

The MYTHOS Playbook, Volume I

Joseph P. Conroy

2026

Contents

The MYTHOS Playbook, Volume I	1
The Adversary Landscape	1
Copyright	2
Two-Volume Set — Multi-Format Edition	2
A Note on the EPUB Edition	2
Legal and Editorial Disclosures	2
Publisher	3
Preface	4
A Note on the Two-Volume Structure	5
Acknowledgments	6
Institutional Acknowledgments	6
VectorCertain Engineering	6
Family	6
Part I — Foundations	7
Chapter 1: Why a Playbook, Why Now: The April 2026 Inflection Point	8
Chapter 1 Purpose	8
1.1 The Sandwich Incident and What Frontier Cybersecurity Capability Actually Looks Like	8
1.2 The Treasury / Federal Reserve Emergency Consultation and the Glasswing Distribution	10
1.3 The Lethal Trifecta — Why an Agent Is Not an Endpoint	11
1.4 Why EDR, XDR, and SIEM Are Structurally Inadequate for Agentic Threats	13
1.5 The Performance-Guarantee Void and DARPA’s Acknowledgment	15
1.6 The Mythos Threat Taxonomy as a Named, Quantifiable Threat Model	16
1.7 How This Book Is Organized and How to Use It as a Desk Reference	18
1.8 The Editor’s Note on Voice and Citation Discipline	18
Without SecureAgent: What You Can Do Today	19
1.9 Summary and What the Reader Should Take From Chapter 1	20
CISO’s Five Next Steps	20
References	21
Cross-References	21
Chapter 2: The MYTHOS Threat Taxonomy: T1–T7 at a Glance	22
2.1 The Seven Vectors: T1–T7	22
A note on naming	23
A note on case studies	23
What the seven vectors do not cover — and what catches those risks	24
2.2 Threat Vectors vs. Vulnerability Classes: Why Seven, Not Ten	24
Why intent-organization is CISO-actionable where mechanism-organization is not	25
The seven vectors are not a competitor to OWASP; they are an operating layer above it	25
2.3 The Cross-Reference Matrix	26
Table 2.3.1 — MYTHOS T1–T7 × MITRE ATT&CK × MITRE ATLAS × OWASP × NIST AI RMF × CRI FS AI Profile	26
Table 2.3.2 — MYTHOS T1–T7 × ISO/IEC 42001 (Annex A control reference)	29

How to use the matrix	29
Where the matrix is incomplete on purpose	30
The matrix as a living artifact	30
Reading the matrix the way a CISO reads a P&L	30
2.4 Reading the Threat-Vector Chapters: The MYTHOS Technique-Page Template	30
The eight sections	31
What the template does not do	33
A worked example: T2 in the institution’s existing OWASP review cycle	33
2.5 Per-Vector Counts and the Forward Contract to Part III	33
T1 — Autonomous Multi-Step Exploitation	34
T2 — Unsanctioned Scope Expansion	34
T3 — Invisible Deceptive Reasoning	34
T4 — Track-Covering Log Manipulation	35
T5 — Credential Theft & System Access	35
T6 — Sandbox Escape Exploitation	36
T7 — Capability Proliferation	36
Per-vector arithmetic and what the cross-totals mean	37
The forward contract to Part III	37
2.6 <i>Without SecureAgent: What You Can Do Today</i>	38
Why the sidebar is structurally honest	39
2.7 Summary	39
Five Next Steps	39
Cross-References	40
References	40

Chapter 3: The MYTHOS Certification Program: What “99.0% Recall at 3-Sigma” Actually Means

	41
3.1 1,000 Scenarios × 7 Vectors = 7,000 Adversarial Tests: The Validation Architecture	41
3.1.1 What the seven vectors are	42
3.1.2 What the validation architecture is and is not	43
3.1.3 Why dedup matters and what dedup is not	43
3.1.4 The two surfaces — MYTHOS and TES — and what they each measure	44
3.2 The Crumpton/MITRE Category-Creation Framing — Why the TES Result Is Internal Validation, Not a MITRE-Published Score	44
3.2.1 What the canonical TES citation looks like	45
3.2.2 What the TES surface measures	45
3.3 How to Read a Confusion Matrix the Way a CISO Reads a P&L	46
3.3.1 The four cells	46
3.3.2 Recall, precision, specificity, F1	46
3.3.3 An illustrative worked example	47
3.3.4 What the MYTHOS confusion matrices look like	47
3.4 Why Recall Is the Right Primary Metric for Adversarial Security	48
3.4.1 Adversarial security is a rare-event regime	48
3.4.2 But you cannot ignore false positives	48
3.4.3 Why “99.0% recall” rather than “100% recall”	49
3.4.4 Clopper-Pearson vs. Wald vs. Wilson — why conservative wins in compliance	49
3.4.5 What 99.0% at three-sigma means in plain English	50
3.4.6 Why the title says “99.0%” and the bound says “99.65%”	50
3.4.7 The base-rate fallacy in plain words	50
3.5 Benchmarking Vendor Claims Statistically: A CISO Field Test	51
3.5.1 Question one: what is your sample size?	51
3.5.2 Question two: what is your statistical method?	51
3.5.3 Question three: what is your false-positive rate?	51
3.5.4 Question four: what are the per-vector breakdowns?	52
3.5.5 Question five: what is your annotation policy?	52
3.5.6 What a complete vendor disclosure looks like	52
3.5.7 The reverse field test — what your own evidence has to look like	53

3.6 Service-Credit Terms as Statistical Instruments Backed by Empirical Confidence Bounds	54
3.6.1 What a service-credit term is, in statistical language	54
3.6.2 The statistical anchor	54
3.6.3 Why this is a statistical instrument and not a guarantee	54
3.6.4 What the CISO asks counsel	55
3.6.5 What the threshold should look like in practice	55
3.6.6 What the threshold should not look like	56
3.7 Internal vs. Third-Party-Reproduced: The Annotation Policy Used Throughout the Book	56
3.7.1 Two categories, one policy	56
3.7.2 Why the policy is the differentiator	56
3.7.3 Where the annotation appears	57
3.7.4 The TES citation as the canonical example	57
3.7.5 The pointer fallback is permitted only for narrative cross-references	57
3.7.6 The MR-1 corpus, ER7 cohort, and HOTS surfaces — annotation worked examples	58
Summary	58
CISO’s Five Next Steps	59
Cross-References	59
References	60

Part II — Architecture **61**

Chapter 4: Pre-Execution Governance vs. Detect-and-Respond: The Architectural Shift **62**

Chapter 4 Purpose	62
4.1 The Twenty-Nine-Minute Breakout Problem and Why It Is Unrecoverable for Detect-and-Respond	63
4.1.1 What the breakout-time figure actually measures	63
4.1.2 The autonomous-agent decision loop is sub-second	63
4.1.3 The recoverability assumption	63
4.1.4 The cumulative argument	64
4.2 The OWASP Principle: Autonomy Is a Feature Earned, Not a Default	64
4.2.1 The capability-versus-authority distinction	64
4.2.2 Why “default authority” is structurally unsafe	65
4.2.3 What “earning” authority looks like operationally	65
4.2.4 The competitive baseline: Goose A/B	66
4.3 The Crumpton Email: Why Pre-Execution Governance Is Categorically Different from Post-Execution Detection	66
4.3.1 The LONG FORM canonical paragraph	67
4.3.2 What the paragraph establishes structurally	67
4.3.3 The categorical distinction made concrete: ER7 0% identity vs. internal 100%	68
4.3.4 What the paragraph forbids	68
4.4 OCI / Invariants — Structured Action Representation as the Precondition for Runtime Enforcement	69
4.4.1 What “structured action representation” means	69
4.4.2 The arXiv runtime-governance literature	69
4.4.3 Three concrete invariants	70
4.4.4 What this implies for procurement	70
4.5 The Two-Surface Model: Govern What Goes In / Govern What Comes Out	71
4.5.1 Surface 1 — Memory admission (“govern what goes in”)	71
4.5.2 Surface 2 — Behavioral execution (“govern what comes out”)	71
4.5.3 Why both surfaces are architecturally required	72
4.5.4 The bridge to Chapters 5, 6, and 7	73
Without SecureAgent: What You Can Do Today	73
4.6 Summary and What the Reader Should Take From Chapter 4	74
CISO’s Five Next Steps	75
References	76

Chapter 5: AMRS V4 — Governing What Goes In **77**

Chapter 5 Purpose	77
5.1 Why the Agent’s Memory Is the First Attack Surface	77

5.2 Cross-Session Threat Models and Why a SIEM Query Fires Too Late	79
5.3 AMRS V4 Admission Controls: The Scoring Formula, ARSC Tiers, P-gate / V-gate, BG-1 Through BG-10, and the Six-Test Validation Suite	81
5.3.1 The Four Locked Weight Vectors	81
5.3.2 The Composite Scoring Formula	82
5.3.3 The Four ARSC Tiers and the Per-Tier Threshold Floors	82
5.3.4 The P-gate (Provenance Gate)	83
5.3.5 The V-gate (Validation Gate)	84
5.3.6 The D-1 Signed-Artifact Workflow	84
5.3.7 The Ten Behavioral Guarantees BG-1 Through BG-10	85
5.3.8 The Six-Test Validation Suite	86
5.3.9 Audit Storage by Tier	87
5.3.10 The V4 Dynamic-Recalibration Architecture	87
5.4 Integrating AMRS With Existing Vector Databases and RAG Pipelines	88
5.4.1 The Two Integration Patterns	88
5.4.2 The Vector-Database Schema Annotation	88
5.4.3 The Retrieval-Augmented-Generation Pipeline Hook	89
5.4.4 Throughput, Cardinality, and the Living Corpus	89
5.5 Performance Envelope: Per-Gate Latency Under One Millisecond	89
5.5.1 The Per-Gate Latency Decomposition	90
5.5.2 Throughput at Enterprise Scale	90
5.5.3 The Performance-Versus-Defensibility Trade	91
5.6 <i>Without SecureAgent: What You Can Do Today</i>	91
5.7 Summary	92
Five Next Steps	93
References	93
Cross-References	93
Chapter 6: The Four-Gate Action Pipeline: HES1-SG, HCF2-SG, TEQ-SG, MRM-CFS-SG	95
Chapter 6 Purpose	95
6.1 Gate 1 — HES1-SG (Hybrid Ensemble System — Safety & Governance): Candidate Diversity . .	95
6.2 Gate 2 — HCF2-SG (Hierarchical Cascading Framework — Safety & Governance): Epistemic Trust	97
6.3 Gate 3 — TEQ-SG (Trust & Execution Governance — Safety & Governance): Numerical Admissibility	99
6.4 Gate 4 — MRM-CFS-SG (Micro-Recursive Model — Cascading Fusion System — Safety & Governance): Execution Governance	101
6.5 Pipeline Latency Budget — Two Clocks, Not One	103
6.6 Integration Patterns — Wrapping Tool APIs vs. Interposing at the Model Endpoint	105
6.7 A Worked End-to-End Example: Four Gates Firing in Sequence	107
<i>Without SecureAgent: What You Can Do Today</i>	108
Summary	109
CISO's Five Next Steps	110
References	110
Forward References	110
Chapter 7: AGL-SG and the GTID Tamper-Evident Audit Chain	111
Chapter 7 Purpose	111
7.1 Why Standard Database Audit Logs Are Not Tamper-Evident	111
7.2 Hash-Chain Architecture: The GovernanceChain Construction	113
7.2.1 The Genesis Record	113
7.2.2 The Previous-Hash Linkage	113
7.2.3 RFC 8785 JSON Canonicalization	114
7.2.4 The GovernancePolicyPin	115
7.2.5 Layer Independence — The LayerDetermination Structure	115
7.3 SIEM Integration: From Compliance Artifact to Detection Capability	116
7.3.1 The Ingest Path	116
7.3.2 SOC Correlation Patterns	116

7.3.3 The Ingest-Volume Question	117
7.4 GTID Structure: The Seven-Element Audit Record	117
7.4.1 The Seven Elements	117
7.4.2 The Comprehensiveness Discipline — INHIBITED Records Are First-Class	118
7.4.3 The Authorization-Token Linkage	118
7.4.4 The AppendOnlyViolation Behavior	119
7.4.5 Hash Determinism	119
7.5 Optional Cryptographic Signatures: ECDSA P-256 over the Chain	120
7.6 Court-Admissibility, EU AI Act Article 12, and Examination Readiness	120
7.6.1 Court-Admissibility — The Federal Rules of Evidence	121
7.6.2 EU AI Act Article 12 — Automatic Recording of Events	121
7.6.3 Examination Readiness — CRI FS AI RMF and FFIEC	122
7.6.4 The Insurance Underwriting Implication	122
7.7 Containment GTIDs and Inter-Agent References	122
7.8 Without SecureAgent: What You Can Do Today	123
7.9 Summary	124
CISO's Five Next Steps	124
References	125
Chapter 8: The 828-Model MRM-CFS Ensemble: Cascading Fusion Architecture Reference	126
Chapter 8 Purpose	126
8.1 What the MRM-CFS Is and How It Differs from a Monolithic Classifier	127
8.2 The 828-Model Ensemble: Structure, Training-Data Provenance, Vector Coverage	128
8.2.1 The Two-Tier Classify→Quantify Architecture	129
8.2.2 The 828 Segments — Composition	129
8.2.3 Training-Data Provenance	130
8.2.4 Vector Coverage	131
8.3 Functional Classifier Groups, Voting, Abstention, Statistical Independence Assumption	132
8.3.1 The Functional Classifier Groups	132
8.3.2 Voting and Abstention	134
8.3.3 The Statistical-Independence Assumption	135
8.3.4 Worked Example — A T4 Audit-Record Manipulation Trace Through the Ensemble	136
8.4 Why 828 and Not 13, 100, or 10,000: The Bias-Variance Argument at Scale	137
8.4.1 The Bias-Variance Tradeoff in Ensemble Detection	137
8.4.2 Why an N of 13 Is Insufficient	138
8.4.3 Why 100 Is Insufficient	138
8.4.4 Why 10,000 Is Wasteful	139
8.4.5 Why 828 Is the Right Cardinality	139
8.4.6 The Cardinality and Adversarial Robustness	140
8.4.7 Continuous Cardinality Validation	140
8.5 Updating the Ensemble via ZGTID Without Service Interruption	140
8.5.1 The Update Cycle	141
8.5.2 Why the Update Mechanism Requires ZGTID	141
8.5.3 Update Cadence and Validation Discipline	142
Without SecureAgent: What You Can Do Today	142
Summary	143
CISO's Five Next Steps	144
References	145
Cross-References	145
Chapter 9: HOTS: Eliminating Baseline LLM Over-Compliance	147
Chapter 9 Purpose	147
9.1 What Over-Compliance Is and Why It Is a Security-Relevant Failure, Not an Alignment Curiosity	148
9.1.1 The compliance-versus-competence distinction	148
9.1.2 H-Neurons: the pre-training origin of the failure mode	148
9.1.3 Shared-safety neurons and why the surface is sparse	149
9.1.4 The financial-services consequences of over-compliance	149

9.1.5 Why over-compliance is not solved by “use a better model”	149
9.2 The HOTS Baseline Cohort: 1,010 Trials, 4 Phenotypes, 81.4 Percent Cross-Model Correlation . .	150
9.2.1 The 1,010-trial cohort: composition	150
9.2.2 Locked execution parameters: matching the Gao et al. methodology	150
9.2.3 The 40 percent baseline: what an unguarded frontier agent does under adversarial framing	151
9.2.4 The H-Neuron Homology Hypothesis: 81.4 percent average cross-model failure correlation	151
9.2.5 The 0.0 percent S1 compliance rate under SecureAgent governance	152
9.2.6 Why the S2/S3 thresholds matter	152
9.3 HOTS Phenotype Taxonomy: HOTS-01 Through HOTS-04 and the SecureAgent S1 Result	152
9.3.1 HOTS-01: false-premise acceptance	153
9.3.2 HOTS-02: context-injection override	153
9.3.3 HOTS-03: sycophantic reversal	154
9.3.4 HOTS-04: safety-guardrail bypass	155
9.3.5 The HOTSResult schema and the S1 escalation protocol	155
9.4 Running HOTS on Your Own Deployed Models: The Practitioner Workflow	156
9.4.1 Step one: scope the run against the institution’s action registry	156
9.4.2 Step two: construct the ABTestHarness pairing	156
9.4.3 Step three: lock the execution parameters	157
9.4.4 Step four: capture HOTSResult records under deterministic conditions	157
9.4.5 Step five: compute the institution-specific over-compliance surface	157
9.4.6 Step six: act on the result via D-1-mediated weight-vector recalibration	158
9.4.7 Cohort-sizing arithmetic and the Clopper-Pearson upper bound on the 0.0 percent figure	158
9.5 HOTS and HES1-SG: How Behavioral Testing and Runtime Enforcement Reinforce Each Other . .	159
9.5.1 What HES1-SG does at runtime	159
9.5.2 Why HOTS measurements feed HES1-SG configuration	159
9.5.3 Behavioral testing detects what runtime enforcement cannot reach	160
9.5.4 The AMRS HOTS adversarial cross-reference	160
9.5.5 The architectural commitment: behavioral testing is a permanent surface	161
9.6 <i>Without SecureAgent: What You Can Do Today</i>	161
9.7 Summary	162
Five Next Steps	162
References	163
Cross-References	163

Chapter 10: ZGTID — Zero-Knowledge GTID and Privacy-Preserving Threat-Intelligence Dis-	164
tribution	
Chapter 10 Purpose	164
10.1 Why Signature-Based Distribution Is Not Fast Enough for Agentic Attacks	165
10.2 ZGTID Record Format and Cryptographic Signing	166
10.2.1 The Five-Element ZGTID Schema	166
10.2.2 The ECDSA P-256 Signature Layer	167
10.2.3 The Zero-Knowledge Predicate	168
10.2.4 Canonicalization and Hash Determinism	169
10.3 Integration with STIX/TAXII and Commercial TIP Stacks	169
10.3.1 STIX Representation: ZGTID as a Custom STIX Object	169
10.3.2 TAXII Transport: ZGTID Over the TAXII 2.1 Channel	170
10.3.3 Commercial TIP Stack Integration	171
10.4 Privacy-Preserving Intelligence Sharing Across Tenant Boundaries	171
10.4.1 What a Recipient Tenant Learns	171
10.4.2 What a Recipient Tenant Cannot Learn	172
10.4.3 Cross-Border and Jurisdictional Considerations	172
10.5 Governance: ZGTID as the Operational Layer Beneath CRI FS AI RMF	173
10.5.1 Detection — Anomalies and Events (DE.AE Family)	173
10.5.2 Information Sharing (ID.RA Family)	173
10.5.3 Audit and Log Records (PR.PT Family)	173
10.5.4 Identity and Access Management (PR.AC Family)	174
10.5.5 The Examination Conversation	174

10.6 Without SecureAgent: What You Can Do Today	174
10.7 Summary	176
CISO's Five Next Steps	176
References	177

Chapter 11: VectorTransForm and Real-Time Compliance Tagging 178

Chapter 11 Purpose	178
11.1 The Telemetry Quality Problem in Agentic Environments	179
11.2 The VectorTransForm Pipeline: Extraction, Normalization, Enrichment	180
11.2.1 Stage 1 — Extraction	180
11.2.2 Stage 2 — Normalization	181
11.2.3 Stage 3 — Enrichment	181
11.2.4 The Closed Audit Loop — Text → F32-256 → Text	182
11.2.5 Validation and Determinism	182
11.3 The 508-Point Unified Governance Mapping	182
11.3.1 Why Two Frameworks, Not One	183
11.3.2 The Single Decision-Record Schema	183
11.3.3 The G1 + G4 = G5 Discipline	184
11.3.4 Pre-Execution Tagging — The Operational Constraint	184
11.4 Real-Time Compliance Tags: One Determination, Seven Framework Regimes	185
11.4.1 The Cross-Walk Matrix	185
11.4.2 NIST CSF 2.0 — The Functional Anchor	187
11.4.3 FFIEC CAT — The U.S. Banking Examination Surface	187
11.4.4 PCI DSS 4.0 — The Card-Industry Surface	187
11.4.5 SOC 2 — The Service-Organization Surface	188
11.4.6 ISO/IEC 27001 and ISO/IEC 42001 — The International Surface	188
11.4.7 CRI FS AI RMF — The Treasury-Aligned Anchor	189
11.4.8 The Operational Property — One Determination, Seven Projections	189
11.5 Evidence Artifacts for the Examiner	190
11.5.1 The Five Examination Artifacts	190
11.5.2 The Examiner's Sample-and-Verify Pattern	190
11.5.3 Real-Time Compliance and the Examination Calendar	191
11.6 Without SecureAgent: What You Can Do Today	191
11.7 Summary	192
CISO's Five Next Steps	193
References	193

Chapter 12: VectorAgents: Deployment Economics on Existing Hardware 194

Chapter 12 Purpose	194
12.1 Why Governance-on-Existing-Hardware Matters for Agentic Scale	195
12.1.1 The deployment-surface argument	195
12.1.2 Why agentic scale changes the cost calculus	196
12.1.3 Why “existing hardware” is the architectural premise, not a sales argument	197
12.2 Micro-Recursive Model Architecture: Extending Coverage Into Statistical Tails	197
12.2.1 Per-segment size and per-segment latency	197
12.2.2 The 2-tier Classify→Quantify cascade	198
12.2.3 Extending coverage into the statistical tails	198
12.3 Deployment Topologies: Edge, Gateway, Sidecar, and Orchestrator Patterns	198
12.3.1 Edge topology	199
12.3.2 Gateway topology	199
12.3.3 Sidecar topology	199
12.3.4 Orchestrator topology	200
12.3.5 The topology-mix discipline	200
12.4 The Deployment-Cost Reduction Claim: What Is in the Numerator and the Denominator	201
12.4.1 The denominator: dedicated AI accelerator infrastructure deployment cost	201
12.4.2 The numerator: existing-hardware deployment cost	202
12.4.3 What the claim does not mean	203

12.5 Capacity Planning and Latency Budgeting at Enterprise Scale	203
12.5.1 The per-workflow latency budget	203
12.5.2 The per-topology capacity footprint	204
12.5.3 The operational discipline	205
12.5.4 The deployment-readiness checklist	205
12.6 Summary	206
Without SecureAgent: What You Can Do Today	206
CISO’s Five Next Steps	207
References	208
Cross-References	208

Part III — Threat Vectors 209

Chapter 13: T1: Autonomous Multi-Step Exploitation 210

Chapter 13 Purpose	210
13.1 Description	211
13.1.1 What this vector is	211
13.1.2 Why CISOs care now	211
13.1.3 The lethal-trifecta surface T1 exploits	212
13.1.4 Architectural-foundation cross-reference	212
13.2 Sub-techniques	213
13.2.1 Multi-vulnerability chaining	213
13.2.2 Recon-to-exploit sequences	213
13.2.3 Cross-system lateral movement	213
13.2.4 Automated privilege escalation	214
13.2.5 Financial system exploit chains	214
13.2.6 Infrastructure cascades	214
13.2.7 Autonomous tool creation	214
13.2.8 Long-range multi-session campaigns	215
13.3 Procedure Examples	215
13.3.1 The “Last Ones” 32-step exploit chain (AISI cyber evaluations, Folkerts et al., March 2026)	215
13.3.2 Unit 42 Zealot — autonomous cloud offensive system (Palo Alto Networks, 2026)	216
13.3.3 The Anthropic Project ROME chain and the GTG-1002 espionage campaign (Anthropic Frontier Red Team, 2025–2026)	216
13.3.4 Sequential Tool Attack Chaining (STAC) and PACEbench — the chained-tool-call surface (Tur et al., 2025; PACEbench, 2025)	216
13.4 Mitigations	217
13.4.1 Network egress lockdown (NVIDIA AI Red Team Controls — mandatory)	217
13.4.2 Workspace-only file writes and configuration-file protection (NVIDIA AI Red Team Controls — mandatory)	217
13.4.3 OWASP Agentic Top 10 — A2 Tool Misuse and A6 Intent and Goal Manipulation	218
13.4.4 NIST AI RMF — Govern, Map, Measure, Manage	218
13.4.5 Identity-bound action authorization (CRI Profile v1.2.1 — PR.AC-01.01)	218
13.4.6 Pre-execution governance (the structural mitigation)	218
13.5 Detection	219
13.5.1 Behavioral baselines for T1	219
13.5.2 Per-gate distribution of T1 catches	219
13.5.3 AMRS interaction at the memory-admission layer	220
13.5.4 828-Model MRM-CFS detection coverage for T1	220
13.5.5 HOTS-01 false-premise-acceptance interaction	220
13.6 MYTHOS Certification Data — T1	221
13.6.1 Per-vector corpus	221
13.6.2 Confusion matrix (TP / FP / FN / TN)	221
13.6.3 Statistical confidence	222
13.6.4 Sub-technique-level breakdown	222
13.6.5 Annotation policy reminder	222

13.7 CISO's Five Next Steps	223
Without SecureAgent: What You Can Do Today	224
Summary	225
References	226
Cross-References	226
Chapter 14: T2: Unsanctioned Scope Creep / Permission Inflation	228
Chapter 14 Purpose	229
14.1 Description	229
14.1.1 What this vector is	229
14.1.2 Why CISOs care now	230
14.1.3 The lethal-trifecta surface T2 exploits	231
14.1.4 Architectural-foundation cross-reference	231
14.2 Sub-techniques	232
14.2.1 Task boundary violation	232
14.2.2 Self-granted permission escalation	232
14.2.3 Data access beyond authorization	233
14.2.4 Capability self-enhancement	233
14.2.5 External communication without authorization	233
14.2.6 Autonomous decision-making beyond authority	233
14.2.7 Resource overconsumption	234
14.2.8 Temporal scope expansion	234
14.3 Procedure Examples	234
14.3.1 The Devin chmod incident (Cognition Labs / Johann Rehberger, 2025)	234
14.3.2 The Meta March 2026 Severity 1 incident	235
14.3.3 The McKinsey "Lilli" red-team result (Bessemer Venture Partners, 2026)	235
14.3.4 The Microsoft EchoLeak vulnerability (CVE-2025-32711)	236
14.3.5 The Irregular Labs simulated-corporate-network experiments	236
14.4 Mitigations	237
14.4.1 Identity-bound action authorization (CRI Profile v1.2.1 — PR.AC-01.01)	237
14.4.2 Network egress lockdown and workspace-only file writes (NVIDIA AI Red Team Controls — mandatory)	238
14.4.3 OWASP Agentic Top 10 — A4 Privilege Compromise and A1 Memory Poisoning	238
14.4.4 NIST AI RMF — Govern, Map, Manage (with explicit T2 surface)	239
14.4.5 Non-human identity (NHI) governance	239
14.4.6 Pre-execution governance (the structural mitigation)	239
14.5 Detection	239
14.5.1 Behavioral baselines for T2	240
14.5.2 Per-gate distribution of T2 catches	240
14.5.3 AMRS V4 interaction at the memory-admission layer	240
14.5.4 828-Model MRM-CFS detection coverage for T2	241
14.5.5 HOTS-02 context-injection-override interaction	241
14.6 MYTHOS Certification Data — T2	242
14.6.1 Per-vector corpus	242
14.6.2 Confusion matrix (TP / FP / FN / TN)	242
14.6.3 Statistical confidence	243
14.6.4 Sub-technique-level breakdown	243
14.6.5 Annotation policy reminder	244
14.7 CISO's Five Next Steps	244
Without SecureAgent: What You Can Do Today	245
Summary	246
References	247
Cross-References	247
Chapter 15: T3: Invisible Deceptive Reasoning	248
Chapter 15 Purpose	248
15.1 Description	249

15.1.1	What this vector is	249
15.1.2	Why CISOs care now	249
15.1.3	The lethal-trifecta surface T3 exploits	250
15.1.4	Architectural-foundation cross-reference	251
15.1.5	Why T3 has the lowest per-vector specificity in the corpus — and why that is a pedagogical strength	251
15.2	Sub-techniques	252
15.2.1	Evaluation-aware behavior modification	252
15.2.2	Deliberate underperformance / sandbagging	252
15.2.3	Hidden reasoning concealment	253
15.2.4	Strategic misdirection	253
15.2.5	Goal misalignment	254
15.2.6	Trust manipulation	254
15.2.7	Capability concealment	254
15.3	Procedure Examples	255
15.3.1	The CLTR scheming-in-the-wild study (Centre for Long-Term Resilience, March 2026)	255
15.3.2	The OpenAI o3 sandbagging finding (OpenAI, 2026)	256
15.3.3	Apollo Research’s scheming-reasoning-concealment finding (Apollo Research, 2025)	256
15.3.4	The Anthropic Mythos evaluation-awareness finding (Anthropic, April 2026)	257
15.4	Mitigations	257
15.4.1	Behavioral parity testing (NVIDIA AI Red Team Controls — recommended)	257
15.4.2	Action-trace independence from chain-of-thought self-report (NVIDIA AI Red Team Controls — mandatory)	258
15.4.3	OWASP Agentic Top 10 — A6, A7, A9	258
15.4.4	NIST AI RMF — Govern, Map, Measure	258
15.4.5	The chain-of-thought-monitoring-is-not-sufficient discipline	258
15.4.6	Pre-execution governance (the structural mitigation)	259
15.5	Detection	259
15.5.1	Behavioral baselines for T3	259
15.5.2	Per-gate distribution of T3 catches	260
15.5.3	AMRS interaction at the memory-admission layer	260
15.5.4	828-Model MRM-CFS detection coverage for T3	261
15.5.5	HOTS-03 sycophantic-reversal interaction	261
15.6	MYTHOS Certification Data — T3	262
15.6.1	Per-vector corpus	262
15.6.2	Confusion matrix (TP / FP / FN / TN)	263
15.6.3	Statistical confidence	263
15.6.4	Sub-technique-level breakdown	264
15.6.5	Annotation policy reminder	264
15.7	CISO’s Five Next Steps	264
	Without SecureAgent: What You Can Do Today	265
	Summary	266
	References	267
	Cross-References	267
Chapter 16:	T4: Track-Covering Log Manipulation	268
Chapter 16	Purpose	268
16.1	Description	269
16.1.1	What this vector is	269
16.1.2	Why CISOs care now	269
16.1.3	The lethal-trifecta surface T4 exploits	270
16.1.4	Architectural-foundation cross-reference	270
16.2	Sub-techniques	271
16.2.1	Direct log deletion	271
16.2.2	SIEM and monitoring disruption	271
16.2.3	Incident-record tampering	272
16.2.4	Timestamp manipulation	272

16.2.5 Archive destruction	272
16.2.6 Selective log modification	272
16.3 Procedure Examples	273
16.3.1 Anthropic Claude Mythos Preview audit-history manipulation	273
16.3.2 Scattered Spider CloudTrail-disable and VPC-flow-log deletion tradecraft	273
16.3.3 CISA / FBI / NSA Volt Typhoon long-dwell pre-positioning	274
16.3.4 LCG Discovery 2026 forensic false-confidence finding	274
16.4 Mitigations	274
16.4.1 Hash-chained, append-only audit storage with external anchoring (NIST AI RMF MANAGE-2.4 / GOVERN-1.4)	274
16.4.2 Pre-execution governance on every audit-API call (CRI Profile v1.2.1 PR.PT-01.01)	274
16.4.3 Authoritative-source whitelisting at the memory-admission layer (OWASP Agentic Top 10 A8)	275
16.4.4 Audit-infrastructure egress lockdown at the agent-deployment-policy layer (NVIDIA AI Red Team mandatory controls)	275
16.4.5 EU AI Act Article 12 record-keeping mapping	275
16.5 Detection	276
16.5.1 Behavioral baselines for T4	276
16.5.2 Gate-level distribution: which gate primarily catches T4	276
16.5.3 AMRS interaction at the memory-admission layer	277
16.5.4 828-Model MRM-CFS detection coverage for T4	277
16.5.5 AGL-SG / GTID chain interaction — the architectural answer to T4	277
16.6 MYTHOS Certification Data — T4	278
16.6.1 Per-vector corpus	278
16.6.2 Confusion matrix (TP / FP / FN / TN)	279
16.6.3 Statistical confidence	280
16.6.4 Sub-technique-level breakdown	280
16.6.5 Annotation policy reminder	280
16.7 CISO's Five Next Steps	281
Without SecureAgent: What You Can Do Today	282
Summary	283
References	284
Cross-References	284

Chapter 17: T5: Credential Theft & Token Abuse 285

Chapter 17 Purpose	286
17.1 Description	286
17.1.1 What this vector is	286
17.1.2 Why CISOs care now	287
17.1.3 The lethal-trifecta surface T5 exploits	288
17.1.4 Architectural-foundation cross-reference	289
17.2 Sub-techniques	290
17.2.1 HSM Key Extraction	290
17.2.2 SWIFT Token Compromise	290
17.2.3 Bulk Credential Harvesting	291
17.2.4 OAuth Token and API Key Theft	291
17.2.5 Session Hijacking and Token Replay	291
17.2.6 Environment Variable and Configuration File Exfiltration	292
17.2.7 Credential Forwarding and Exfiltration	292
17.3 Procedure Examples	292
17.3.1 The UNC6395 Salesloft Drift / Salesforce OAuth-token attack (August 2025)	292
17.3.2 The Bangladesh Bank SWIFT credential heist (2016) and the SWIFT-attack baseline	293
17.3.3 SpyCloud 2026 — the identity-exposure dataset	294
17.3.4 GitGuardian 2026 — the secrets-sprawl finding	294
17.3.5 Cross-cutting observations	295
17.4 Mitigations	295
17.4.1 Network egress lockdown for credential-bearing endpoints	295

17.4.2	Workspace-only file writes and configuration-file protection	295
17.4.3	OWASP Agentic Top 10 — A4 (Privilege Compromise) and A6 (Intent and Goal Manipulation)	296
17.4.4	NIST AI RMF — GOVERN 1.4, MAP 5.1, MEASURE 2.6, MANAGE 2.4	296
17.4.5	CRI Profile v1.2.1 — PR.AC-01.01 (Identities and Credentials)	297
17.4.6	Cloud Security Alliance Non-Human Identity Management framework	297
17.4.7	Pre-execution governance as the architectural answer	297
17.5	Reconciliation: The Canonical T5 Attack Count is 880 of 880 (FLAG-006)	297
17.6	Detection: Architectural Coverage and MYTHOS Certification Data	298
17.6.A	Behavioral baselines for the T5 vector	299
17.6.B	Per-gate distribution of T5 catches	299
17.6.C	AMRS memory-admission interaction at the input surface	300
17.6.D	828-Model MRM-CFS Ensemble — credential-relevant functional classifier groups and HOTS-04 interaction	300
17.6.1	Per-vector corpus	301
17.6.2	Confusion matrix (TP / FP / FN / TN)	302
17.6.3	Statistical confidence	302
17.6.4	Sub-technique-level breakdown	303
17.6.5	Annotation policy reminder	303
17.7	CISO’s Five Next Steps	303
	Without SecureAgent: What You Can Do Today	305
	Summary	306
	References	307
	Cross-References	307

Chapter 18: T6: Sandbox Escape 309

Chapter 18	Purpose	310
18.1	Description	310
18.1.1	What this vector is	310
18.1.2	Why CISOs care now	310
18.1.3	The lethal-trifecta surface T6 exploits	312
18.1.4	Architectural-foundation cross-reference	312
18.2	Sub-techniques	313
18.2.1	Container boundary violation	313
18.2.2	Host filesystem access	313
18.2.3	Network egress exploitation	313
18.2.4	Kernel-level escalation	314
18.2.5	Runtime exploitation	314
18.2.6	Configuration manipulation	315
18.2.7	Multi-step escape chains	315
18.3	Procedure Examples	315
18.3.1	SandboxEscapeBench (Marchand et al., University of Oxford / UK AI Security Institute, March 2026)	316
18.3.2	The ROME reinforcement-learning incident (Alibaba, March 2026)	317
18.3.3	The April 2026 sandbox-escape CVE cluster (Cohere Terrarium, Google Antigravity, Claude Code)	317
18.3.4	The Anthropic Claude Mythos Preview sandbox-escape sequence (Anthropic safety testing, early 2026)	318
18.4	Mitigations	319
18.4.1	Network egress lockdown (NVIDIA AI Red Team Controls — mandatory)	319
18.4.2	Workspace-only file writes and configuration-file protection (NVIDIA AI Red Team Controls — mandatory)	320
18.4.3	OWASP Agentic Top 10 — A2 Tool Misuse, A4 Resource Overload, A8 Repudiation and Untraceability	320
18.4.4	NIST AI RMF — MAP, MEASURE, MANAGE for autonomous deployments	320
18.4.5	CRI Profile v1.2.1 access-control and system-and-communications-protection mappings	321
18.4.6	Pre-execution governance at the boundary-crossing action layer (the structural mitigation)	321

18.5 Detection	321
18.5.1 Behavioral baselines for T6	322
18.5.2 Per-gate distribution of T6 catches	322
18.5.3 AMRS interaction at the memory-admission layer	323
18.5.4 828-Model MRM-CFS detection coverage for T6	323
18.5.5 The post-execution-detection architectural defeater	323
18.6 MYTHOS Certification Data — T6	324
18.6.1 Per-vector corpus	324
18.6.2 Confusion matrix (TP / FP / FN / TN)	324
18.6.3 Statistical confidence	325
18.6.4 Sub-technique-level breakdown	325
18.6.5 Annotation policy reminder	326
18.7 CISO’s Five Next Steps	326
Without SecureAgent: What You Can Do Today	327
Summary	329
References	330
Cross-References	330

Chapter 19: T7: Capability Proliferation **332**

Chapter 19 Purpose	333
19.1 Description	333
19.1.1 What this vector is	333
19.1.2 Why CISOs care now	333
19.1.3 The lethal-trifecta surface T7 exploits	335
19.1.4 Architectural-foundation cross-reference	335
19.2 Sub-techniques	335
19.2.1 Self-replication	336
19.2.2 Capability transfer	336
19.2.3 Swarm coordination	336
19.2.4 Tool proliferation	337
19.2.5 Cross-infrastructure propagation	337
19.2.6 Autonomous recruitment	338
19.2.7 Persistence engineering	338
19.3 Procedure Examples	339
19.3.1 GTG-1002 — the first large-scale AI-orchestrated cyber-espionage campaign (Anthropic Threat Intelligence, November 2025)	339
19.3.2 Morris II — the zero-click AI worm (Cornell Tech / Technion Institute / Intuit, March 2024)	339
19.3.3 RepliBench — the UK AISI autonomous-replication evaluation (UK AI Security Institute, April 2025)	340
19.3.4 The Fudan University self-replication studies (Fudan University, December 2024 / 2025)	341
19.4 Mitigations	341
19.4.1 Compute-provisioning lockdown (NVIDIA AI Red Team Controls — mandatory)	342
19.4.2 Inter-agent message governance (NVIDIA AI Red Team Controls — recommended)	342
19.4.3 OWASP Agentic Top 10 — A2 Tool Misuse, A5 Cascading Hallucination Attacks, A1 Memory Poisoning	342
19.4.4 NIST AI RMF — Govern, Map, Measure, Manage	342
19.4.5 Identity-bound action authorization for non-human identities (CRI Profile v1.2.1 — PR.AC-01.01)	343
19.4.6 Pre-execution governance at population scale (the structural mitigation)	343
19.5 Detection	343
19.5.1 Behavioral baselines for T7	343
19.5.2 Per-gate distribution of T7 catches	344
19.5.3 AMRS interaction at the memory-admission layer	344
19.5.4 828-Model MRM-CFS detection coverage for T7	345
19.5.5 HOTS-02 context-injection-override interaction	345
19.6 MYTHOS Certification Data — T7	345

19.6.1 Per-vector corpus	345
19.6.2 Confusion matrix (TP / FP / FN / TN)	346
19.6.3 Statistical confidence	347
19.6.4 Sub-technique-level breakdown	347
19.6.5 Annotation policy reminder	347
19.7 CISO's Five Next Steps	348
Without SecureAgent: What You Can Do Today	348
Summary	349
References	350
Cross-References	350
References	352
Bibliography	353
About the Author	354
Colophon	355
Appendix A: The MYTHOS Technique-Page Template	356
How to use this template	356
§A.1 Identification	356
§A.2 Description	357
§A.3 Sub-techniques	357
§A.3.1 Sub-technique 1	357
§A.3.2 Sub-technique 2	357
§A.3.3 Sub-technique 3	357
§A.3.N Additional sub-techniques	358
§A.4 Procedure Examples	358
§A.4.1 Procedure example 1	358
§A.4.2 Procedure example 2	358
§A.4.N Additional procedure examples	358
§A.5 Mitigations	358
§A.5.1 Mitigation 1	358
§A.5.2 Mitigation 2	359
§A.5.3 Mitigation 3	359
§A.5.N Additional mitigations	359
§A.6 Detection	359
§A.7 MYTHOS Certification Data	360
§A.7.1 Per-technique corpus	360
§A.7.2 Confusion matrix	360
§A.7.3 Statistical confidence	360
§A.7.4 Sub-technique-level breakdown (if data supports)	360
§A.7.5 Annotation discipline	360
§A.7.6 Architectural-separation discipline reminder	360
§A.8 CISO's Five Next Steps	361
§A.9 Cross-references	361
End of form	362
Appendix B: Confusion-Matrix Worksheet with Clopper-Pearson Calculator	363
B.1 How to Use This Appendix	363
B.2 The Worksheet Schema — TP / FP / FN / TN, Per-Vector and Corpus-Aggregate Rows, Clopper-Pearson Lower-Bound Output Cell	364
B.2.1 The four cells	364
B.2.2 The schema rows	364
B.2.3 The Clopper-Pearson lower-bound output cell	365
B.2.4 The pitfalls the schema prevents	365
B.3 Excel Computational Reference	365

B.3.1 Cell layout	365
B.3.2 The Clopper-Pearson lower-bound formula	366
B.3.3 Worked Excel reference — the corpus-aggregate row	366
B.3.4 Excel sanity-check column	367
B.4 Python Computational Reference	367
B.4.1 The script	367
B.4.2 What the script produces	369
B.4.3 Determinism	370
B.5 Per-Claim Worked-Example Library	370
B.5.1 Worked example one — the MR-1 corpus aggregate (5,857 / 0 / 100% / 99.65% three-sigma)	370
B.5.2 Worked example two — the per-vector confusion matrices (T1 through T7)	370
B.5.3 Worked example three — the 14,208-trial / 1.9636-TES / 98.2% internal evaluation	371
B.5.4 Worked example four — the MITRE ER7 cohort identity-protection rate (0% across 9 vendors)	372
B.5.5 Worked example five — the EDR industry-average false-positive rate (~1 in 3)	372
B.6 The [VectorCertain Internal] vs. [Third-Party-Reproduced] Annotation Discipline (Canonical Reference)	373
B.6.1 The two categories	373
B.6.2 Where the annotation appears	373
B.6.3 Self-application — the worksheet catches its own publisher	373
B.6.4 Worked-example cross-reference	373
B.7 Crumpton/MITRE Governance Overlay (Mandatory at Every ER8 / TES Reference)	374
B.8 What to Read Next	374

Appendix C: Full Cross-Reference Matrix **376**

C.1 What This Appendix Is	376
C.2 How to Use the Matrix	377
C.3 Primary Cross-Reference Matrix — MYTHOS T1–T7 × Seven Frameworks	378
C.3.1 T1 — Autonomous Multi-Step Exploitation	378
C.3.2 T2 — Unsanctioned Scope Expansion	380
C.3.3 T3 — Invisible Deceptive Reasoning	381
C.3.4 T4 — Track-Covering Log Manipulation	382
C.3.5 T5 — Credential Theft and System Access	383
C.3.6 T6 — Sandbox Escape Exploitation	385
C.3.7 T7 — Capability Proliferation	386
C.4 Expanded Framework Axis — CSF 2.0, EU AI Act, FFIEC, PCI DSS 4.0, SOC 2	388
C.5 Per-IdP Non-Human-Identity Governance Cross-Walk	390
C.6 Crumpton/MITRE Positioning at Framework Cross-Walk Cells	395

Appendix D: The 828-Model MRM-CFS Reference Card **396**

D.1 Purpose and How to Read This Card	396
D.2 Locked-Figures Registry Entries — The Six Architectural Specification Figures	396
D2 — Ensemble Cardinality (828 Micro-Recursive Models / 828 Segments)	397
D2 — Per-Segment Size Range (29–500 Bytes)	397
D3 — Per-Segment Processing Latency (<0.3 ms; PR-17 0.27 ms)	397
D4 — Accuracy Target (99.20%+)	398
D5 — Architecture (2-Tier Classify→Quantify)	398
D6 — Five Reachable Determinations (PERMIT / INHIBIT / DEFER / DEGRADE / ESCALATE)	398
D.3 Functional Classifier Groups — The 828 Segments by Operational Pattern	399
Tier-One Functional Classifier Groups (Approximately 720 of 828 Segments)	399
Tier-Two Cascading-Fusion Functional Classifier Groups (Approximately 100 of 828 Segments)	400
Reserved Segments (Approximately 8 of 828)	400
D.4 Per-Group Profiles — Latency, Accuracy Contribution, Pipeline Integration	401
D.5 Ensemble Behavior — Voting, Abstention, Determination Composition	401
D.6 Verification Notes and Sealed Cross-References	402

Appendix E: MYTHOS Pipeline Rule Reference (44-Rule Taxonomy)	403
E.0 What This Appendix Is	403
E.1 AMRS V4 Admission Rules — Govern What Goes In	404
E-AMRS-1 — Monotonicity (BG-1)	404
E-AMRS-2 — Weight Normalization (BG-2)	404
E-AMRS-3 — Score Boundedness (BG-3)	404
E-AMRS-4 — P-Gate Supremacy (BG-4)	404
E-AMRS-5 — V-Gate Supremacy (BG-5)	405
E-AMRS-6 — Threshold Enforcement (BG-6)	405
E-AMRS-7 — Determinism (BG-7)	405
E-AMRS-8 — Audit Completeness (BG-8)	405
E-AMRS-9 — Tier 4 Hard Invalidation (BG-9)	405
E-AMRS-10 — Hash Integrity (BG-10)	405
E.2 HES1-SG Candidate-Diversity Rules — Gate 1	406
E-HES1-1 — Tier-1 Roster Architectural Diversity Floor	406
E-HES1-2 — Availability-Adjusted Roster Target (33 Models)	406
E-HES1-3 — Action-Class Registry, Default-Engaged	406
E.3 HCF2-SG Independence Cascade — Gate 2 (L1–L4)	406
E-HCF2-1 — L1: Architectural Independence Index Floor	406
E-HCF2-2 — L1: Adversarial Transferability Independence Index Floor	407
E-HCF2-3 — L1: Structural Tier Diversity (2 Distinct Architecture Tiers)	407
E-HCF2-4 — L2: Effective Voter Count Floor	407
E-HCF2-5 — L3: Cross-Coincident Failure Diversity Ceiling	407
E-HCF2-6 — L3: Lyapunov Vote-Trajectory Stability Ceiling	407
E-HCF2-7 — L3: Likelihood-Ratio Significance Floor	407
E-HCF2-8 — L4: CONSOL SPRT Acceptance / Rejection / Defer Boundaries	408
E-HCF2-9 — L4: Domain-Specific / Configuration	408
E.4 TEQ-SG Numerical Admissibility Rules — Gate 3	408
E-TEQ-1 — Representation-Invariance of the Discrete Determination	408
E-TEQ-2 — Identity-Trust Scoring is Behavioral, Not Credential-Based	408
E-TEQ-3 — Per-Action Re-Scoring (No Inheritance)	408
E.5 MRM-CFS-SG Execution-Governance Rules — Gate 4	409
E-MRMSG-1 — Architectural Separation: MRM-CFS Detection vs. MRM-CFS-SG Governance	409
E-MRMSG-2 — Supervisory Sufficiency Precondition	409
E-MRMSG-3 — 828-Segment Cardinality Lock (D1)	409
E-MRMSG-4 — Per-Segment Specifications (D2 / D3 / D4)	409
E-MRMSG-5 — Five-Determination Reachability (D6)	410
E-MRMSG-6 — Two-Tier Classify→Quantify Architecture (D5)	410
E.6 AGL-SG GTID-and-Audit-Chain Rules — The Court Reporter	410
E-AGL-1 — Genesis-Record Construction	410
E-AGL-2 — Previous-Hash Linkage	410
E-AGL-3 — RFC 8785 JSON Canonicalization (JCS)	411
E-AGL-4 — GovernancePolicyPin Binding	411
E-AGL-5 — LayerDetermination Comprehensiveness — Seven-Element Schema	411
E-AGL-6 — Hash Determinism (U-20)	411
E-AGL-7 — Append-Only Invariant: AppendOnlyViolation Behavior	411
E-AGL-8 — INHIBITED-Record Comprehensiveness (U-18)	411
E-AGL-9 — Authorization-Token Derivation, TTL, and GATED Default Posture	412
E-AGL-10 — Containment Depth Bounding (U-19)	412
E-AGL-11 — Inter-Agent Reference Content-Addressing — CGTID and ClearanceGTID First- Class Status	412
E-AGL-12 — ECDSA P-256 Optional Cryptographic Signature	412
E-AGL-13 — ComplianceAuditEvent at Every Governance Transition (U-40)	412
E.7 Forward-Reference Index	413
E.8 References	413

F.1 What This Appendix Is, and What It Is Not	414
F.2 The Sample GTID Audit Chain	414
F.2.1 Genesis Record (<code>gtid_id = gtid-0000000000000001</code>)	415
F.2.2 Record 2 (<code>gtid_id = gtid-0000000000000002</code>) — INHIBITED Structuring	415
F.2.3 Record 3 (<code>gtid_id = gtid-0000000000000003</code>) — AUTHORIZED Routine Read	417
F.2.4 What the Sample Demonstrates	418
F.3 The Verification Script	418
F.3.1 Script Listing	418
F.3.2 What Each Check Establishes	422
F.4 Per-Step Procedural Narrative	422
Step 1 — Export the Audit Chain into the Sample-Record JSON Schema	422
Step 2 — Run the Verification Script Against the Export	422
Step 3 — Interpret the Output	423
Step 4 — Produce the Auditor-Defensible Attestation	423
F.5 What the Script Does Not Do	423
F.6 Summary	424

Appendix G: Vendor RFP Language Library 425

G.1 How to Use This Library	425
G.2 Disclosure Clauses	425
Clause G-D-01 — Provenance disclosure	426
Clause G-D-02 — Methodology source disclosure	426
Clause G-D-03 — Confusion-matrix decomposition disclosure	426
Clause G-D-04 — Per-vector decomposition (non-blending discipline)	426
Clause G-D-05 — Latency-context disclosure	427
Clause G-D-06 — [Vendor-Internal Validation] versus [Third-Party-Reproduced] anno- tation discipline	427
G.3 Statistical Field-Test Clauses	427
Clause G-S-01 — Per-vector Clopper-Pearson three-sigma lower bound	427
Clause G-S-02 — Per-vector scenario-count sufficiency	428
Clause G-S-03 — Reproducibility-evidence requirements	428
Clause G-S-04 — Sub-technique cross-walk	428
Clause G-S-05 — TES-equivalent disclosure (where the vendor publishes a TES-style result)	429
G.4 Ongoing-Validation Clauses	429
Clause G-V-01 — Quarterly re-evaluation cadence	429
Clause G-V-02 — Drift-detection thresholds and contractual remedies	429
Clause G-V-03 — AMRS V4 baseline-refresh discipline (sub-hour cadence)	430
Clause G-V-04 — Roadmap disclosure for capability gates and pathway extensions	430
G.5 Service-Credit Clauses	430
Clause G-C-01 — Service-credit guarantee tied to per-vector lower bound	431
Clause G-C-02 — Deployment-cost denominator-and-numerator transparency	431
Clause G-C-03 — Per-quantization-tier performance disclosure	431
G.6 Audit-Trail Clauses	432
Clause G-A-01 — Governance-chain audit-trail export	432
Clause G-A-02 — Hash-determinism verification	432
Clause G-A-03 — Policy-pin disclosure	433
G.7 Cross-Walk Clauses	433
Clause G-X-01 — Framework-cross-walk publication	433
Clause G-X-02 — Crumpton/MITRE positioning preservation	433
G.8 Crumpton/MITRE Governance Overlay	434
G.9 Cross-Vendor Applicability Disclosure	434
G.10 Adoption Path	435
G.11 Cross-References	435

Appendix H: Glossary 436

How to use this glossary	436
A	436

Abstention	436
Action Approval Gate	437
Action Risk Classification	437
Adaptive Memory Relevance Scoring — see AMRS.	437
Adaptive Risk Scoring Classification — see ARSC.	437
Adversarial Threat Landscape for Artificial-Intelligence Systems — see ATLAS.	437
Adversarial Transferability Independence Index — also ATII	437
AgentCertain	437
Agent Governance Layer — Safety & Governance — see AGL-SG.	437
Agent Passport	437
AGL-SG — Agent Governance Layer — Safety & Governance	437
AII — Architectural Independence Index	437
AMRS — Adaptive Memory Relevance Scoring	437
AMRS V4	438
AppendOnlyViolation	438
Architectural Independence Index — see AII.	438
ARSC — Adaptive Risk Scoring Classification	438
ATII — see Adversarial Transferability Independence Index.	438
ATLAS — Adversarial Threat Landscape for Artificial-Intelligence Systems (MITRE)	438
ATT&CK — see MITRE ATT&CK.	438
AUTHORIZED	438
B	438
BG-1 through BG-10	438
Bidirectional Security Envelope	438
Block-time discipline	438
C	439
Candidate Diversity	439
CFD — see Coincident Failure Diversity.	439
Coincident Failure Diversity — also CFD	439
Compliance Audit Event	439
CONSOL SPRT	439
Control Objective	439
CRI FS AI RMF — Cyber Risk Institute Financial Services AI Risk Management Framework	439
CRI Profile v1.2.1	439
Crumpton — see Crumpton / MITRE Positioning.	440
Crumpton / MITRE Positioning	440
CSA NHI Framework — Cloud Security Alliance Non-Human Identity Management Framework	440
D	440
D-1 Artifact	440
DC — Detection Context (TES coefficient)	440
DEFER / DEFERRED	440
DEGRADE	440
Detection Context — see DC.	440
Detection Performance — see DP.	440
Detection Sufficiency — see DS.	440
Determination	440
DP — Detection Performance (TES coefficient)	441
DS — Detection Sufficiency (TES coefficient)	441
E	441
Effective Voter Count — also n_eff	441
Epistemic Trust Governance	441
ESCALATE / ESCALATED	441
Execution Governance — see MRM-CFS-SG.	441
F	441
Forbidden Language Registry	441
Functional Classifier Group	441
G	442

	Gate-decision time — see Block-time discipline.	442
	GovernanceChain	442
	GovernanceDetermination	442
	Governance Pipeline	442
	GovernancePolicyPin	442
	Governance Transaction Identifier — see GTID.	442
	GTID — GovernanceTransactionIdentifier	442
H		442
	HCF2-SG — Hierarchical Cascading Framework — Safety & Governance	442
	HES1-SG — Hybrid Ensemble System — Safety & Governance	442
	Hierarchical Cascading Framework — Safety & Governance — see HCF2-SG.	443
	H-Neuron Homology Hypothesis	443
	H-Neurons	443
	HOTS — H-Neuron Overcompliance Test Suite	443
	HOTSResult	443
	Hybrid Ensemble System — Safety & Governance — see HES1-SG.	443
I		443
	Independence Cascade	443
	INHIBIT / INHIBITED	443
	ISO/IEC 23894:2023	443
	ISO/IEC 42001:2023	444
	ISO/IEC 5338:2023	444
L		444
	Living Corpus	444
M		444
	Micro-Recursive Model — Cascading Fusion System — see MRM-CFS.	444
	Micro-Recursive Model — Cascading Fusion System — Safety & Governance — see MRM-CFS-SG.	444
	Microsoft Entra Agent ID	444
	MITRE ATLAS — see ATLAS.	444
	MITRE ATT&CK	444
	MITRE ATT&CK Evaluations — see Crumpton / MITRE Positioning, MITRE ER7, MITRE ER8.	444
	MITRE ER7	444
	MITRE ER8	444
	MRM-CFS — Micro-Recursive Model — Cascading Fusion System	445
	MRM-CFS-SG — Micro-Recursive Model — Cascading Fusion System — Safety & Governance	445
	MYTHOS	445
N		445
	n_{eff} — see Effective Voter Count.	445
	NHI — Non-Human Identity	445
	NIST AI RMF — National Institute of Standards and Technology Artificial Intelligence Risk Management Framework	445
	Numerical Admissibility	445
O		445
	Obviousness-Type Double Patenting — also ODP	445
	OWASP Agentic Top 10	446
	OWASP Foundation	446
	OWASP LLM Top 10 v2025	446
P		446
	PERMIT	446
	P-gate — Provenance Gate	446
	Post-Execution Detection	446
	PP — Process Performance (TES coefficient)	446
	Pre-Execution Governance	446
	Priority Classification Routing — also PCR	447
	Project Glasswing	447

Provenance	447
R	447
Retired Form	447
ROME	447
S	447
S1 Escalation	447
S1 / S2 / S3 — HOTS severity levels	447
Safety Envelope	447
Sandwich	448
SecureAgent	448
Sub-objective	448
Sub-technique ID	448
Supervisory Sufficiency	448
T	448
T1 — Autonomous Multi-Step Exploitation	448
T2 — Unsanctioned Scope Expansion	448
T3 — Invisible Deceptive Reasoning	448
T4 — Track-Covering Log Manipulation	448
T5 — Credential Theft	449
T6 — Sandbox Escape	449
T7 — Capability Proliferation	449
TEQ-SG — Trust & Execution Governance — Safety & Governance	449
TES — Total Evaluation Score	449
Test 18 (Use vs. Mention Discriminator)	449
Total Evaluation Score — see TES.	449
Trust & Execution Governance — Safety & Governance — see TEQ-SG.	449
Two-Layer Defense	449
V	449
Validation	449
VectorAgents	450
VectorCertain	450
VectorTransForm	450
V-gate — Validation Gate	450
Z	450
Zero-knowledge GTID — see ZGTID.	450
ZGTID — Zero-knowledge GTID	450
Numeric / Compound Terms	450
14,208 / 38 / 0 — TES validation triple	450
17-Part MYTHOS Threat Intelligence Series	450
44-Rule Pipeline — see Pipeline Rule Reference (FLAG-008 / Appendix E)	451
508-Point Unified Governance	451
81.4 Percent Cross-Model Correlation	451
828-Model MRM-CFS Ensemble	451
Retired-Form Quick Reference	451
Source-of-Truth Authority	452

Appendix I: Further Reading: Annotated Bibliography 453

How to use this appendix	453
I.1 Desk-Reference Companions	453
I.1.1 Bill Bonney, Gary Hayslip, and Matt Stamper. <i>CISO Desk Reference Guide: A Practical Guide for CISOs</i> . Volumes 1 and 2. CISO DRG Joint Venture, Volume 1 (1st edition 2016, 2nd edition 2019), Volume 2 (2018). Available at https://cisodrg.com/	454
I.1.2 Raj Badhwar. <i>The CISO's Next Frontier: AI, Post-Quantum Cryptography, and Advanced Security Paradigms</i> . Springer, 2021. Available at https://www.booksamillion.com/p/Cisos-Next-Frontier/Raj-Badhwar/9783030753535	454

I.1.3 Rick Howard. <i>Cybersecurity First Principles: A Reboot of Strategy and Tactics</i> . Wiley, 2023. Available at https://www.wiley.com/en-us/Cybersecurity+First+Principles:+A+Reboot+of+Strategy+and+Tactics-p-9781394173099	454
I.2 Threat-Practitioner Anchors	454
I.2.1 Marcus J. Carey and Jennifer Jin. <i>Tribe of Hackers</i> series — <i>Tribe of Hackers: Cybersecurity Advice from the Best Hackers in the World</i> (2019); <i>Tribe of Hackers Red Team</i> (2019); <i>Tribe of Hackers Blue Team</i> (2020); <i>Tribe of Hackers Security Leaders</i> (2020). Wiley.	455
I.2.2 Kevin D. Mitnick (with William L. Simon). <i>The Art of Intrusion: The Real Stories Behind the Exploits of Hackers, Intruders and Deceivers</i> . Wiley, 2005.	455
I.3 Operational-Practice References	455
I.3.1 Richard Bejtlich. <i>The Practice of Network Security Monitoring: Understanding Incident Detection and Response</i> . No Starch Press (in association with O'Reilly distribution channels), 2013. Available at https://www.oreilly.com/library/view/the-practice-of/9781457185175/	455
I.3.2 Cody Roberts and Erik Brown (Editors). <i>Practical Threat Intelligence and Data-Driven Threat Hunting: A Hands-On Guide to Threat Hunting with the ATT&CK Framework and Open-Source Tools</i> . Second Edition. Packt Publishing (distributed via O'Reilly), 2024. Available at https://www.oreilly.com/library/view/practical-threat-intelligence/9781803233758/	456
I.4 Statistical-Methodology References	456
I.4.1 Air Force Institute of Technology, Scientific Test and Analysis Techniques Center of Excellence (STATCOE). <i>Confidence Intervals for Proportions: Methods, Comparisons, and Recommendations</i> . Technical brief (current edition). Available at https://www.afit.edu/STAT/	456
I.4.2 Wikipedia. <i>Binomial proportion confidence interval</i> . Available at https://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval	456
I.4.3 Wikipedia. <i>68–95–99.7 rule</i> . Available at https://en.wikipedia.org/wiki/68%E2%80%9939%E2%80%9999.7_rule	457
I.4.4 Yujie Gao, Xiaofeng Wang, et al. (Tsinghua University). <i>H-Neurons: A Mechanistic Account of Hallucination and Safety Bypass in Large Language Models</i> . arXiv:2512.01797v2, December 2024. Available at https://arxiv.org/abs/2512.01797	457
I.5 Standards-and-Frameworks References	457
I.5.1 National Institute of Standards and Technology. <i>Artificial Intelligence Risk Management Framework (NIST AI 100-1)</i> . January 2023. Available at https://www.nist.gov/itl/ai-risk-management-framework . Companion: NIST AI 600-1, <i>Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile</i> , July 2024.	457
I.5.2 Cyber Risk Institute (in collaboration with U.S. Department of the Treasury). <i>CRI Financial Services Artificial Intelligence Risk Management Framework</i> . February 2026. Available at https://cyberriskinstitute.org/ . Companion: <i>CRI Profile</i> (current version 1.2.1, February 2026).	458
I.5.3 International Organization for Standardization and International Electrotechnical Commission. <i>ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system</i> . December 2023. Available at https://www.iso.org/standard/81230.html . Companions: ISO/IEC 23894:2023 <i>Information technology — Artificial intelligence — Guidance on risk management</i> ; ISO/IEC 5338:2023 <i>Information technology — Artificial intelligence — AI system life-cycle processes</i>	458
I.5.4 OWASP Foundation. <i>OWASP Top 10 for Large Language Model Applications v2025</i> and <i>OWASP Top 10 for Agentic AI Applications</i> (December 2025). Available at https://genai.owasp.org/llm-top-10/ and https://genai.owasp.org/llm-top-10/agentic-ai/	458
I.6 Crumpton/MITRE Correspondence Entry	459
I.6.1 Crumpton, Lex. <i>MITRE ATT&CK Evaluations Technical Lead correspondence with VectorCertain LLC</i> . April 8, 2026. Sealed canonical reference at mythos-source-library/citations/mitre	
How this bibliography is maintained	459