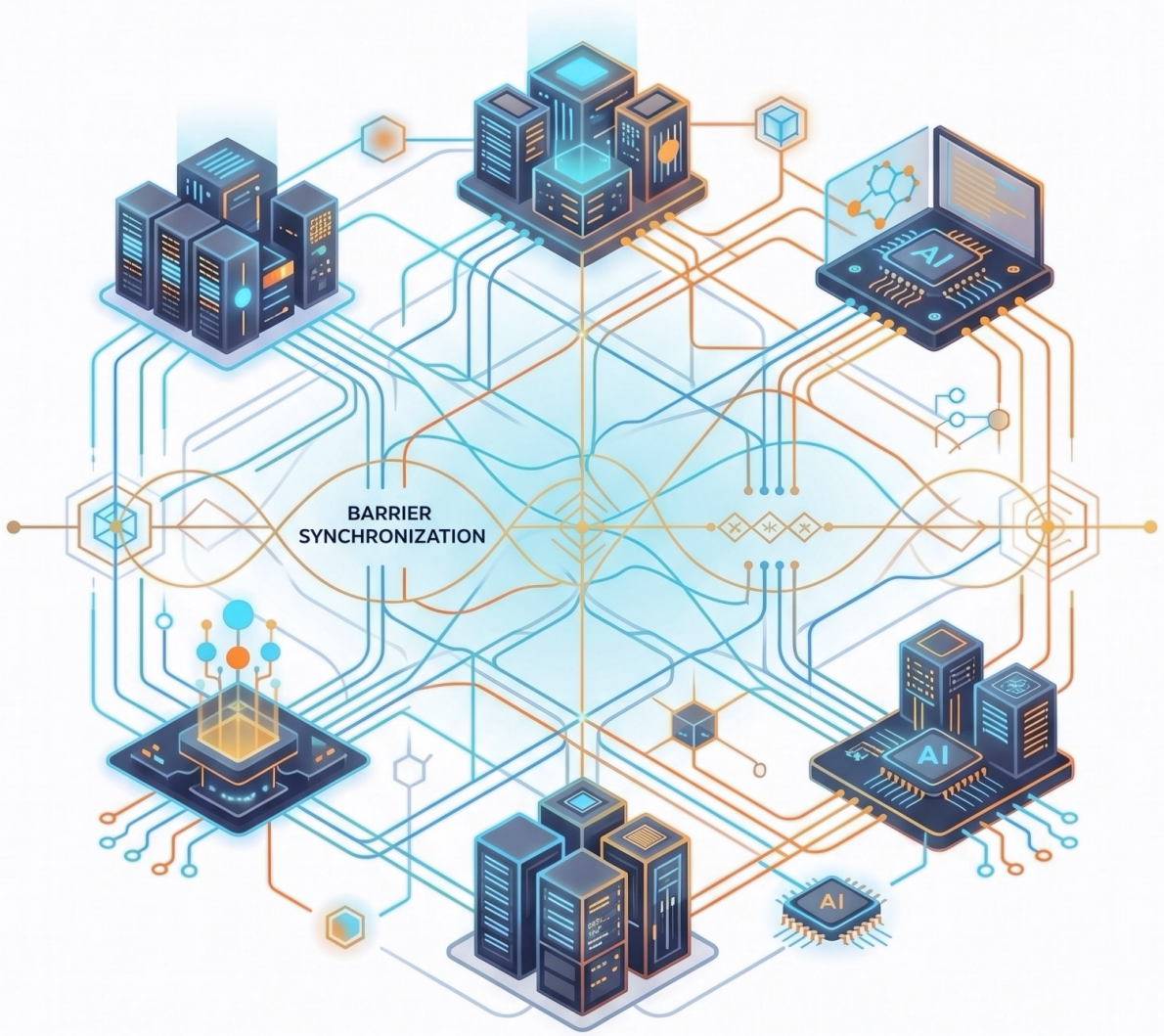


BEYOND UPTIME

RELIABILITY ENGINEERING FOR AI INFRASTRUCTURE



SREEJITH KAIMAL

Copyright © 2026 All Rights Reserved

Tweet This Book!

Please help Sreejith Kaimal by spreading the word about this book on [Twitter](#)!

The suggested hashtag for this book is [#forensic-ai](#).

Find out what other people are saying about the book by clicking on this link to search for this hashtag on Twitter:

[#forensic-ai](#)

Contents

Beyond Uptime : Reliability Engineering for AI Infrastructure	ii
Disclaimer & Limitation of Liability	v
1: How AI Infrastructure Actually Works	1
What is an AI Model	1
Why this requires GPU's rather than CPU's	2
Visualizing the AI Factory Floor	2
How Tiny Delays Compound in AI	4
The Modern AI Stack : A Glossary Through the Same Factory Lens . .	5
Beyond Network Rules : A Systems-Thinking Argument	5
2: Operation Slowburn	6
How can a system be healthy and still be failing?	6
3: The Synchronization Problem	7
An Intuitive Model: Barrier Synchronization	7
From the Table to the Training Run	7
Why this is unavoidable	7
A Worked Capacity-Planning Illustration	7
4: Operation Slowburn : A Forensic Reconstruction	8
Incident Premise and Cluster Configuration	8
The Forensic Investigation	9
5: Specialized Networking Technologies	11
Limits of Ordinary Networking	11
RDMA, RoCE, and InfiniBand	11
A Cautionary Case: Fast Transport Is Not Sufficient Alone	11
Comparing the Three Fabrics Along the Dimensions That Matter . . .	11

Transport Reliability Semantics: What RDMA Guarantees and What It Does Not	11
6: Lossless Fabric Bottlenecks	13
The Congested Highway	13
Priority Flow Control and Cascading Pause Frames	13
Hash Polarization: A Second Congestion Mechanism	13
The Diagnostic Implication	13
7: The Anatomy of Silent Data Corruption	15
Transient Voltage Fluctuations & Thermal Noise	15
Cosmic Radiation & Single-Event Upsets (SEUs)	15
Manufacturing Variance & “Mercurial Cores”	15
Passive Hardware Telemetry Checklist	15
Memory Subsystem Degradation	15
The Structural Argument	15
8: Systematic RCA - The Failure Graph	17
The Fleet Aggregation Paradox	17
Observability as Evidence Collection	17
Systematic , Layer-by-Layer Method	17
A Divergent Presentation of the Same Failure Class	17
The Failure Graph as a Formal Structure	17
9: Generalizing the Incident :A Practitioner Framework	18
Why Generalization Is Possible at All	18
From Forensics to Foresight	18
Toward Automated Detection	18
A Concrete Detection Formalism: From Fleet Averages to Per-Rank Tail Statistics	19
Preventive Engineering Practices	19
Organizational Practices: Building the Team Around the Method	19
10: AI Investigating AI : Towards Autonomous Infrastructure	20
Machine Learning as the Investigator, Not Only the Instrument	20
Toward Autonomous AI Infrastructure	20
An example of a Harnessed Agent Cluster running an AI model	20
An Agenda for Further Research	20

References 21

Beyond Uptime: Reliability Engineering for AI Infrastructure

Sreejith Kaimal

Beyond Uptime : Reliability Engineering for AI Infrastructure

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

Beyond Uptime: Reliability Engineering for AI Infrastructure

This book is available at <https://leanpub.com/forensic-ai>

This version was published on 2026-09-27 ISBN 979-8-9984068-9-8



This is a [Leanpub](#) book. Leanpub empowers authors and publishers with the Lean Publishing process. [Lean Publishing](#) is the act of publishing an in-progress ebook using lightweight tools and many iterations to get reader feedback, pivot until you have the right book and build traction once you do.

© 2026 Sreejith Kaimal

Disclaimer & Limitation of Liability

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

Original Research & Independent Works

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

Reader Discretion & Assumption of Risk

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

No Warranties & Exclusion of Liability

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

1 - How AI Infrastructure Actually Works

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

1.1 - What is an AI Model

An AI model is not a reasoning mind; instead, it is an incredibly large pattern-finding system. It is a computer program trained against millions or billions of examples, words, images, sounds, or videos, from which it does not memorize individual examples but instead extracts statistical regularities, mathematical relationships between inputs and the outputs those inputs tend to produce. When the trained model is later given a new prompt, it draws on those extracted regularities to predict the most probable continuation, not by consulting anything resembling stored facts, but by computing a probability distribution over possible next outputs and selecting from it. A model does not think in the sense a human does; it performs, at a very large scale, a very sophisticated form of statistical prediction. Picture this mechanism concretely as an enormous equalizer board, the kind used in audio mixing, except instead of a few dozen sliders, this board has billions or trillions of them, each called a parameter or a weight.

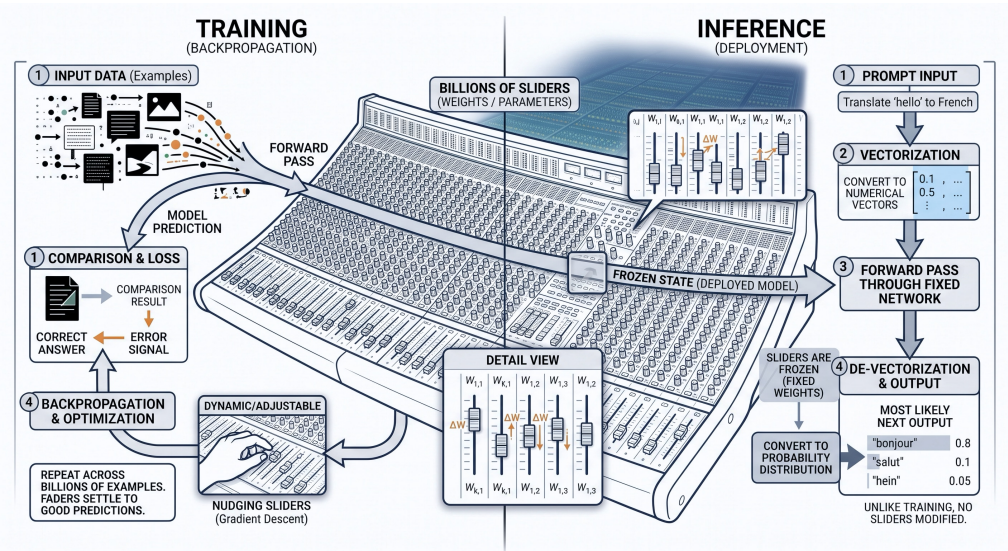


Figure 1.1. Non Technical View of an AI Model

During training, the model is shown an example and produces a guess, for instance, a predicted next word in a sentence, and is told whether that guess was correct. When it is wrong, a procedure called back propagation nudges a very large number of these sliders by a small amount in the direction that would have made the correct answer slightly more probable next time¹. This nudging repeats across billions of examples until the sliders settle into a configuration that produces good predictions consistently. Once training concludes, those sliders are frozen; this frozen state is what practitioners mean when they refer to a **model's weights**. At inference time, a prompt is converted into numerical vectors, passed through this enormous fixed network of sliders, and converted back into a probability distribution over the most likely next output, a process that, unlike training, does not modify any of the sliders themselves.

1.2 - Why this requires GPU's rather than CPU's

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

¹Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. Nature 1986, 323, 533–536.

1.3 - Visualizing the AI Factory Floor

Later in this chapter, the analogy of a huge, fast factory assembly line comes up again and again, and it's worth stating it explicitly and completely here in order to make every subsequent discussion on synchronization, on congestion, and on silent hardware degradation much easier to follow.

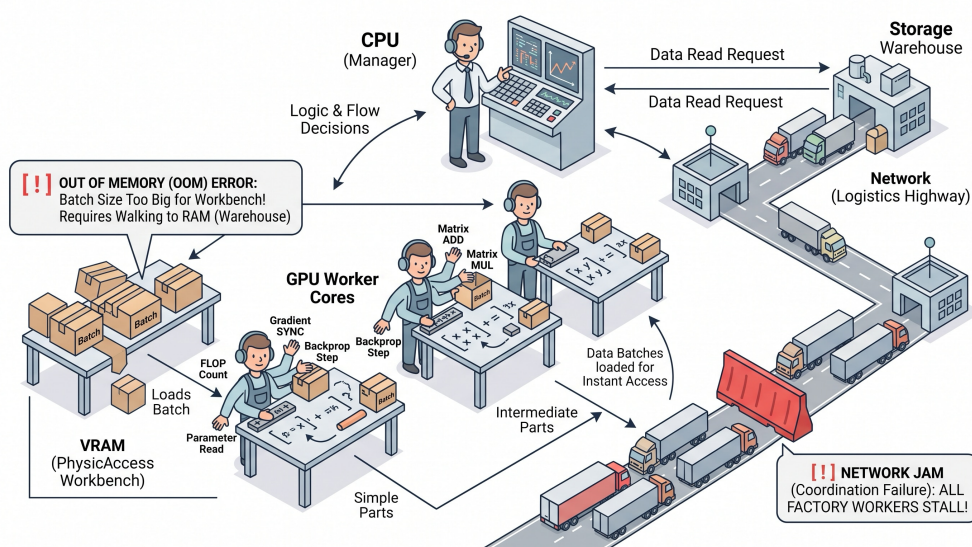


Figure 1.2. Graphic Representation of the Idea (Generated in Collab by G.Gemini)

The CPU will now act as the factory manager. It is smart and very flexible, but it cannot do the manufacturing work itself; it is only used to make a decision about the logic, the flow of work, and what to do next.

The GPU acts as the workforce of a factory where each worker is simply doing the same part of the job, but they are capable, as a group, of doing this at vast numbers and extremely high efficiency, in this case, a simple arithmetic operation.

The GPU's built-in memory (VRAM) acts as the workbench directly next to each worker. Any raw material put on this workbench is instantly available to the worker; if the workbench is too small to work on, the worker's choice is to either walk to a distant warehouse or ordinary storage to retrieve more material or to walk back to the workbench. This is exactly how the Out of Memory error occurs when a training job's batch size (amount of material loaded onto the workbench at one time) is larger than the physical capacity

of the workbench, and a practitioner reduces the batch size or adds GPUs with larger VRAMs, a larger workbench.

Networking is like a logistical freeway that links several small factory structures that are spread out around the country, with each building being a single physical server node with a lot of GPUs in it. The significance of this role is enormous to the argument that follows in the rest of this chapter: even if each manager and every worker in each individual factory building were working at perfect efficiency, the moment the trucks carrying intermediate parts between factories got stuck in the traffic jams on the road the entire multi-factory production line would come to a standstill, regardless of how each building was doing on its own. This same intuition is captured in Section 3, albeit in a much more explicit manner, in the form of barrier synchronization.

This same factory language applied to the data path used to move one batch of training data from disk to a GPU's workbench, breaks the process into three specific steps, and these steps are each named explicitly because they will be encountered in the following text, under their appropriate technical names.

1. **Data Loading:** the manager of the factory, the CPU, gets raw materials from the warehouse. The training framework makes a read request to the operating system, asking it to return the addresses in the physical storage where the data is located, and the CPU puts the request on a list and passes it to the storage drive's own hardware controller, which instructs it to start performing the transfer.
2. **Data staging:** The material is in the storage drive's own controller, and with a method known as Direct Memory Access, it is moved directly to the workbench (or host system memory) and only notifies the CPU when the transfer is finished, while the factory manager is not involved in the transfer process and can attend to other work.
3. **Data Processing:** processing of the delivered material, the mass processing, is performed by the assembly line laborers (the GPU), where the parameters of the model would be updated as outlined in Section 1.1, based on the thousands of delivered units. Not just a delay inside the GPU, but anywhere of the three delays above can spell trouble for a whole fleet of healthy workers, as the worked example that follows in Section 3.2 shows when training is CPU-bound.

1.4 - How Tiny Delays Compound in AI

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

1.5 - The Modern AI Stack : A Glossary Through the Same Factory Lens

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

1.6 - Beyond Network Rules : A Systems-Thinking Argument

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

2 - Operation Slowburn

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

2.1 - How can a system be healthy and still be failing?

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

3 - The Synchronization Problem

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

3.1 - An Intuitive Model: Barrier Synchronization

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

3.2 - From the Table to the Training Run

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

3.3 - Why this is unavoidable

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

3.4 - A Worked Capacity-Planning Illustration

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

4 - Operation Slowburn : A Forensic Reconstruction

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

4.1 - Incident Premise and Cluster Configuration

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

4.1.0.1: What are these components ?

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

4.1.0.1.1: Confined NVLINK DOMAIN (TP=8)

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

4.1.0.1.2: Two-Tier Leaf-Spine FAT-Tree (200 Gbps RoCE v2)

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

4.1.0.1.3: Single-Hop Rail-Optimized Nic-Switch Mapping

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

4.1.0.1.4: GPUDIRECT RDMA & KERNEL-BYPASS

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

4.1.0.1.5: O(N) Latency Scaling in NCCL Default Ring

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

4.1.0.1.6: Silent PCIe Link Downtraining (GEN 5 x16 → GEN 3 x4)

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

4.1.0.1.7: IEEE 802.1Qbb PFC Pause-Frame Cascades & HEAD-OF-LINE (HoL) Blocking

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

4.1.1: The Blind Spot of Single-Node Telemetry

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

4.1.1.0.1: Host CPU Telemetry and the RDMA Paradox

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

4.1.1.0.2: Node-Local Integrity vs. Cluster Coordination

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

4.2 - The Forensic Investigation

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

4.2.0.1: Finding the Ring Topology Bottleneck

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

4.2.1: What did we learn & next steps

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

4.2.1.1: Implement In-Band & Switch-Level Telemetry (gNMI / INT)

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

4.2.1.2: Mandate Pre-Flight Synthetic Canaries

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

4.2.1.3: Shift from Static Thresholds to Statistical Tail Tracking

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

4.2.1.4: Decouple Containment from Hardware Replacement

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

4.2.2: Suggested Multi-Layer Monitoring Reference Matrix

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

4.2.3: From Reactive Firefighting to Architectural Discipline

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

5 - Specialized Networking Technologies

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

5.1 - Limits of Ordinary Networking

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

5.2 - RDMA, RoCE, and InfiniBand

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

5.3 - A Cautionary Case: Fast Transport Is Not Sufficient Alone

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

5.4 - Comparing the Three Fabrics Along the Dimensions That Matter

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

5.5 - Transport Reliability Semantics: What RDMA Guarantees and What It Does Not

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

6 - Lossless Fabric Bottlenecks

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

6.1 - The Congested Highway

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

6.2 - Priority Flow Control and Cascading Pause Frames

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

6.3 - Hash Polarization: A Second Congestion Mechanism

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

6.4 - The Diagnostic Implication

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

6.4.0.1: The Limits of Traditional Interface Telemetry

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

6.4.0.2: Four Reasons Lossless Congestion Escapes Standard Monitoring

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

6.4.0.3: The Required Telemetry Shift

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

7 - The Anatomy of Silent Data Corruption

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

7.1 - Transient Voltage Fluctuations & Thermal Noise

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

7.2 - Cosmic Radiation & Single-Event Upsets (SEUs)

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

7.3 - Manufacturing Variance & “Mercurial Cores”

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

7.4 - Passive Hardware Telemetry Checklist

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

7.5 - Memory Subsystem Degradation

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

7.6 - The Structural Argument

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

8 - Systematic RCA - The Failure Graph

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

8.1 - The Fleet Aggregation Paradox

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

8.2 - Observability as Evidence Collection

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

8.3 - Systematic , Layer-by-Layer Method

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

8.3.1: Diagnostic Method Rules

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

8.4 - A Divergent Presentation of the Same Failure Class

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

8.5 - The Failure Graph as a Formal Structure

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

9 - Generalizing the Incident :A Practitioner Framework

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

9.1 - Why Generalization Is Possible at All

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

9.2 - From Forensics to Foresight

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

9.2.1: A Composite Health Score & A Circuit Breaker

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

9.2.1.1: State Definitions & Lifecycle Dynamics

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

9.2.1.2: Prevention of Threshold Flapping (Hysteresis)

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

9.3 - Toward Automated Detection

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

9.4 - A Concrete Detection Formalism: From Fleet Averages to Per-Rank Tail Statistics

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

9.5 - Preventive Engineering Practices

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

9.6 - Organizational Practices: Building the Team Around the Method

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

10 - AI Investigating AI : Towards Autonomous Infrastructure

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

10.1 - Machine Learning as the Investigator, Not Only the Instrument

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

10.2 - Toward Autonomous AI Infrastructure

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

10.3 - An example of a Harnessed Agent Cluster running an AI model

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

10.4 - An Agenda for Further Research

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.

References

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/forensic-ai>.