Contents

Preface xvii
Hemant Kumar SAINI, Mariya OUAISSA, Yassine MALEH,
Mariyam OUAISSA and Rejwan BIN SULAIMAN

About the Editors xxi
Hemant Kumar SAINI, Mariya OUAISSA, Yassine MALEH,
Mariyam OUAISSA and Rejwan BIN SULAIMAN

Chapter 1. Privacy-preserving ML and Secure AI Architectures 1
Hicham ZMAIMITA, Abdellah MADANI and Driss ABADA

1.1. Introduction 1
1.2. Threat model and leakage taxonomy 3
1.2.1. Assets to protect. 3
1.2.2. Leakage channels 4
1.2.3. Adversaries and scenarios 5
1.2.4. Assumptions. 6
1.3. Privacy-by-design requirements for secure GenAI architectures 6
1.3.1. Data minimization 6
1.3.2. Separation of environments and responsibilities 6
1.3.3. Cryptographic confidentiality 7
1.3.4. Access control and auditability 7
1.3.5. Resilience: detection, response, recovery. 8
1.3.6. Utility constraints (security without breaking user workflows). 8
1.4. Reference architecture. 8
1.4.1. High-level logical view. 9
1.4.2. End-to-end “upload-to-response” data flow 11
1.4.3. Control points and attack surfaces. 12

1.5. Securing file ingestion 15
1.5.1. File validation and hygiene 15
1.5.2. Sandboxing and isolation 16
1.5.3. Sensitivity classification 17
1.5.4. Policy decisioning (policy engine) 18
1.6. Architecture-wide data protection (condensed) 18
1.6.1. Encryption and key management 18
1.6.2. De-identification with TTL and controlled reversibility 19
1.6.3. Secrets and telemetry hygiene 19
1.6.4. Optional: confidential computing 19
1.7. Secure RAG and connector safety 20
1.7.1. Secure RAG design: separation and retrieval-time authorization 20
1.7.2. Prompt-injection defenses for retrieved/uploaded content 20
1.7.3. Retention and life cycle 21
1.8. Runtime controls and operations 21
1.8.1. Output leakage prevention 21
1.8.2. Observability and audit 21
1.8.3. Monitoring and incident response 21
1.9. Evaluation methodology and key metrics 22
1.9.1. Evaluation protocol 22
1.9.2. Metrics 22
1.10. Deployment guidance and checklist 23
1.10.1. Deployment blueprint 23
1.10.2. Practical checklist and common pitfalls 23
1.11. Conclusion and outlook 23
1.12. References 25

Chapter 2. Shadow AI, Insider Threats and Organizational Uploaded Risks 29
Shahed A. AL-TAMIMI and Qasem Abu AL-HAIJA

2.1. Introduction 29
2.1.1. Understanding Shadow AI 31
2.1.2. Motivation 33
2.1.3. Contribution 34
2.1.4. Chapter organization 34
2.2. History of Shadow AI and fundamental concept 35
2.3. Insider threats 36
2.3.1. Types of insider threats 37
2.3.2. Causes and motivations of insider threats 38
2.3.3. AI-based detection of insider threats 38
2.3.4. Mitigation strategies for insider threats 39

2.4. Organizational uploaded risks 44
2.4.1. Data leakage risk 44
2.4.2. Intellectual property exposure 44
2.4.3. Compliance and regulatory risks 44
2.4.4. Security vulnerabilities risks 45
2.4.5. Loss of data governance and control 45
2.5. Threat modeling for Generative AI systems 45
2.5.1. Applications for AI in organizational security 45
2.5.2. Usage of AI technologies in organizations 46
2.6. Case studies 47
2.7. Open challenges 48
2.8. Mitigating Shadow AI and insider threats: policies, governance and human factors 49
2.8.1. Organizational policies and governance 49
2.8.2. Technical mitigation approach 50
2.8.3. Human-centered risk mitigation 50
2.8.4. Risk assessment and continuous improvement 51
2.9. Conclusion and future work 53
2.10. References 53

Chapter 3. Mitigating Data Leakage in Gen-AI Cloud Uploads: Detection Frameworks and Best Practices 57
Lourdu Akhila YERUVA, Shubneet and Anushka Raj YADAV

3.1. Introduction 58
3.2. Foundations of Gen-AI cloud uploads for data protection 61
3.3. Semantic DLP for Gen-AI prompts 63
3.4. File and multimodal upload inspection 66
3.5. Real-time risk scoring and policy evaluation 70
3.6. Data preparation, encoding and false-positive management 73
3.7. Compliance and regulatory risks of Gen-AI cloud uploads 77
3.8. Advanced hybrid and privacy-preserving enforcement 79
3.9. Case studies and pilots of Gen-AI upload protection 83
3.10. Conclusion 88
3.11. References 89

Chapter 4. Threat Modeling for Upload-Induced Data Leakage in Generative AI 95
Nikita SHARMA, Bindiya JAIN, Hemlata JAIN and Manisha GAUR

4.1. Introduction 95
4.2. System model: upload-driven generative AI architecture 97
4.2.1. Architectural components 97
4.2.2. Data flow model 100

4.3. Formal definition and taxonomy of upload-induced data leakage 101
4.3.1. Formal definition 103
4.3.2. Taxonomy of upload-induced leakage 103
4.3.3. Structural characteristics of upload-induced leakage 105
4.3.4. Toward a generative-aware leakage model 106
4.4. Adversary model 106
4.4.1. Threat actors 106
4.4.2. Capability spectrum 107
4.4.3. Capability × leakage matrix 108
4.4.4. Adversarial escalation pathways 109
4.4.5. Probabilistic advantage model 109
4.4.6. Implications for threat modeling 109
4.5. Threat modeling framework for upload-induced leakage 110
4.5.1. Layered threat surface model 110
4.5.2. STRIDE-GAI: extending classical threat modeling 112
4.5.3. Attack tree modeling 113
4.5.4. Generative-specific modeling considerations 114
4.6. Quantitative risk assessment framework for upload-induced leakage 114
4.6.1. Risk dimensions 115
4.6.2. Conceptual risk model 116
4.6.3. Probabilistic amplification factor 116
4.6.4. Operational application 117
4.6.5. Limitations and future refinement 117
4.7. Mitigation mapping: from leakage threats to defensive controls 117
4.7.1. Design principles for mitigation 118
4.7.2. Threat-to-control mapping 118
4.7.3. Architectural mitigations 119
4.7.3.1. Tenant-isolated retrieval 119
4.7.4. Governance controls 120
4.7.5. Mitigation maturity levels 120
4.7.6. From control to assurance 121
4.8. Regulatory and compliance implications 121
4.9. Conclusion 122
4.10. References 124

Chapter 5. Secure Threat Modeling of Upload-based Data Leakage in Generative AI Systems 127
Shahed A. AL-TAMIMI and Qasem Abu AL-HAIJA

5.1. Introduction 127
5.1.1. Overview of threat modeling and data leakage 129
5.1.2. Motivation 130
5.1.3. Contribution 130
5.1.4. Chapter organization 130

5.2. Fundamental background 131
5.2.1. Generative AI architecture overview 131
5.2.2. Data flows in Generative AI platforms 133
5.2.3. Data lifecycle within AI systems 133
5.3. Data leakage cyber threat 134
5.3.1. Data leakage overview 134
5.3.2. Data leakage types 136
5.3.3. Critical data classification 136
5.3.4. Causes of data leakage 137
5.3.5. Data leakage threats in Generative AI. 138
5.3.6. Impact of data leakage 139
5.4. Upload-induced data leakage within Generative AI 140
5.4.1. Upload-induced data leakage overview 140
5.4.2. Channels in AI systems uploaded 141
5.4.3. Data exposure risks 141
5.5. Threat modeling for Generative AI systems. 142
5.5.1. Threat modeling overview 142
5.5.2. Threat modeling methodologies 143
5.5.3. Threat modeling approaches 144
5.5.4. Suggested threat model for upload-induced data leakage 144
5.6. Risk assessment and mitigation. 145
5.6.1. Attack vectors in Generative AI systems 145
5.6.2. Risk analysis. 146
5.6.3. Impact risk assessment 146
5.6.4. Mitigation strategies and security controls 147
5.7. Open challenges 148
5.8. Conclusion and future work. 149
5.9. References 150

Chapter 6. Uncontrolled Generative AI Uploads and Sector-Specific Data Leakage: A Statistical Case Study Across Healthcare, Finance, Legal and Software Domains. 155
N. MANI, V. KAVITHA and B.L. SHIVAKUMAR

6.1. Introduction 156
6.1.1. Generative AI entering enterprise spaces 156
6.1.2. Data leakage concept under upload 156
6.1.3. Difference with the traditional cybersecurity breaches 157
6.2. Theoretical and statistical framework 158
6.2.1. Taxonomy of Gen-AI data leakage 158
6.2.2. Statistical risk modeling technique 160
6.2.3. Regulatory environment of the industry 161

6.3. Case study I: healthcare industry 162
6.3.1. Scenario: Generative AI systems electronic health records (EHR)-upload 162
6.3.2. Nature of sensitive data. 163
6.3.3. Risk amplifiers 163
6.3.4. Statistical assessment 164
6.3.5. Mitigation strategies 165
6.4. Case study II: financial services 166
6.4.1. Trading models and client portfolios uploaded by the company 166
6.4.2. Types of exposed assets 166
6.4.3. Threat vectors 167
6.4.4. Quantitative risk modeling 167
6.4.5. Framework 168
6.5. Case study III: legal sector 169
6.5.1. Scenario: uploading of confidential client documents 169
6.5.2. Ethics and professional responsibilities 169
6.5.3. Privilege breach probability modeling 170
6.5.4. Risk of precedent exposure 170
6.5.5. Safeguards 171
6.6. Case study IV: software and technology industry. 172
6.6.1. Situation: upload of proprietary source code 172
6.6.2. Risks 173
6.6.3. Statistical evaluation 173
6.6.4. Mitigation strategies 174
6.7. Risk severity matrix 175
6.7.1. Sectoral vulnerability ranking 176
6.8. An integrative statistical risk-scoring framework 178
6.8.1. Gen-AI Upload Risk Index (GAURI) construction 178
6.8.2. Weighted scoring model 179
6.8.3. Implementation guidelines 179
6.9. Implications of policy and governance. 181
6.9.1. AI governance models in organizations. 181
6.9.2. Mandatory upload protocols 182
6.9.3. Auditing and compliance systems 183
6.9.4. Future research directions 183
6.10. Conclusion 184
6.11. References 185

Chapter 7. Role of Digital Forensics Techniques in Sustaining Generative AI Systems 189
Muhammad Ahsan QURESHI

7.1. Introduction 189

7.2. Foundations: digital forensics and AI ecosystems 191
7.2.1. Fundamentals of digital forensics 191
7.2.2. The evolving landscape of AI forensics. 193
7.2.3. Architecture of GenAI systems 193
7.2.4. Convergence challenges 194
7.3. Role of AI in digital forensics 195
7.3.1. NLP in forensic analysis 195
7.3.2. AI-driven threat analysis and protection 196
7.3.3. AI for evidence classification and triage 197
7.3.4. AI in malware and exploit attribution 197
7.3.5. AI-driven attribution of social media artifacts 198
7.3.6. Deepfake and synthetic media forensics 198
7.3.7. AI for timeline reconstruction 198
7.4. Forensic readiness in GenAI systems 199
7.4.1. Logging architecture for GenAI systems 200
7.4.2. Prompt and interaction logging 200
7.4.3. Retrieval traceability in RAG systems 200
7.4.4. Cryptographic evidence integrity and secure time synchronization. 201
7.5. Evidence sources in GenAI ecosystems 202
7.5.1. Data-layer artifacts 202
7.5.2. Model-layer artifacts 203
7.5.3. Prompt artifacts 203
7.5.4. Vector database artifacts 203
7.5.5. Runtime and infrastructure logs 204
7.5.6. Third-party and supply-chain artifacts 204
7.6. Acquisition and preservation in AI environments. 205
7.6.1. Live evidence acquisition in AI systems 205
7.6.2. Capturing ephemeral inference states 206
7.6.3. Handling encrypted and confidential artifacts 206
7.6.4. Reproducibility and deterministic replay 207
7.6.5. Distributed chain-of-custody mechanisms 207
7.7. AI-powered tools and technologies in digital forensics 208
7.7.1. Machine learning for log analysis and event correlation 208
7.7.2. AI-driven malware detection and classification 209
7.7.3. Deepfake detection and synthetic media analysis 209
7.7.4. NLP for evidence analysis 210
7.7.5. Graph-based investigation and relationship analysis. 210
7.7.6. AI-augmented security information and event management systems 210
7.8. Advanced forensic analysis for GenAI. 211
7.8.1. Prompt injection tracing 211
7.8.2. Retrieval poisoning detection 211

7.8.3. Model tampering detection 212
7.8.4. Output watermark and provenance verification 213
7.9. Sustaining GenAI systems through continuous forensic assurance 213
7.9.1. Continuous auditing and compliance 214
7.9.2. Simulation of forensic intelligence 214
7.9.3. Drift monitoring as a forensic signal 215
7.10. Legal, ethical and governance considerations 215
7.10.1. Tradeoffs between privacy and logging 216
7.10.2. Cross-border evidence challenges 216
7.10.3. AI explainability versus forensic transparency 217
7.10.4. Standards and regulatory alignment 217
7.11. The future of AI in digital forensics 218
7.11.1. XAI for court-admissible evidence 218
7.11.2. Privacy-preserving and federated forensics 218
7.11.3. Quantum-resilient evidence integrity 219
7.12. Conclusion 219
7.13. References 220

Chapter 8. Forensic-Driven Security for Generative AI Detecting Upload-Induced Data Leakage 223
Mritunjay Kr. RANJAN, Parimala MANI, Pranjal SONJE, Rajashri RIKAME, Disha ARSUDE and Ankita PATIL

8.1. Introduction 224
8.1.1. Objective. 226
8.2. Related study 226
8.3. Problem statement 228
8.4. Methodology 230
8.5. System architecture 236
8.6. Results and discussion. 238
8.7. Conclusion 243
8.8. Future work 244
8.9. References 244

Chapter 9. DeepFakeGuard: A Multimodal Framework for Detecting Misinformation Driven by GenAI Data Leakage 247
Bayan ALFARRAYEH, Mais ALASHQAR, Marya SABEHA, Farah Abu AL-SOUD and Qasem Abu AL-HAIJA

9.1. Introduction 247
9.2. The deepfake threat landscape 250
9.2.1. Evolution of generative architectures 251
9.2.2. Multimodal expansion of deepfake techniques 251

9.2.3. Deepfakes within the misinformation ecosystem 252
9.2.4. Cybersecurity and socio-technical implications 253
9.2.5. Necessity of multimodal detection 253
9.3. Related work and research gaps 253
9.3.1. Evolution of deepfake generation and detection 253
9.3.2. Unimodal versus multimodal detection 254
9.3.3. Robustness limits 255
9.3.4. Dataset bias 257
9.4. Datasets and generator artifacts. 258
9.4.1. The datasets 258
9.4.2. Categorizing generator artifacts: GANs, diffusion and transformers 261
9.4.3. Dataset preprocessing and label harmonization 261
9.5. DeepFakeGuard system architecture 261
9.5.1. The input acquisition and preprocessing layers 262
9.5.2. Modular modality-specific analysis pipelines 263
9.5.3. The logic and rationale for late fusion. 264
9.5.4. System requirements and stakeholder integration 266
9.5.5. Strategic implementation and scalability 266
9.6. Transformer-based text deepfake detection 267
9.6.1. Theoretical foundations of transformer-based detection. 267
9.6.2. DeepFakeGuard text detection architecture. 268
9.6.3. Model selection and comparative strategy 268
9.6.4. Statistical and linguistic indicators of synthetic text 269
9.6.5. Robustness challenges and concept drift 270
9.6.6. Limitations and future directions 270
9.7. Multimodal fusion and decision strategy 271
9.7.1. Overview of the fusion design 271
9.7.2. Independent modality pipelines and decision-level outputs 271
9.7.3. SoftMax-based probabilistic decision modeling 274
9.7.4. Formal late fusion formulation. 274
9.7.5. Robustness to missing and noisy modalities 274
9.7.6. Implications for practical deployment. 275
9.8. Cybersecurity and societal implications 275
9.8.1. Deepfakes as an emerging cybersecurity threat 275
9.8.2. Shift from static to AI-driven threat vectors 275
9.8.3. Multimodal detection for social engineering defense 276
9.8.4. Societal impact and digital trust 276
9.8.5. Digital forensics and institutional implications. 277
9.9. Evaluation strategy performance validation 277
9.9.1. Performance metrics 277
9.9.2. Cross-validation and statistical stability. 278
9.9.3. Cross-dataset generalization 278

9.9.4. Multimodal fusion evaluation 278
9.9.5. Scalability and deployment analysis 279
9.9.6. Security and adversarial considerations 279
9.10. Limitations and future directions 279
9.10.1. Scope and modality integration limitations 279
9.10.2. Dataset bias and generalization constraints 279
9.10.3. Computational and scalability challenges 280
9.10.4. Future directions for system enhancement 280
9.10.5. Long-term research and deployment outlook 280
9.11. Conclusion 281
9.12. References 281

Chapter 10. A Post-Quantum Secure Automata-Governed Framework for Preventing Upload-Induced Data Leakage in Generative AI Systems Using Lattice Cryptography 283
Rohit GUPTA, Amit KUMAR, Shilpa BHOJNE, Jyotsna KULKARNI, Ankita PATIL and Mritunjay Kr. RANJAN

10.1. Introduction 284
10.1.1. Objective 285
10.2. Related study 285
10.2.1. AI-driven post-quantum security framework 287
10.2.2. Lattice-based post-quantum cryptography for secure data systems 287
10.2.3. Post-quantum cyptography integration in AI and distributed learning . . 287
10.3. Problem statement 287
10.4. Methodology 288
10.5. System architecture 298
10.6. Results and discussion 299
10.7. Conclusion 302
10.8. Future work 303
10.9. References 303

Chapter 11. An Automata-Governed Self-Healing Cybersecurity Architecture Integrating Machine Learning and Upgradeable Blockchain Smart Contracts 307
Pankaj PATIL, Mritunjay Kr. RANJAN, Monali JADHAV, Jayashree KUNDE, Ishwari RAKIBE and Disha ARSUDE

11.1. Introduction 308
11.1.1. Objectives 309
11.2. Related study 309
11.2.1. Building cyber resilience in modern digital infrastructures 310
11.2.2. ML-based cybersecurity framework for intelligent threat detection 310

11.2.3. Blockchain-based secure data management for distributed systems . . . 310
11.3. Problem statement . . . 312
11.4. Methodology . . . 314
11.5. System architecture . . . 322
11.6. Results and discussion . . . 323
11.7. Conclusion . . . 326
11.8. Future work . . . 327
11.9. References . . . 328

Chapter 12. A Privacy-Aware Hybrid Automata–Machine Learning Framework for Secure Real-Time Fraud Detection in Satellite-enabled Financial Systems . . . 331
Vivek NAKHATE, Mritunjay Kr. RANJAN, Sarika KONDEKAR, Gaurav PATIL, Rajashri RIKAME and Pranjal SONJE

12.1. Introduction . . . 331
12.1.1. Objective . . . 333
12.2. Related study . . . 333
12.3. Problem statement . . . 335
12.4. Methodology . . . 335
12.5. System architecture . . . 344
12.6. Results and discussion . . . 345
12.7. Conclusion . . . 348
12.8. Future work . . . 349
12.9. References . . . 349

Chapter 13. Mitigating Gen-AI Upload-induced Data Leakage with Federated TinyML on ESP32: A Privacy-Preserving Edge Architecture for Rural Biogas Production Forecasting . . . 353
Amine CHIGUER, Mohammed SADIK and Abderrahim FAIL

13.1. Introduction . . . 354
13.2. Materials and methods . . . 356
13.2.1. Embedded hardware platform . . . 356
13.2.2. FL system architecture . . . 356
13.2.3. Hybrid data acquisition strategy . . . 357
13.2.4. ML problem formulation . . . 358
13.2.5. Local training and backpropagation on the edge . . . 358
13.2.6. Federated aggregation protocol. . . . 359
13.2.7. Communication, compression and security . . . 359
13.2.8. Evaluation protocol . . . 360
13.3. Results. . . . 361
13.3.1. Baseline TinyML (non-federated) . . . 361
13.3.2. Federated versus local-only versus centralized oracle . . . 363

13.3.3. Communication-accuracy trade-offs 364
13.3.4. Resource usage and payload characterization 364
13.3.5. Robustness to non-IID data and client dropout 365
13.4. Discussion. 366
13.4.1. Interpretation of results 366
13.4.2. Privacy-by-architecture and Gen-AI data leakage mitigation 367
13.4.3. Practical deployment considerations 369
13.4.4. Limitations and lessons learned 369
13.4.5. Future work 370
13.5. Conclusion 371
13.6. References 372

List of Authors 377

Index. 383