AI Security

A Practitioner's Field Guide — Volume 1: Foundations and the Threat Landscape

Sylvan Press

Sylvan Press

Field-tested, plain-English references for the people who do the work and stand behind their call.

Contents

AI Security

A Practitioner's Field Guide — Volume 1: Foundations and the Threat Landscape

Published by Sylvan Press, the security imprint of Sylvan Assurance, LLC — Vermont, USA.

ISBN (paperback / hardcover / ebook): to be assigned prior to publication.

Print and ebook ISBNs are registered through Bowker before release.

First edition, 2026.

Disclaimer. This book provides general guidance and recommended security and compliance practices, not legal, financial, or professional advice, and it does not create any advisory or consulting relationship between the author, the publisher, and the reader. Every recommendation is optional — a widely accepted way to reduce a common risk — and the reader decides which to adopt, adapt, or decline. Following this guidance reduces exposure to common risks but does not guarantee any particular outcome and does not prevent any particular breach, incident, or loss. Where this book refers to laws, regulations, standards, or commercial products, it does so for general awareness only; confirm specific obligations with qualified counsel, and note that product references are illustrative and do not constitute endorsement.

Sylvan Press — sylvanassurance.com

Preface

AI security has a hype problem in both directions. One camp says the machines are about to end us; the other says the whole field is a rebranding exercise. Both stories are worth distrusting. In the working world this book is written for, AI systems are simply becoming infrastructure — attackable, defendable, auditable infrastructure — faster than most security programmes can absorb. The interesting questions are not philosophical. They are operational: what is this system, what can go wrong with it, who would want it to, and what can actually be done about it before something does.

That operational view is the one that is hardest to find. The available material tends to sit at one of two extremes — research papers written for the people who build models, or executive briefings written to alarm the people who sign budgets. The security engineer in the middle, the one who has been doing threat modelling and incident response for a decade and has just been handed "the AI thing," is left to translate. Much of that translation, it turns out, is reassuring: a great deal of AI security is the security you already know, wearing unfamiliar vocabulary. The rest — the genuinely novel failure modes — is what deserves real attention, and this book works hard to tell the two apart.

The judgement in this book comes from more than two decades of practitioner experience inside a top-25 global company — across network security, computer security incident response, and product secu-

rity incident response.

This first volume does the foundational work: what AI security actually is and is not, how to treat AI systems as security objects, the vulnerability taxonomy, threat modelling that fits these systems rather than fighting them, and the maturity spectrum for reading an organisation honestly. Then it walks the attack surface in depth — prompt injection, jailbreaks, training-data poisoning, model extraction and inversion, embedding-space attacks, the AI supply chain, and agentic systems with tool use. The second volume takes up the working programme that lives around all of this; the two are designed to be read together, but this one stands on its own.

The recommendations are framed in the Sylvan defensibility style: practices that reduce risk, never guarantees. The field is young, the standards are still settling, and anyone selling certainty is selling something else. What the book offers instead is a way of seeing these systems clearly enough to defend them — and to know which of the day's alarming headlines is a problem the discipline already solved years ago under a different name.

How to Use This Book

0.0.1 The shape of this volume

Part I — Foundations (Chapters 1 through 5). What AI security is, AI systems as security objects, the taxonomy, threat modelling, and the maturity spectrum. Read in order if you are new to the field.

Part II — The Attack Surface (Chapters 6 through 12). Prompt injection, jailbreaks, poisoning and provenance, extraction and inversion, embedding-space attacks, supply chain, and agentic systems. Self-contained; read what your systems expose first.

Part III — The Expanding Surface (Chapters 13 through 17). Multimodal and vision-language attacks, retrieval-layer security, fine-tuning and alignment-stripping, the open-weights trade-off, and evaluating robustness.

The AI Security set. This book is complete on its own: the foundations and the full threat landscape, from the taxonomy and threat modelling through the attack surface to the robustness evaluation that tests the defences. Two companion books extend the practice for readers who want them. *Running the Programme* covers vendor due diligence, incident response, red-teaming, tabletops, defensive AI, governance, the regulatory horizon, and maturity. *The AI Security Operator's Workbook* holds the templates, runbooks, checklists, the AI Security Glossary, and the Maturity Rubric. Each stands alone; none assumes you have read the others. The complete set is under $30 in ebook.

About This Book

A few conventions are worth flagging. Acronyms are spelled out at first use in each major section, not only at first use in the book. Where the book recommends a practice, the recommendation is framed as an approach that reduces risk — not as a guarantee of any outcome. Nothing in this book is legal advice; regulatory and contractual specifics belong with qualified counsel, and the landscape moves — verify current requirements against authoritative sources.

Chapter 1 begins on the next page.

1 What AI Security Actually Is (And Isn't)

The voice on the line was unmistakably Adaeze Nwankwo's — the cadence, the trailing vowels, the half-laugh she used to soften a hard question. It was telling a producer at a rival podcast network that she would be glad to record a guest spot on Tuesday, and could they send the brief to an old Gmail address.

Adaeze had not made the call. She was in Lagos, asleep, eight time zones from the laptop her voice was coming out of.

The mechanics were ordinary. A freelance editor, no longer on the books but still in the shared workspace, had kept access to Halftone Audio Collective's voice-cloning tool. They cloned Adaeze's voice from episodes she had recorded for the network, generated thirty seconds of her accepting an invitation she had never received, and played it down a phone line. The tool did exactly what it was sold to do.

Halftone is twelve people and six shows. Daniel Okafor, one of the two founders, hosts a show, runs the IT, and — because nobody else would — had signed off on the voice-cloning licence eighteen months earlier. They had been careful. Every patched line was disclosed on air; consent forms were on file for every host. None of it helped, because the misuse never went through the consent process.

What nobody at Halftone had been paid to ask was the security

question. Who could use this tool? Who did? What could it be used for, how would we know, and who owns the answer? Those questions went unasked because nobody had framed them as security questions in the first place.

This book is about how to frame them — and how to answer them.

Here is the definition it builds on: AI security is the security discipline applied to systems that include AI or ML components. That is less of a departure than it sounds, and it is the central argument of this book. AI security is not a new field. It is an extension of one the profession has practised for thirty years — asset inventory, threat modelling, access management, vendor due diligence, incident response — stretched to cover a new kind of object. There is genuinely new material, and the chapters ahead give it the room it deserves. But it is a bounded, learnable body of work laid on top of a craft many readers will already have.

Halftone is where we begin because it is small enough to see whole. Larger organisations arrive later, where the same incidents run for weeks instead of hours, across systems no one person can hold in their head. The scale changes. The shape of the problem does not.

1.0.1 The working definition

Stay with that definition, because it is doing a lot of quiet work. The discipline it names is the same one the field has practised for thirty years: asset inventory, threat modelling, identity and access management, secure development, vulnerability management, incident response, vendor due diligence, governance and assurance. The objects of that discipline have changed. A modern AI system is not a single piece of software in the way a database server is a single piece of software. It is a base model, often produced by a third party from training data the user of the model will never see; a fine-tuning corpus, often assembled from a mixture of curated and ambient sources; a retrieval layer that draws on documents the system was not originally trained on; a system prompt that constitutes part of the system's behaviour and is also the most fragile part of it; a set of tool integrations through which the system can act on the world; and an inference infrastructure that may

be on-premises, in a cloud the organisation pays for, or behind an API the organisation does not control.

Each of those layers is an attack surface. Each is also a layer at which a security professional with twenty years of experience already has serviceable instincts. The base model is a third-party dependency — and the field has thirty years of practice asking hard questions about third-party dependencies. The fine-tuning corpus is a data asset — and the field has well-developed practice around the security classification, handling, and provenance of data assets. The retrieval layer is, mechanically, a database with a search interface — and databases with search interfaces are familiar ground. The system prompt is a configuration artefact, and configuration artefacts have always required version control, change management, and access control. The tool integrations are API calls subject to authorisation decisions, and we have been writing authorisation policies for as long as we have been writing applications. The inference infrastructure is a workload on a platform, and workloads on platforms are the bread and butter of every cloud security programme on earth.

What is new — and there is a genuine, narrow body of what is new, which the rest of this book will treat with the respect it deserves — is the handful of failure modes that arise from the combination of these layers and the specific way modern AI systems mix instructions and data. Prompt injection. Training-data poisoning. Model extraction. Embedding inversion. Excessive agency in agentic systems. These are not abstract concerns. They have practitioner names because practitioners have been responding to them in production environments since at least 2023. They will be the spine of Part II of this book.

But they are a narrower part of the problem than the marketing of AI security suggests. The single most common AI security incident I have seen in two decades of practice — and the great majority of the AI incidents quietly handled by the security functions of organisations I have worked alongside — is not a prompt injection or a training-data poisoning. It is a person who pasted something sensitive into a chat interface they should not have used, in a tool the organisation had not catalogued, under a contract nobody had read. That is a shadow-IT problem with an AI label on it. The discipline that addresses it is the

discipline the field has been practising on shadow IT for a decade.

The working definition holds, then, in both directions. AI security is the security discipline applied to systems that include AI or ML components. The discipline is mostly the one you already practise. The systems are mostly the ones you already know how to think about. The new parts are real, are bounded, and are learnable.

1.0.2 What AI security is not

The boundary work matters because the field is currently being asked to absorb three adjacent concerns under the same label, and the conflation is doing real harm to working security programmes.

The first adjacent concern is AI safety. AI safety, in the sense the term is used by researchers at the major model laboratories and at the universities and institutes that work alongside them, is the project of ensuring that increasingly capable AI systems behave in ways that are consistent with human intent and human flourishing, including in deployment scenarios that have not yet been observed. It is a serious and important research programme. It is concerned with questions of alignment, of capability evaluation, of long-horizon agent behaviour, of catastrophic and existential risk from systems that do not yet exist at the scale being modelled. It is not what the security function of a podcast network, a legal-tech vendor, a regional supermarket chain, or a global freight company is being asked to do.

When a CEO asks the CISO whether their AI is "safe", they are almost never asking about alignment in the research sense. They are asking a security question — is the system going to leak data, get jailbroken into producing something embarrassing, be exploited to take an action it should not have taken? — wearing a safety word. The CISO who answers the alignment question rather than the security question will be heard as evasive and will lose the conversation. The CISO who clarifies the question and answers the security one will be heard as competent and will keep the seat at the table. The conflation is harmful in both directions: it loads the security function with a research mandate it cannot deliver, and it dilutes the seriousness of the safety research community's actual concerns.

The second adjacent concern is AI ethics in the philosophical sense — the body of work concerned with the morally appropriate uses, design choices, and societal effects of AI systems. This is also a serious field. It is also not what a security function does, even if individual security professionals hold informed opinions about it. The security function's contribution to an organisation's AI ethics posture is, properly, the assurance work — can we say with confidence what the system did, who decided to deploy it, what data it used, who has access to its outputs — that allows the organisation's ethics, legal, and policy functions to do their work credibly. The security function is not the venue in which the substantive ethical questions are decided. The CISO who tries to make it that venue will exhaust their function and underdeliver on the security work the organisation actually depends on them for.

The third adjacent concern is AI policy advocacy — the work of shaping the public and regulatory conversation about how AI systems should be developed, deployed, and constrained. Some of the most thoughtful people in the field do this work and the field is better for it. It is not security work either. The security function may inform the organisation's policy positions — by providing evidence about what works, what fails, what is feasible to assure, and what is not — but it does not own them.

There is a fourth boundary that matters in the other direction. AI security is not AI governance, although the two are close cousins and frequently get organisationally confused. Governance, in the sense used in this book, is the framework of policies, standards, roles, controls, and assurance activities by which an organisation manages its AI estate against its risk appetite and its regulatory obligations. The security function owns a meaningful share of the controls that governance frameworks rest on — incident response, vulnerability management, vendor due diligence, red-teaming, monitoring — but does not own the framework as a whole. Where governance functions exist (and Greenfield & Marrow's, which we will meet shortly, is one of the better ones), the security function partners with them. Where governance functions do not yet exist, the security function is sometimes asked to stand in. That is workable as a temporary measure. It is not a stable

end-state, and the companion guide *Running the Programme* returns, in its governance chapter, to the question of how the two functions properly relate.

Drawing these boundaries crisply is not a turf exercise. It is the precondition for the security function being able to do its actual job. A security programme that tries to be a safety research lab, a philosophy department, a policy shop, and a governance secretariat all at once will deliver none of those things and will also fail to defend the AI systems the organisation has put into production. A security programme that knows where its remit starts and ends can — and the rest of this book is the argument that it can — do the security work to a recognisable professional standard.

1.0.3 The thesis

I stated the thesis in this chapter's opening: AI security is not a new field, but an extension of an existing one. It is not a popular position to take in 2026. The market is full of vendors who would like to convince organisations that the AI security problem requires a wholly new tool category, staffed by a new species of expert, sitting under a new function with a new budget line. The conferences are full of talks that imply the same thing in subtler language. The thesis is unpopular because there is genuine money to be made if practitioners and procurers can be persuaded that nothing they know transfers.

The thesis is unpopular and, in my experience over two decades of security practice, it is also wrong on the evidence. The security professionals I have watched move into AI security work over the last three years — and I have watched a fair number — have not had to throw away their craft. They have had to add a layer to it. The threat modeller who knew how to walk a STRIDE analysis across a payments system has been able to walk a STRIDE analysis across an inference pipeline within a quarter. The vulnerability manager who ran a working software bill of materials (SBOM) programme for a software estate has been able to extend the same discipline to AI components, with the genuinely new wrinkle that the bill of materials includes training data and the field's standards for describing that lineage are still maturing.

The incident responder who has run a working IR programme has been able to handle an AI incident with the same lifecycle, with the genuinely new wrinkle that the containment options include rolling back a model and the evidence-handling options include preserving embeddings. The vendor due diligence specialist who has spent a decade asking software vendors hard questions has been able to extend the questionnaire to AI vendors, with the genuinely new wrinkle that the answers about training data are often "we don't know" and the playbook for that answer is still being written.

In each case the existing craft has done the heavy lifting. The new material — the genuinely new material that practitioners actually have to learn — has been the narrower body of failure modes I named earlier: prompt injection and jailbreaks, training-data poisoning and provenance, model extraction and inversion, embedding-space attacks, AI-specific supply-chain reasoning, agentic-system risk, AI-specific incident response patterns, and the regulatory framings that are now bearing down on all of these. That is a real body of new knowledge. It is also a finite, learnable body of new knowledge. Most of it can be acquired by an experienced security professional over the course of a year of deliberate study and practical engagement, and the rest of this book is one route through that year.

The practitioner's anxiety, in my experience, is the wrong shape. The anxiety in the security community in 2026 is that AI has rendered existing expertise irrelevant. It has not. AI has rendered existing expertise necessary-but-no-longer-sufficient, which is a very different proposition, and one the field has weathered before. The arrival of cloud computing in the late 2000s did not render network security expertise obsolete; it required network security expertise to be supplemented with cloud-specific reasoning. The arrival of microservices and containers in the mid-2010s did not render application security expertise obsolete; it required AppSec to absorb new patterns of trust and deployment. The arrival of AI in production at scale in the mid-2020s requires the same extension move. The practitioners who treated cloud and containers as wholesale replacements for what they already knew did badly. The practitioners who treated them as extensions did well. The pattern, on the evidence so far, is the same here.

I want to be careful with this argument. It would be easy to read the thesis as a complacent one — as the suggestion that AI security is no big deal and that any competent security professional can sleepwalk through it. That is not the claim. The claim is that the work is hard, that there is real new material to learn, that the failure modes are genuine and have already produced material harm in production environments, and that none of that constitutes a reason to discard the discipline the field has built over thirty years. The work is hard in the way that the work has always been hard. It does not require, and the organisations being told it requires this are being sold something, a wholesale reinvention.

This will be the spine of the book's argument throughout. Where the existing discipline transfers, the chapters will say so plainly and show how. Where the existing discipline needs extension, the chapters will name the extension, describe its mechanics, and show what working practice around it looks like. Where the existing discipline cracks under the weight of an AI-specific problem — and there are places where it does — the chapters will say so honestly and describe what is being tried instead.

1.0.4 The four tiers

The book uses four recurring case-study organisations to ground its arguments in practice. They are introduced fully in Chapter 5, but a one-paragraph sketch of each here will help the reader orient.

Halftone Audio Collective, our Tier 1 organisation, is the podcast network we have already met. Twelve employees, six shows, one technical co-founder wearing every infrastructure hat, and an AI footprint adopted piecemeal by individual producers without central planning. Halftone is the case study for organisations whose AI security challenge is not to build a sophisticated programme but to stop being reckless — to inventory what is in use, to put basic acceptable-use policies around it, to ask vendors the small set of questions that matter at their scale, and to recover from incidents like the voice-cloning misuse without theatre and without losing trust with the people the business depends on. Halftone's arc across the book is not transformation. It is the

much smaller and more achievable goal of becoming a small business that has stopped being a sitting target.

Lexstream Legal Technologies, our Tier 2 organisation, is a two-hundred-and-fifty-person legal-tech vendor in Dublin selling AI contract-review software to law firms. Their product is the AI system; their customers are sophisticated buyers with their own CISOs and their own questionnaires; their regulator is the Irish Data Protection Commission with the EU AI Act now bearing down. Lexstream's security programme is solid on traditional application security and is visibly catching up on AI security, having been pushed into urgency six months before the book opens by a customer-demo incident in which a prospect's CISO crafted a clause that extracted the product's system prompt and a fragment of another tenant's indexed data. Lexstream is the case study for organisations that have a real AppSec discipline and are extending it credibly but not gracefully into the AI layer.

Greenfield & Marrow Stores, our Tier 3 organisation, is an FTSE 250 regional grocery chain operating three hundred and forty stores across the north of England, Scotland, and Northern Ireland. Their AI estate includes in-house demand forecasting, inventory routing, computer-vision CCTV procured before the governance function existed and never since reassessed, and a recently piloted generative-AI assistant for store managers. They have a working AI governance function — eighteen months old, reporting jointly to the CIO and the General Counsel, with a real model risk inventory — and a mature traditional security programme. They are our case study for organisations that are mostly doing it properly, that fail in the predictable places (vendor AI, legacy systems procured before the governance net existed, shadow AI in non-technical teams), and that recover from those failures without theatre. The Greenfield & Marrow chapters are the book's clearest worked picture of what a healthy AI security programme looks like in practice.

Veridian Continental Freight, our Tier 4 organisation, is a Geneva-headquartered global logistics company with sixty thousand employees across seventy countries and a €14 billion revenue line. Their AI estate is mature in places and fragmented in others: route and load

optimisation in production globally, fraud detection saving them tens of millions of euros a year, a customer-shipment-data benchmarking product that raises questions their lawyers are working through, autonomous-vehicle pilots in two yard-operations sites, and a sprawl of agentic workflows in finance, procurement, and customer service that nobody has a complete inventory of. They have a dedicated AI security team of fourteen sitting inside a three-hundred-person global security function. Veridian is our case study for organisations operating AI at scale, where the problem is not whether to do AI security but how to consolidate a fragmented estate into a single defensible posture across multiple jurisdictions and under the enforcement window of the EU AI Act.

The four tiers are not a ladder up which every organisation should be climbing. One of the book's quieter arguments — developed at length in Chapter 5 — is that a small organisation deliberately staying small, with a tight AI footprint and strong vendor discipline, can be more defensible than a mid-sized organisation that has stood up a model registry and a governance committee but cannot enumerate what its product teams have wired into their development environments this week. Tier is not maturity. Tier is the scale and shape of the AI estate the organisation has to defend. Maturity is how well they are defending it, and the relationship between the two is not the straight line the consulting decks like to draw.

1.0.5 The shape of the rest of the book

This volume — the first of the three-volume AI Security set — runs to seventeen chapters in two parts. The operational, governance, and regulatory material described at the end of this section is the work of Volume 2, *Running the Programme*; the templates live in Volume 3, the *Operator's Workbook*.

Part I — the chapters you are currently reading the first of — establishes the foundations. After this chapter, Chapter 2 examines AI systems as security objects, walking the architecture layer by layer and naming the attack surfaces at each seam. Chapter 3 surveys the vulnerability taxonomy, reconciling the three vocabularies a working

programme will encounter — MITRE ATLAS, the OWASP LLM Top 10, and the NIST AI Risk Management Framework's adversarial categories — into a single working map. Chapter 4 covers threat modelling for AI systems, showing how the disciplines you already use stretch and where they crack. Chapter 5 sets out the four-tier maturity rubric the book uses throughout.

Part II is the vulnerability deep-dive: twelve chapters across the AI-specific attack surface — prompt injection, jailbreaks and content-policy evasion, training-data poisoning and provenance, model extraction and inversion, embedding-space attacks, the AI supply chain, and agentic systems and tool-use exploits, then the wider surface of multimodal and vision-language models, RAG and the retrieval layer, fine-tuning and alignment-stripping attacks, and the open-weights versus API trade-off, closing with how to evaluate model robustness. These are the chapters that contain most of the genuinely new material a security professional has to learn.

Volume 2, *Running the Programme*, carries the arc from there. Its opening chapters move to operational practice — vendor due diligence for AI, AI incident response, AI red-teaming, AI tabletop scenarios, and the use of AI itself to defend the rest of the security estate — mapping the new material back onto the operational rhythms a working security function already runs. Its later chapters address governance and the regulatory horizon — the NIST AI RMF and ISO/IEC 42001 frameworks, and the regulatory landscape spanning the EU AI Act, GDPR, sector-specific regimes, the US patchwork, and Singapore's Model AI Governance Framework — and close on maturity, cadence, and what comes next: the annual rhythm of a healthy AI security programme across the four tiers, and the author's closing argument that the practitioners who do well in AI security over the next five years will be the ones who treat it as security first, AI second.

Two reference artefacts ride with the Workbook volume. The glossary (Volume 3, Appendix C) covers the AI-security-specific terms used throughout, with cross-references to the *PSIRT Field Guide*, *Vulnerability Management Field Guide*, and *GDPR Compliance Field Guide* glossaries for terms shared across the Sylvan canon. The maturity rubric (Volume 3, Appendix D) sets out the full AI security maturity

rubric introduced in Chapter 5.

A companion book — the *AI Security Operator's Workbook* — collects the templates, decision trees, scoring matrices, runbooks, and tabletop scripts referenced throughout this book in a form that can be used directly by a working team. Where this book argues a discipline, the workbook provides the artefacts to start practising it.

1.0.6 A note on defensibility

A word on what this book is and is not, for the avoidance of doubt and for the protection of the reader.

This book offers practitioner framings drawn from real operations across the security disciplines its author has practised over more than two decades. The framings are intended to support security professionals in thinking clearly about AI systems and in extending their existing craft to defend them. They are not, and are not intended to be read as, legal advice. They are not regulatory determinations. They are not guarantees of any particular AI security outcome in any particular organisation. The AI security field is moving quickly, the regulatory environment is moving quickly, and the threats are evolving with the underlying technology. A book is a snapshot of working practice at the moment of its writing. The reader carries the responsibility of applying the framings to their own context, of consulting qualified legal counsel on questions of regulatory obligation, and of treating every recommended practice as a starting point for considered judgement rather than as a substitute for it.

What this book does claim — and the claim is bounded — is that the disciplines it describes are recognisable as security disciplines, that they have been practised in production environments by working security functions, and that an organisation that adopts them in a form appropriate to its scale and risk appetite will be in a meaningfully better position to defend its AI systems than one that does not. That is a defensible claim. It is the claim the rest of the book is written to support.

Recap: AI security is the security discipline applied to systems that include AI or ML components — mostly an extension of the craft the field has practised for thirty years, with a narrower and learnable body of genuinely new failure modes layered on top. It is distinct from AI safety, AI ethics, AI policy, and AI governance, and the book uses four recurring organisations — Halftone, Lexstream, Greenfield & Marrow, and Veridian — to ground its arguments across tiers of scale and maturity.

Next: Chapter 2 — AI Systems as Security Objects. Before the security function can defend an AI system, it must understand what kind of object it is defending. Chapter 2 walks the architecture layer by layer — base model, fine-tuning corpus, retrieval layer, system prompt, tool integrations, inference infrastructure, monitoring — and names the attack surface at each seam.

You've just read Chapter 1 of

AI Security

The full book runs 324 pages across 17 chapters,
with a complete back-of-book index.

Published by Sylvan Press, the book imprint of Sylvan Assurance — field-tested, plain-English references for the people who do the work.

Scan to see the full book — and where to buy it.
sylvanassurance.com/go/ai-security

Not ready for the full book? See where you stand first — free, a few minutes:

sylvanassurance.com/go/free-ai-security-assessment