

AN EXECUTIVE BRIEFING

Secure by Design in the Age of AI™

Engineering velocity is up an order of magnitude.
Security review capacity is not.

Stephen R. Jordan

Founder, SRJ Consulting & Services LLC

Nobody skipped a gate.
Every gate had been passed.

WHERE MOST PRODUCT SECURITY TEAMS ARE

The lifecycle is current. The backlog is a hundred items deep.

The feature shipped on a Thursday. Nine days later a customer's security team found it would apply a credit to an account it had never been asked about. The director pulled the release record expecting to find the gate someone had gone around.

Every gate had been passed. The sprint had closed forty-one changes, thirty of them generated with an assistant, and the two reviews the lifecycle required had each taken about eleven minutes.

Four questions the lifecycle cannot answer:

- 01** What is in the product that is not code, and when did it last change?
- 02** Which feature can act without a human, and at what consequence?
- 03** What did we actually verify at the moment this release was cut?
- 04** If a customer reports behavior rather than a defect, who owns it?

Four parties will ask you to prove the product is safe. On their schedule.

4

The enterprise customer

The security questionnaire now carries an AI section, and their own team will test the product the way Volume V taught them to.

The certification auditor

Asks which model processes customer data, what it can do with it, and how you know when it changed.

The regulator

Product security obligations arrive on a calendar: reporting clocks, support periods, conformity procedures, documentation.

The acquirer

Diligence now asks what the models were trained on, and whether the product's security can be reconstructed from its records.

None of these are going away. The only variable is whether you arrive with an operating model or with a backlog chart.

The Dual-Impedance Problem

Two mismatches arrived together, and each remedy for one makes the other worse.

RATE

Submission capacity rose by an order of magnitude. Review capacity did not. Pull request volume roughly doubles after assistants are adopted, review time lengthens by about ninety percent, and delivery metrics stay flat.

KIND

The defects sit in components that did not exist: prompts, retrieval, tools, agents, memory. Whether a system prompt holds under injection cannot be determined by reading it. The defect is not in the diff.

Hiring reviewers addresses rate and loses the race. Lengthening the checklist addresses kind and deepens the queue. Buying model-based review detects a fraction of what a human flags and lengthens closure time. The problem is not a gate that failed. It is a gate sized for a volume that no longer arrives.

THE STRUCTURE OF THE PROGRAM

Five frameworks. One question each.
See, sort, verify, decide, sustain.

SEE THE SURFACE

One row per boundary crossing,
validated against what is running.

SORT BY CONSEQUENCE

Five dimensions, four tiers, so
judgment goes where the blast radius
is.

VERIFY IN TWO LANES

A machine lane that blocks. A human
lane reserved by tier.

DECIDE AND RECORD

Evidence at the release timestamp,
three states, a named acceptor.

05 SUSTAIN AFTER RELEASE

Six standing functions from the first report from outside through retirement. The rung that has never run on a product is the hole in the ladder.

Per Boundary. Never Per Feature.

A product security model written per feature is obsolete the moment an integration adds a tool.

A per-feature model asks whether this assistant is approved. It has to be reopened on every release, every integration, and every capability a vendor adds. It was already out of date at Larkfield when a tool registry change added bulk crediting, because that was not a code change and nothing in the lifecycle read it.

A per-boundary model asks what enters the context window, what enters the corpus, what the output can invoke, what receives it, and what persists between sessions. Those five questions survive the next integration, because a new tool is a new row rather than a new model.

Untrusted content may inform the model. It must never authorize a side effect. That single sentence is what the enforcement point exists to hold, and it does not change when the feature does.

WHAT CARRIES THE PROGRAM

Three instruments convert intention into evidence a customer can read.

01 The AI Attack Surface Ledger

One row for every input, retrieval source, tool, model, output, memory and update path. Each row carries an owner, a mitigation, a classification and a validation date. Validated against the running product rather than the design.

02 The Two-Lane Assurance Model

A deterministic lane that blocks every change, a human lane reserved by tier, and model-based review in an advisory band that never gates. The classification column on the ledger decides what goes where.

03 The Ship Decision Record

Evidence at the release timestamp rather than at some point in the sprint. Three states, promote, hold, or roll back, and a named executive on the residual risk.

A Gate Is Not a Verification

The distinction the customer, the auditor and the acquirer are all testing for.

THE GATE

A review was scheduled.

The scanner passed.

The tool is deployed.

The threat model was approved.

THE VERIFICATION

A review ran, against this change, and its record cites the change it examined.

The scan ran at the release timestamp, against the artifact being cut, cited by run identifier.

The check blocks the merge. A warning that does not stop a merge is a warning nobody is required to read.

Every change since opened a delta, and each delta closed with its re-verification attached.

SELF-DIAGNOSTIC

Three signs the lifecycle has already stopped describing the product.

01 Nobody can say what the product is made of.

Ask for the list of models, prompts, retrieval sources and tools in one product, with the date each last changed. If the answer is assembled rather than read from a register, the ledger does not exist.

02 The merge is never refused.

Ask how many merges the pipeline blocked last month. If the number is zero, the tooling is advisory, and advisory tooling verifies nothing.

03 The release decision lives in a chat thread.

Ask which version was tested for the release that shipped six months ago, what was open, and who accepted it. If that cannot be answered in a minute, there is no ship record.

Categories Outlive Vendors

Not one vendor is named anywhere in this program.

The tooling market for AI security consolidates faster than a pipeline can be rebuilt. A lane-one configuration anchored to a named product has to be revisited every time that product is acquired, renamed or discontinued.

Anchored to a tool class and an architectural pattern, the configuration survives. The frameworks are cited only where they bind: NIST SP 800-218 and 800-218A, NIST AI 100-2, ISO/IEC 27034, MITRE ATLAS, CISA Secure by Design, and the OWASP LLM and Agentic lists.

Every claim in the book is sourced. Where the evidence does not settle a question, the book says so rather than filling the gap with confidence.

The Review Capacity Ratio

The number that turns a backlog complaint into a decision a CFO can act on.

Security-relevant changes in the period, divided by reviewer hours available at the skill the tier requires.

TARGETS

0.5 at tier one, 1.0 at tier two, 4.0 at tier three. Set by what a review requires, not by what the calendar allows.

WHAT IT REVEALS

Larkfield opened at 3.4 on tier-one work: eighteen minutes of judgment against a two-hour requirement.

THE THREE CLOSURES

Automate, tier, or hire. Each with a cost and a date by which the ratio reaches target. That is a decision, not a plea.

PILLAR 5

The First Ninety Days

Sequenced by dependency. Each week produces the artifact the next week reads.

WK 1

Score

The baseline on one product, dated and reviewed by the CISO.

WK 2

Measure

The Review Capacity Ratio for that product line, with the three closures costed.

WK 3

Intake

A ticket type with three fields, and the Seven Scoping Questions run against the roadmap.

WK 4

Tier

Every AI-enabled feature in development, on five dimensions, published to engineering.

WK 5

Ledger

Built for the highest-tier feature and validated against the running product.

WK 12

Ship

The first release whose evidence can be reconstructed by someone not in the room.

FRAMEWORK 01

See the Whole Surface

A threat model describes the product on one day. A ledger describes it on this one.

WHAT THE PROGRAM INSTALLS

The AI Attack Surface Ledger

One row per boundary crossing, with an owner, a mitigation and a validation date.

The Five AI Trust Boundaries

Model, retrieval, tool, output, state. Every finding files against one of them.

The Component Admission Standard

One set of conditions for libraries, models, datasets, frameworks, servers and skill bundles.

The Threat Model Delta

Eight triggers. A change opens a delta; the delta closes when its re-verification runs.

Sort Before You Design

Human judgment is the scarce resource. The tier map is the rule that decides where it goes.

WHAT THE PROGRAM INSTALLS

The AI Feature Tier Map

Five dimensions, four tiers, derived by rule and recorded with the scores.

The Seven Stop Conditions

A refusal with a rule behind it, and the design change that would clear it.

Intake before design

Three fields a product manager can answer. Two minutes to route, forty-five to assess.

The generated plan

Activities follow from the tier, so nothing is negotiated feature by feature.

Verify at the Rate the Product Changes

Rate goes to the machine. Kind goes to the human. Neither lane does the other's work.

WHAT THE PROGRAM INSTALLS

The Two-Lane Assurance Model

A deterministic lane that blocks on every change, a human lane reserved by tier.

The advisory band

Model-based review measured for signal and never permitted to gate.

Sampling by defect density

Rates set per generating tool from the organization's own code, revisited quarterly.

The reviewer's hour

Specification, boundary, failure behavior, mitigation. Not formatting the tools already check.

Decide, Then Sustain

Evidence produced after a feature ships is remediation evidence, not release evidence.

WHAT THE PROGRAM INSTALLS

The Ship Decision Record

Three states and a named acceptor, attached to the release tag.

The Post-Release Response Ladder

Six standing functions from the first outside report through retirement.

The Product Security Case

Bounded, verifiable, recoverable, each argued from evidence and naming its model dependence.

The Behavioral Vulnerability Protocol

For the report that has no code defect. Five steps, ending in the regression suite.

The no-review share

Beside the ratio sits the share of merged changes carrying no recorded human review.

Changes merged with no human review

In open source, about eighty percent of merged AI-co-authored pull requests from contributors without ownership carried none.

Tier-one changes merged without lane two

The number the sprint agreement exists to hold. It should be zero, and it is auditable.

Undocumented boundary crossings found at validation

Larkfield found five on a feature whose diagrams showed six. It should fall toward zero.

Silent model version changes per quarter

Marlow Ridge found three versions in one quarter and had been told about one.

The Product Attestation Pack

One assembly for four audiences. Every part is an artifact a phase already produced.

PART	WHAT IT PROVES
The Product Security Case	That the product is bounded, verifiable and recoverable, argued rather than asserted.
The baseline score, dated	Where the program stands, and how it has moved since the last score.
The ledger summary	What the product is made of, and when it was last validated against what runs.
The ship record and both bills of materials	What was true at the release timestamp, and what the product contains.
The disclosure record	That the response function runs, with closed reports and response times.

How the operating model is stood up

Five stages across the first four quarters, run against your release calendar.

- | | | |
|-----------|---------------------------|---|
| 01 | Baseline | Score the program on seven areas, compute the ratio, and put both in front of the CISO. |
| 02 | Sort | Install intake, tier the roadmap, and publish the human-lane forecast the tiering produces. |
| 03 | See | Build the ledger for the highest-tier product and validate it against what is running. |
| 04 | Verify | Make one machine-verifiable check block the merge, then move the classification column into lane one. |
| 05 | Decide and sustain | Produce the first ship record, then stand up the six rungs with owners and cadences. |

What you operate afterward

Thirty-three working instruments, each with a named owner and a cadence.

- | | | | |
|-----------|---|-----------|---|
| 01 | The Defensible Product Security Baseline scorecard | 02 | The Review Capacity Ratio worksheet and the three closures |
| 03 | The accountability and risk-acceptance table, by tier | 04 | The AI Attack Surface Ledger template and validation record |
| 05 | The AI Feature Tier Map and the Seven Stop Conditions checklist | 06 | The Threat Model Delta record and its trigger list |
| 07 | The lane assignment table and sampling policy | 08 | The Ship Decision Record and the Product Security Case template |
| 09 | The Post-Release Response Ladder, six rungs with cadences | 10 | The Definition of Secure Done and the Seven Scoping Questions |
| 11 | The model inventory, configuration record and canary suite | 12 | The laboratory manifest and the judgment budget |

The first ninety days

Sequenced by dependency. Each week produces the artifact the next week reads.

DAY 1–30

Score and measure

The baseline on the highest-revenue AI-enabled product, dated by the CISO. The ratio for that product line, with three costed closures. Neither requires a purchase.

DAY 31–60

Sort and see

Intake as a ticket type. The roadmap tiered on five dimensions. The ledger built for the highest-tier feature, then validated against what is running. Count the crossings nobody documented.

DAY 61–90

Verify and decide

One machine-verifiable mitigation made blocking, with its block rate measured weekly. Then the first release that ships with a complete Ship Decision Record.

TO SET EXPECTATIONS CLEARLY

What this program is not

This is the engineering discipline for products you build and ship to others.

- ✗ **Not a repeat of the audit**
Volume V finds the exposure across the estate. This builds the product so the exposure is designed out.
- ✗ **Not the runtime program**
Runtime determinism and behavior in production belong to Volume VIII. Infrastructure belongs to Volume IX.
- ✗ **Not a tool selection**
No vendor is named. Tools are sorted into a class that may gate and a class that may only advise.
- ✗ **Not a promise of zero risk**
Prompt injection is not fully solvable. The program bounds the consequence rather than claiming the behavior.

BUILT FOR

Who this program is for

Written for the practitioner who owns the work, not for the board and not for a general reader.

READERS

CISOs, and the directors and vice presidents of product security who answer to them. Security architects and assurance engineers who configure the pipeline. Engineering leaders who sign the definition of done.

PRECONDITION

You build software that other people run. If AI is only something your organization uses, Volumes V and VI are the ones to read first. This Volume begins where a company becomes answerable for what it ships.

Cedar Hollow became a product organization without noticing: an assistant built for its own physicians, adopted by three affiliated practices under a services agreement that mentioned it in one line.

ABOUT THE AUTHOR



Stephen R. Jordan

Founder, SRJ Consulting & Services LLC

Author of Secure by Design in the Age of AI™, the seventh volume of The Operating Discipline for AI Library™ and the third of Pillar II, AI Risk Governance & Security™. The Library is a nine-volume series on installing AI as a governed business function rather than a collection of pilots.

The books are operating manuals, structured around instruments a reader can put into service on a Monday: ledgers, tier maps, decision records and the cadences that keep them current. Every framework carries an owner, a cadence, and the decision it feeds.

No vendor is named anywhere in any volume, and no claim is made without a source. He lives and works in Frisco, Texas.

Ready to build the product so the exposure is designed out?

Book a conversation with Stephen R. Jordan.

srjconsultingservices.com

SRJ Consulting & Services LLC

Dallas–Fort Worth · Nationwide