consent. People undoing agent actions is the clearest signal that autonomy was granted where it was not wanted.

Two things to add to the board pack

Cost per resolution, with a horizon. Quote today's figure alongside the 2030 expectation, so nobody builds a case on unit costs falling forever.

Time to first measurable value. If the program cannot see anything inside sixty days, the scope was wrong — not the technology.

ONE MEASURE TO STOP USING

Deflection. It names the goal as getting rid of the customer, and everyone who manages to it eventually gets what they measured.

WEEKLY • THE BUILD TEAM

Trips and errors

Every guardrail trip read, every tool error triaged. Thirty minutes, no slides.

MONTHLY • THE OWNER

Bands and thresholds

The six measures against their bands. One decision — widen, hold or narrow — and the reason recorded against the role card version.

QUARTERLY • LEADERSHIP

Journeys and exclusions

Which journey goes next, and whether anything has moved on or off the never list.

ON INCIDENT • IMMEDIATELY

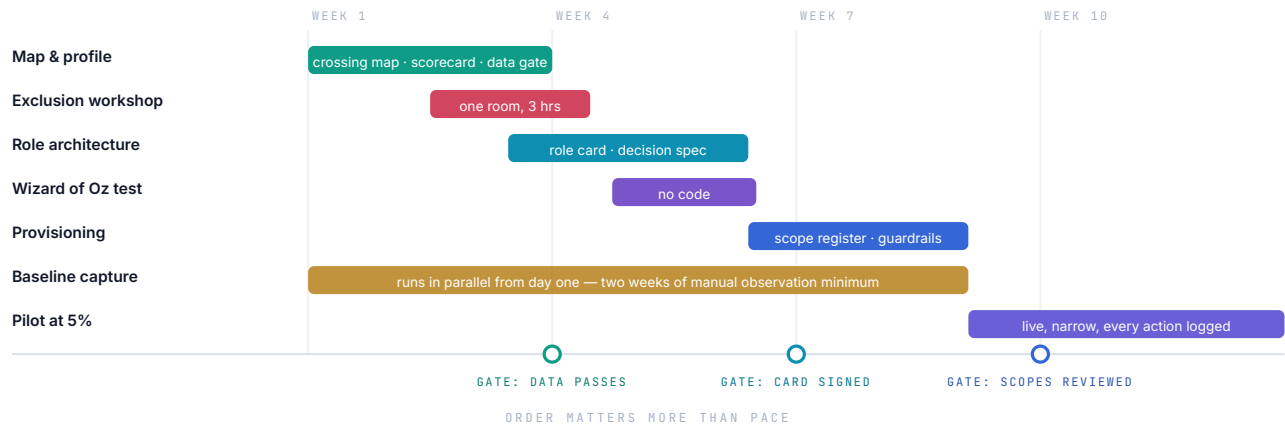
Replay, then decide

Pull the record, reconstruct the decision, and narrow authority the same day if it was too wide.

Ninety days, in this order

None of this requires a platform decision, and none of it is a technology project until week seven. The sequence below is what turns agentic ambition into something you can defend to a customer, a regulator and a board.

FIGURE 22 • THE NINETY-DAY PLAN — FOUR PEOPLE, NINE ARTIFACTS, ONE JOURNEY



Teams that start at provisioning build a guardrail layer for an authority model nobody has agreed. Teams that skip the exclusion workshop eventually automate a moment they should never have touched. The first six weeks produce no working software and they are the six that decide the outcome.

What good looks like at day ninety

- One journey where the agent acts, and a written list of journeys where it never will.
- A signed role card with a named owner, a version number and a review date.
- A scope register a security reviewer can read in five minutes.
- Six measures on one page, with bands rather than targets.
- A replayable record, and four people who may press stop.

The honest caveat

This field is young enough that the pattern vocabulary is still moving. Treat this blueprint — and any framework in this space — as a working instrument to be revised against your own evidence.

What is not moving is the underlying requirement. Customers will delegate to a system that discloses what it is, explains what it did, and lets them undo it. Everything else is implementation.

WHERE THIS GOES NEXT

Build it. Bound it. Prove it before you widen it.

f/cx.ai is a customer experience design and innovation agency for the AI era. We design, prototype and prove the journeys, service models and AI agents enterprise organizations run on.

IF YOU HAVE NOT STARTED

Run phase one

The crossing map and the exclusion workshop take two weeks and no engineering. They also decide most of what happens later.

IF YOU ARE STALLED IN PILOT

Find the missing artifact

Stalled pilots are almost always missing one of the nine. It is usually the role card, the scope register, or the baseline nobody captured.

IF YOU ARE ALREADY LIVE

Instrument, then widen

Stand up the six measures and the guardrail queue. Raise one band on one journey and watch what reversion does.

An agent is only as autonomous as its worst unguarded path.
Build the stop first, and the speed is free.

THE NINE ARTIFACTS, IN ORDER

- | | | |
|---------------------------|------------------------|----------------------------|
| 01 Workflow crossing map | 02 Candidate scorecard | 03 Data readiness report |
| 04 Agent role card | 05 Decision logic spec | 06 Exclusion list |
| 07 Scope & token register | 08 Baseline worksheet | 09 Exit criteria scorecard |

THE CONVERSATION WE HAVE MOST OFTEN

Which moments an agent may act in, under what authority, and what has to be true before you widen. If that is the question in front of your leadership team, it is the one we are set up for.

CONTACT

Skip Vanderburg

Founder & Chief Experience Officer

innovate@fractionalcx.ai

fractionalcx.ai

Where every external number came from

Every figure in this paper is original. External data points are attributed on the page where they appear and listed here in full. Three notes on care: the MIT NANDA figures circulated in a paraphrased form that overstates them; one Salesforce trust dataset is a 2024 baseline, flagged as such; and vendor-published performance data is self-interested, so it is paired with independent sources wherever a claim rests on it. Illustrative values inside the worked build — handle times, volumes, thresholds, costs — are constructed, not a client's results.

ADOPTION, SCALING AND THE PRODUCTION GAP

MIT NANDA 2025	<i>The GenAI Divide: State of AI in Business.</i> Task-specific enterprise AI tools: 60% evaluated, 20% piloted, 5% in production. Widely paraphrased; the headline finding concerns zero measurable return, not pilot failure.
Gartner 2025	<i>Press release: agentic AI project cancellations.</i> Over 40% of agentic AI projects expected to be canceled by end of 2027; "agent washing"; roughly 130 genuine agentic vendors.
McKinsey 2025	<i>Seizing the agentic AI advantage.</i> Fewer than 10% of deployed use cases pass pilot; up to 80% of common incidents resolvable autonomously under process reinvention; 60–90% resolution-time reduction.
McKinsey 2026	<i>Rewiring customer experience for the agentic era.</i> 41% of customer-facing AI deployments fully scaled, 3.5× the rate of other domains; named program outcomes including +24% first-contact resolution.

SERVICE PERFORMANCE, DATA READINESS AND ECONOMICS

Salesforce 2026	<i>AI service agents and customer satisfaction.</i> Agent adoption in service rose from 39% to 66%; 70% see value within 60 days; data readiness cited as a major blocker by 72%.
Salesforce 2026	<i>Agentic Enterprise Index.</i> Seven in ten conversations resolved without escalation; 32% escalation rate held steady across 170× conversation growth.
Gartner 2026	<i>GenAI cost per resolution forecast.</i> Cost per resolution to exceed \$3 by 2030, surpassing many offshore human agent costs; regulatory change to raise assisted interactions 30% by 2028.

CUSTOMER TRUST, MEMORY AND DISCLOSURE

Zendesk 2026	<i>CX Trends.</i> 95% expect an explanation for an AI-made decision; 74% frustrated by repeating themselves; 81% expect continuity rather than restart.
Salesforce 2024	<i>State of the AI Connected Customer.</i> ~75% want to know they are dealing with an agent; 45% more likely to use one given a clear escalation path; 44% given explained logic. 2024 baseline.
Intercom 2026	<i>AI Sentiment Report.</i> 54% would trust an agent with simple issues only; easy escalation and upfront disclosure are the top two comfort factors.

GOVERNANCE, FAILURE MODES AND CONTROL

Microsoft AI Red Team 2026	<i>Taxonomy of failure modes in agentic AI systems, v2.</i> Human-in-the-loop bypass among the most consistently exploited failure modes, largely via consent fatigue; goal hijacking; session context contamination.
IBM Institute for Business Value 2026	<i>The AI control gap.</i> 77% say adoption outpaces governance; two in three accountable for systems they do not control; control-mature organizations report 25% fewer incidents and deploy 16× more agents.
European Union 2024	<i>Regulation (EU) 2024/1689, Article 50.</i> Transparency obligations: notice at first interaction, clear and distinguishable, plus machine-readable marking of synthetic output. Confirm dates against the Regulation.

Prepared by f/cx.ai. Companion to *Agentic AI Customer Experience & Innovation* (Volume I, 2026). Zendesk, Shopify and Stripe are named as an illustrative mid-market stack, not as endorsements; verify scope and permission names against each vendor's current documentation, and regulatory dates against the primary legal text.