Securing AI Using Zero Trust Principles



CINDY GREEN-ORTIZ
ZIG ZSIGA
SASKIA LAURA SCHRÖER

ciscopress.com

FREE SAMPLE CHAPTER |

Securing AI Using Zero Trust Principles

Cindy Green-Ortiz
Zig Zsiga
Saskia Laura Schröer

Cisco Press

Securing AI Using Zero Trust Principles

Cindy Green-Ortiz, Zig Zsiga, Saskia Laura Schröer

Copyright © 2026 Cisco Systems, Inc.

Published by:
Cisco Press
Hoboken, New Jersey

All rights reserved. This publication is protected by copyright, and permission must be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or likewise. For information regarding permissions, request forms, and the appropriate contacts within the Pearson Education Global Rights & Permissions Department, please visit https://www.pearson.com/global-permission-granting.html.

No patent liability is assumed with respect to the use of the information contained herein. Although every precaution has been taken in the preparation of this book, the publisher and authors assume no responsibility for errors or omissions. Nor is any liability assumed for damages resulting from the use of the information contained herein.

$PrintCode

Library of Congress Control Number: 2026934064

ISBN-13: 978-0-13-836341-3

ISBN-10: 0-13-836341-2

Warning and Disclaimer

This book is designed to provide information about securing AI using Zero Trust principles. Every effort has been made to make this book as complete and as accurate as possible, but no warranty or fitness is implied.

The information is provided on an "as is" basis. The authors, Cisco Press, and Cisco Systems, Inc. shall have neither liability nor responsibility to any person or entity with respect to any loss or damages arising from the information contained in this book or from the use of the discs or programs that may accompany it.

The opinions expressed in this book belong to the authors and are not necessarily those of Cisco Systems, Inc.

Trademark Acknowledgments

All terms mentioned in this book that are known to be trademarks or service marks have been appropriately capitalized. Cisco Press or Cisco Systems, Inc., cannot attest to the accuracy of this information. Use of a term in this book should not be regarded as affecting the validity of any trademark or service mark.

Feedback Information

At Cisco Press, our goal is to create in-depth technical books of the highest quality and value. Each book is crafted with care and precision, undergoing rigorous development that involves the unique expertise of members from the professional technical community.

Readers' feedback is a natural continuation of this process. If you have any comments regarding how we could improve the quality of this book, or otherwise alter it to better suit your needs, you can contact us through email at feedback@ciscopress.com. Please make sure to include the book title and ISBN in your message.

We greatly appreciate your assistance.

Head of IT & Professional Learning, Enterprise Learning and Skills: Julie Phifer

Alliances Manager, Cisco Press: Caroline Antonio

Executive Editor: James Manly

Managing Editor: Sandra Schroeder

Development Editor: Ellie C. Bru

Senior Project Editor: Tonya Simpson

Copy Editor: Chuck Hutchinson

Technical Reviewers: Petar Radanliev, PhD. Jennifer Vaniel

Cover Designer: Chuti Prasertsith

Compositor: codeMantra

Indexer: Timothy Wright

Proofreader: Donna E. Mulder



Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices.**

Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: www.cisco.com/go/trademarks. Third party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1110R)

About the Authors

Cindy Green-Ortiz is a globally recognized cybersecurity strategist, principal architect, and trusted advisor to Fortune 100 enterprises and public sector leaders. With 40 years of experience in security and technology leadership, Cindy has guided complex organizations across industries—including financial services, healthcare, hospitality, military, energy, and manufacturing—in transforming their security posture and achieving sustainable business outcomes.

At Cisco, Cindy is a principal security architect, leading global Zero Trust initiatives. She co-led Cisco's AI program, Wintermute, and co-led the Post Quantum Resistant Cryptography Working Group. As a Cisco Press author, Cindy translates technical depth into practical frameworks that inspire engineers, architects, and executive leaders alike. Her ability to bridge cutting-edge technology with security management has made her a sought-after public speaker and educator.

Cindy has delivered impactful presentations and workshops at Cisco Live, Offensive Summit, Cisco Secure, ISC2, ISACA, and WiCyS, engaging audiences from the boardroom to the classroom. She is committed to mentoring the next generation of cybersecurity professionals and frequently collaborates with academic institutions to advance cybersecurity education, research, and policy.

Publications:

- Cisco Live Speaker:
  - BRKXAR-2008: Navigating the Future of Cybersecurity: AI, Quantum-Resistant Cryptography, and Zero Trust (2025)
  - BRKXAR-2008: Exploring the Paradigm Shift in Security: AI and Quantum Cryptography's Influence on Zero Trust (2025)
- Co-Author: *Zero Trust Architecture* (2023)
- Technical Editor: *In Zero Trust We Trust* (2024)
- Cisco Insider Advocate: "Ask Me Anything" (2023)
- Co-Author of a Cisco Whitepaper: "IPv6 Addressing Analysis for the US Army" (Clearance required, 2023)
- Cisco Live Speaker: "BRKXAR-2008 Zero Trust Segmentation" (2022–2024)
- DHS Whitepapers:
  - "Going Dark: Impacts of Encryption" (2017)
  - "Digital Blackmail (Ransomware) as an Emerging Tactic" (2016)
  - Wang Laboratories, Chairman's Golden Circle

Zig Zsiga, CCDE 2016::32, CCIE #44883, CISSP, has been in the networking industry for 20 years. He is currently a principal architect supporting the Cisco CX U.S. public sector business and customers. Zig holds an active CCDE and two CCIE certifications, one in Routing and Switching and the second in Service Provider. He also holds a BS in computer science from Park University. He is a father, a husband, a United States Marine, a gamer, a nerd, a geek, and a big soccer fan. Zig loves all technology and can usually be found in the lab learning and teaching others. This is his second published book, and he is also the host of the Zigbits Network Design Podcast (ZNDP), where he interviews leading industry experts about network design. All of Zig's content is located at https://zigbits.tech.

Publications:

- Author: *Cisco Certified Design Expert (CCDE 400-007) Official Cert Guide* (2023)
- Cisco Insider Advocate: "Ask Me Anything" (2023)
- Cisco Live Speaker:
  - BRKRST-2044: Enterprise Multi-Homed Internet Edge Architectures (2017–Present)
  - LTRENT-2016: Learning IPv6 in the Enterprise for Fun and (Fake) Profit: A Hands-On Lab (2018–Present)
  - TECCRT-3005: CCDE Techtorial (2019–Present)
  - LTRCRT-3000: CCDE Practical Exam Practice Lab (2020–Present)
  - LTRENT-2016: Learning VxLAN in the Enterprise for Fun and (Fake) Profit: A Hands-On Lab (2024–Present)

Saskia Laura Schröer holds a PhD with a specialization in Information Systems and an MSc focusing on data science. Saskia is certified in CCNP Enterprise and is a DevNet Associate, a WiCyS speaker, and a Cisco Live speaker. Saskia is a security consulting engineer in Cisco's EMEA Cybersecurity Centre of Excellence with almost 10 years of experience in consulting, IT audit, network engineering, and security. Her focus lies on the technical and organizational aspects of cybersecurity, across various sectors.
In her PhD, supervised by Prof. Pavel Laskov, Saskia has developed a core expertise at the intersection of cybersecurity and artificial intelligence, which she is leveraging to drive innovation at Cisco.

Publications:

- Cisco Live Speaker: BRKXAR-1009: Exploring the Paradigm Shift in Security: AI and Post-Quantum's Influence on Zero Trust (2026)
- Schröer, S. L., Apruzzese, G., Human, S., Laskov, P., Anderson, H. S., Bernroider, E. W., … & Wang, G. (2025, April). "SoK: On the Offensive Potential of AI." In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)* (pp. 247–280). IEEE.

- Schröer, S. L., Pajola, L., Castagnaro, A., Apruzzese, G., & Conti, M. (2025). "Exploiting AI for Attacks: On the Interplay Between Adversarial AI and Offensive AI." *IEEE Intelligent Systems*.
- Schröer, S. L., Canevascini, N., Pekaric, I., Widmer, P., & Laskov, P. (2025, June). "The Dark Side of the Web: Towards Understanding Various Data Sources in Cyber Threat Intelligence." In *2025 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)* (pp. 79–89). IEEE.
- Schröer, S. L., Seideman, J. D., Luo, S., Apruzzese, G., Dietrich, S., & Laskov, P. (2025). "Using a Stack to Find an AI Needle: Topic Modeling for Cyber Threat Intelligence." *Digital Threats: Research and Practice*.
- Weinz, M., Schröer, S. L., & Apruzzese, G. (2024, September). "'Hey Google, Remind Me to Be Phished': Exploiting the Notifications of the Google (AI) Assistant on Android for Social Engineering Attacks." In *2024 APWG Symposium on Electronic Crime Research (eCrime)* (pp. 109–122). IEEE.

About the Technical Reviewers

Dr. Petar Radanliev lectures and supervises postgraduate master's students' research dissertations on AI and cybersecurity at the Department of Computer Science, University of Oxford. He is also a Lecturer/Instructor at Pearson and O'Reilly (USA), while conducting research on digital identity system security at the Alan Turing Institute, based at the British Library in London.

After completing his PhD in 2013–14, Petar held postdoctoral research appointments at Imperial College London, the University of Cambridge, the Massachusetts Institute of Technology, and the Department of Engineering Science at the University of Oxford, where he remained for seven years before moving to his current position.

His work spans artificial intelligence, cybersecurity, post-quantum security, and blockchain security. This research has led to an H-index of 25 (as indexed by Web of Science and Scopus), more than 3,700 citations, more than 100 peer-reviewed publications, and four authored books.

In recognition of his contributions, Petar has received major funding awards, including a Fulbright Fellowship and the Prince of Wales Innovation Award.

Jennifer Vaniel, GMLE, is a security software engineer with more than 25 years of experience. As a software engineering technical lead at Cisco, Jennifer focuses on integrating multiple security solutions to deliver secure and efficient outcomes for customers. Prior to joining Cisco, Jennifer spent 12 years at Juniper Networks, focused on network security.

She has deep expertise in Zero Trust segmentation, intrusion detection and prevention, application development, and large-scale network security analytics. Jennifer is dedicated to empowering teams by developing secure automation and tools that incorporate AI/ML processes and enable customer success.

Outside of work, Jennifer spends her time volunteering in the local school district, spending time outdoors, gaming with her husband and daughters, and enjoying live music.

Dedications

To my beloved husband, Erick Ortiz-Alvarenga, every day is a blessing, and I am truly grateful for all your love and support. Your opinions challenge me to think differently or to look at things from an alternate perspective. To our daughters, Angela and Anna, what a bright future you have. Know that you are both loved and supported to reach your dreams. You both inspire me every day! I know you both will change the world. To my Uncle Roger Green and my Aunt Joan Green, you bring me joy and always have a great story from back home. To my parents in heaven, Howard L. Green and Nancy Salyers Green, I would not be where I am without your courage to strive for knowledge, sometimes from the top of a mountain at night taking college courses via radio, which led me to a better life.

—Cindy Green-Ortiz

I dedicate this book to you, the reader. May it help you on your AI journey!

—Zig Zsiga

To my mum, whose early death showed me how short life can be and motivated me to pursue my dreams. I will forever miss you. To my dad, who always encouraged me to step out of my comfort zone and made me believe I could achieve anything I dream of in life. And to my boyfriend, Tom, I could not imagine a better partner by my side. I would not be where I am today without your support.

—Saskia Laura Schröer

Acknowledgments

With 40 years in this field, I have too many to thank for your help, guidance, and support. To name but a few, I would like to give special recognition to my lifelong mentor Denis McDuff, to my mentors and colleagues: Jennifer Vaniel, Lisa Bobbitt, Ginny Lee, John Strong, James Granozio, Theresa Payton, David Ankeney, Demetria Davis, Jason Penn, Chris Mula, Maurice DuPas, and this authoring team, who have seen me at my best or at my worst and have helped me navigate life or work's ever-changing landscape. I am ever grateful to my high school science teacher, Dr. Demchek, and my piano teacher, Edrie Ballard, who set me on my path.

—Cindy Green-Ortiz

Thank you to my Cisco leadership team: Maurice DuPas and Jim Lien. You continue to give me the space, time, and overwhelming encouragement to make projects like this happen. Thank you for always supporting me and my career aspirations. Thank you, James Manly and Ellie Bru. This entire process would have been truly unbearable without the two of you. You have been with us from the start. With the many changes this book has gone through, it simply wouldn't be successful without the two of you. My co-authors, Cindy and Saskia, I couldn't have asked for a better team to make this book happen. It was an honor to work with the two of you on this project. We did it!

—Zig Zsiga

There are many people who supported me during my career, always believing in me and encouraging me to be the best version of myself, both personally and professionally. To name a few, I would like to thank Martin Schönbacher, David Byrne, Barry Reddan, Meg Rainbow, Tim Wadwha-Brown, Giovanni Apruzzese, and Prof. Pavel Laskov, my PhD supervisor. And of course, a special thank you goes to Cindy and Zig for taking me on this journey to write a book. It has been an honor to learn from and work with you!

—Saskia Laura Schröer

In Memory of David Houck, deceased August 2024: The writing team of *Zero Trust Architecture* spent countless hours together, both physically and virtually as we have navigated our extensive careers. In 2017, I recruited David Houck into Cisco Security Services, seeing his potential, drive, and empathetic nature that he brought to customers while working for Cisco TAC. In his dedication, he wrote "For all who teach, inspire, and mentor us." In intersection with his phrasing, my acknowledgment was meant to include people like David who shared the passion for knowledge sharing, mentorship, and a scrupulous attitude to benefit the industry and our customers. "In a technology career spanning multiple decades, one does not reach the apexes in technical knowledge without exposure and guidance from some of the best managers, mentors, and sounding boards in the industry."

David was one of those sounding boards whom I would reach out to on a regular basis, as I knew he would help challenge ideas, assumptions, and perceptions with a vast set of knowledge gained across all of his roles and experiences. When I moved back to RTP, he was one of the first to welcome me back, barbecue with me, and share his jolly personality with everyone involved. He was passionate about charitable giving, and regularly raised money for food banks and causes he believed in. An entire generation of interns can credit him for guidance he provided during their tenure at Cisco, along with the countless colleagues and strangers he helped along the way. The loss of his knowledge, empathy, and drive can never be replaced, but can only serve as a guiding light for us to aspire to. †

—Patrick Lloyd

In Memory of Nick Russo, deceased June 2024: Nick, you were a beacon of knowledge for all of us, a true leader and pioneer in this industry. We love you and miss you. Nick was my co-worker, my friend, and my brother. He was the person I went to when I was suffering from imposter syndrome. We all suffer from imposter syndrome in this industry. He was my sounding board when I didn't believe in myself, be it a new job, presenting at Cisco Live for the first time, writing my first book, creating my first course, Nick was always there for me. Nick was one of my best friends; I will miss all of him! I think it's fitting to share the traits, tactics, and decisions I have adopted because of Nick's influence:

- Nick taught me that watching TV and playing video games are a waste of time by yourself. We only have so much time in life; we can't waste any of it. It's different when you are doing these things with someone; it's an experience. Live for the experiences in life!
- I signed up for Audible because Nick gave me his famous 20-point debate in two minutes on why it was beneficial. He wasn't wrong; he rarely was wrong in fact. I listen to more books than I read every week, month, and year because of him.
- He showed me that giving more of yourself is how to live, and I have tried to constantly give back to each person I interact with because of him.
- If you are going to do something, do it to the best you can. Never cut corners. All in or not at all.

- Nick had the ability to take something new, understand it rapidly, and make decisions on it in minutes.
- Your brand is everything, and word of mouth travels fast. Nick's brand was him, and it's why I started my own brand.
- Nick's work ethic was impressive, to say the least. One hour of Nick's time was like nine hours of everyone else's time. In that sense, I use Nick as a sounding board and my bartender.

Nick would get up early to get the most important tasks done before he was pulled into anything else. Because of him, I get up at 5 a.m. and am a lifelong insomniac. Nick showed me what time blocking is and why it is a better way of making sure you get done what you need to get done. Nick used to make fun of me; he would say I was trying to make progress on 40 different projects a week, only to move the needle on each project one hour forward. He would say, "Focus on that one project until it's 100% completed." Here is my way to thank Nick: Nick, thank you for all that you have done and the impact you have had on my entire life, this company, and the entire IT community worldwide. I would not be who I am today if it weren't for you. Until we meet again, Semper Fi Brother!"

—Zig Zsiga

Contents at a Glance

Introduction xliv

Part I: Defining Responsible AI and the Evolving AI Landscape

Chapter 1 Overview 1

Chapter 2 Responsible AI and Integrated Awareness 29

Chapter 3 Artificial Intelligence Threat Landscape 67

Chapter 4 Zero Trust Principles and Methods 107

Chapter 5 Securing AI from the Start 173

Part II: Building Operational Resilience—People, Processes, and Infrastructure

Chapter 6 Organizational AI Security Readiness 225

Chapter 7 AI-Ready Data Privacy and Business Impact 255

Chapter 8 Third-Party AI Risk 303

Chapter 9 Build AI-Ready Environments 353

Part III: Defending at Scale—Platform Protection, Monitoring, and Response

Chapter 10 Build and Secure Enterprise-Grade Generative AI Applications: ChatAI, RAG, MCP, Agentic AI, and Embedded AI 401

Chapter 11 Monitor and Respond to AI Security Threat Vectors 447

Chapter 12 Case Study 485

Chapter 13 Conclusion: Vigilance, Policy, Iteration, and Beyond 515

Appendix: Study Guide 545

Glossary 595

Afterword 609

Final Word 613

Index 615

Reader Services

Register your copy at www.ciscopress.com/title/9780138363413 for convenient access to downloads, updates, and corrections as they become available. To start the registration process, go to www.ciscopress.com/register and log in or create an account*. Enter the product ISBN **9780138363413** and click Submit. When the process is complete, you will find any available bonus content under Registered Products.

*Be sure to check the box that you would like to hear from us to receive exclusive discounts on future editions of this product.

Contents

Introduction xliv

Part I: Defining Responsible AI and the Evolving AI Landscape

Chapter 1 Overview 1

Foundations of Zero Trust in AI Security 3

The Origins and Evolution of Zero Trust 3

Zero Trust Principles in AI Security 4

Key Frameworks and Regulations 5

NIST SP 800-207 6

NIST AI Risk Management Framework 6

Key Provisions of AI RMF 7

EU AI Act 7

Key Provisions of the EU AI Act 8

The Intersections of AI and Security 8

From Reactive to Proactive Security with AI 10

Zero Trust as a Paradigm Shift in Securing AI 11

Ways to Build AI-Ready Data Centers and Cloud Architecture 12

Network Design Basics with AI in Mind 13

Key Components Required for an AI-Ready Environment 14

AI Workloads 14

Infrastructure Requirements 15

Compute Infrastructure 15

Storage Systems 16

Networking 17

AI Data Center Deployment Options 17

AI in the Cloud 18

Hybrid Cloud 18

Multi-Cloud 19

Edge Computing and AI at the Edge 19

AI and IoT 20

Summary 21

Key Terms 22

End-of-Chapter Questions and Answers 22

Questions 22

Answers 23

References 24

Chapter 2 Responsible AI and Integrated Awareness 29

Definition and Principles of Responsible AI 29

Ethical AI Development 30

The Landscape of AI: From Basics to Advanced Concepts 31

Foundations of AI Architectures 31

Agentic AI 33

Chain-of-Thought Reasoning Models 33

Key Zero Trust Principles for AI Agents and Reasoning Models 35

Foundational Considerations in AI 35

Ways to Overcome Organizational Barriers to Secure AI Adoption 36

Underdeveloped Use Cases and Unrealized Potential 38

Siloed Decision-Making and Political Gridlock 38

Hype-Driven Investments in Immature Technologies 39

AI/ML Pipeline 39

No Free Lunch Theorem: Common Challenges in AI Development 41

Explainable AI (Is AI a Black Box?) 41

AI in Organizations 43

AI Adoption Framework 43

Essential Skills for the AI Era 44

Ethical Considerations and Bias Mitigation 45

Ethical Frameworks and Guidelines for AI 46

Additional Considerations for AI 47

Emerging Technologies 48

Internet of Things 48

Blockchain 49

Quantum Computing 49

Post-Quantum Cryptography 50

Defining an AI Maturity Model 51

Applying Zero Trust to AI Deployment Models 53

Understanding Risk, Control, and Governance Across the AI Landscape 53

Enterprise-Controlled AI (First-Party and Open-Source AI) 53

Externally Managed AI (Third-Party and Cloud AI) 54

Distributed AI (Federated and Embedded AI) 54

Autonomous Agent AI 55

Securing AI Agents Through Zero Trust Guardrails 55
Principle 1: Allow Known Good. Block Everything Else 55
Principle 2: Make Policies Readable 56
Principle 3: Log Everything 56
Principle 4: Fail Closed, Not Open 57
Principle 5: Use Multiple Layers 57
Summary 58
Key Terms 59
End-of-Chapter Questions and Answers 59
Questions 59
Answers 62
References 63

Chapter 3 Artificial Intelligence Threat Landscape 67
Overview of AI Threats 67
AI as Target: Adversarial Machine Learning 69
Threat Model 71
Integrity: Evasion, Poisoning, and Backdoor Attacks 73
Evasion Attacks 73
Data Poisoning Attacks 74
Backdoor Attacks 75
Confidentiality: Model Inversion, Extraction, and Membership Inference Attacks 76
Model Inversion Attacks 77
Model Extraction Attacks 78
Membership Inference Attacks 78
Differential Privacy 79
Availability: Energy Latency Attacks 79
Other Common Attacks: Supply Chain and Third Party 80
Specific Considerations for Attacks on Generative AI 83
Direct vs. Indirect Prompt Injection 84
Mitigation Strategies 86
Attacking AI Systems vs. AI Models 86
Libraries for Testing AI Models 88
AI Systems vs. AI Models 88
Case Studies of AI Security Events: MITRE ATLAS 89
Overview 89

Selected Case Studies 89
AI Incident Database 91
AI as Attack Vector: Offensive AI in Generative Adversarial Networks 93
Summary 97
Key Terms 97
End-of-Chapter Questions and Answers 98
Questions 98
Answers 100
References 101

Chapter 4 Zero Trust Principles and Methods 107
Benefits of Zero Trust for AI Security: A Proof of Value 107
The Evolution of Zero Trust: A Foundation for Securing AI 108
AI as a Catalyst for Zero Trust Transformation 110
Applying the Five Zero Trust Categories to AI 111
Policy and Governance 112
Change Control 113
Data Governance and Encryption 114
Data Retention 115
Quality of Service 116
Redundancy and Replication 117
Business Continuity 118
Disaster Recovery 119
Risk Classification Policy 120
Segmentation 121
Identity 122
Authentication, Authorization, and Accounting 123
Certificate Authority 123
Network Access Control 124
Provisioning 125
Privileged Access Management 126
Multifactor Authentication 127
Asset Identity 127
Configuration Management Database 128
IP Schemas and Addressing 130
Vulnerability Management 131

Endpoint Protection and Encryption 132
Malware Prevention and Detection 133
Vulnerability Management 133
Vulnerability Scanning with Threat Intelligence 134
Database Change Control 135
Enforcement 135
Cloud Access Security Broker 136
Distributed Denial of Service 136
Data Loss Prevention 137
Email Security 138
Firewall 139
Intrusion Prevention Systems 139
Proxy 140
VPN/Remote Access (RA) 141
SOAR (Security Orchestration, Automation, and Response) 141
File Integrity Monitoring 142
Segmentation 143
Analytics 144
Application Performance Monitoring 144
Auditing, Logging, and Monitoring 145
Change Detection 146
Network Threat Behavior Analytics 147
Security Information and Event Management 148
Threat Intelligence 148
Traffic Visibility 149
Asset Monitoring and Discovery 150
Practical Workshop Design: Zero Trust for AI 150
Risk and Regulation 151
Implementation Guidance 151
Capability Alignment 151
Organizational Dynamics in Zero Trust for AI 152
Risk and Regulation 152
Implementation Guidance 152
Capability Alignment 153
Roadmap: Zero Trust for AI Security Maturity 153

Risk and Regulation 153
Implementation Guidance 154
Capability Alignment 154
Application of Zero Trust: Securing Embodied AI Through Zero Trust 155
The Trust Gap in Embodied AI 155
Securing Perception, Planning, and Action 155
Simulation, Noise, and Real-World Deployment 156
Collaborative, Ethical, and Societal Risks 156
Case Study: Application of Zero Trust—Salt Typhoon and Advanced Threat Campaigns Against Embodied AI 157
Case Study: Application of Zero Trust—Implications of State-Sponsored Network Compromise Campaigns 158
Case Study: Application of Zero Trust—Real-World Technology Shift at Scale to AI-Native Software Development 160
Case Study: Nation-State Espionage, the Quantum Threat, and Harvest Now Decrypt Later 161
HNDL Description and Analysis 162
PQC Recommendations 162
PQC Insights and Business Implications 163
Summary 164
Key Terms 165
End-of-Chapter Questions and Answers 165
Questions 165
Answers 169
References 170

Chapter 5 Securing AI from the Start 173

Importance of Early Data Classification 174
Data Classification Tools 176
Data Classification 177
Data Labels 177
Case Study: Data Labeling Lessons in the Real World 177
Data Sensitivity 180
Governance and Legal Requirements 181
Potential Threats and Consequences from Missing Data Classification 183
Considerations for Data Used for AI Model Training 184
Privacy Risks of Trained Models: Membership Inference and Differential Privacy 185

Considerations for Publicly Released Data 187
Ways to Build Security into the AI Development Lifecycle 189
Proactive vs. Reactive Security Measures 191
Business and Operational Benefits 193
Quantitative and Qualitative Metrics 193
Value Propositions 193
Cost Savings from Early Security Implementation 194
Improved Trust and Compliance 196
Ways to Future-Proof AI Systems by Building Crypto-Agility for Post-Quantum Resilience 197
Scalability and Adaptability of Secure AI 198
The Need to Secure AI from the Start: Challenges and Considerations 199
Securing AI Application Development 200
Securing AI Application Deployment 201
Moving from Software Development to AI Application Development 201
DevOps vs. MLOps: Tailoring Operations for Machine Learning 203
CI/CD: The Engine of Continuous Delivery 204
DevSecOps and MLSecOps: Embedding Security Throughout Development Lifecycles 205
Securing AI Chatbots and Agents 206
Understanding the Advanced Threat Landscape and Mitigation 206
Securing Agentic AI and Retrieval-Augmented Generation 207
AI Security Readiness Framework 209
1. Embedding Security in AI Governance and Strategy 211
2. Strengthening Data Security and Privacy 212
3. Ensuring Model Integrity and Robustness 212
4. Mitigating AI-Specific Threats and Attack Vectors 214
5. Addressing Compliance and Ethical Requirements 214
6. Building a Security-Resilient Infrastructure 215
7. Cultivating a Security-Aware Culture 216
Summary 217
Key Terms 218
End-of-Chapter Questions and Answers 218
Questions 218
Answers 221
References 222

Part II: Building Operational Resilience—People, Processes, and Infrastructure

Chapter 6 Organizational AI Security Readiness 225

Assessing Organizational Readiness 226
- Stakeholder Engagement and Ownership 226
- Security Readiness Assessments 226
- Baseline of Current Capabilities 227
- Risk Assessment and Prioritization 228
- Gap Analysis and Areas for Improvement 229
- Compliance and Regulatory Readiness 230
- Technical Infrastructure and Tooling Evaluation 232
- Culture and Awareness Readiness 233
- Incident Response and Recovery Preparedness 233
- Actionability and Roadmap Development 234

Building a Security-First Culture 234
- Leadership and Commitment 235
- Security Policies and Governance 235
- Risk Management and Accountability 236
- Integration of Security in AI Lifecycle 236
- Cultural Change Strategies 237

Training and Awareness Programs 237
- Tailored Training Programs 238
- Awareness Campaigns 238
- Hands-On Exercises 239
- Continuous Learning 239
- The Reasons to Measure Effectiveness 239

Organizational AI Security Readiness: Challenges and Considerations 240
- AI Model Security Readiness 240
- Data Governance and Privacy for AI Readiness 241
- AI-Specific Incident Response Readiness 243
- Explainable AI (XAI) Readiness 244
- Zero Trust Principles Applied to AI Security Readiness 245
- AI Supply Chain Readiness 246

Summary 247

Key Terms 248

End-of-Chapter Questions and Answers 248
Questions 248
Answers 252
References 253

Chapter 7 AI-Ready Data Privacy and Business Impact 255

The Strategic Value of Data in the Age of AI 256
Data as an Enterprise and National Security Asset 256
The Criticality of Protecting Strategic, Classified, and Proprietary Data 257
How AI-Driven Decision Automation Amplifies Business Impact from Data Compromise 258
The Convergence of AI Data Governance and Digital Sovereignty 258
Evolving Attack Surfaces in AI Ecosystems 259
The Influence of Visionary Fiction on the AI Landscape 259
From Imagination to Implementation: Agentic and Embedded AI 259
AI as a Living System: Expanding the Threat Model 260
Zero Trust for AI Systems Reimagined 260
Science Fiction Realized, Responsibility Required 261
Privacy and Security Challenges in Agentic and Embedded AI 262
Autonomous Data Processing and Contextual Inference Without Human Oversight 263
Data Lineage, Provenance, and Chain of Custody in Distributed AI Environments 263
The Difficulty of Enforcing Access Control and Policy Verification Within Embedded Architectures 263
Risk Propagation Across Cross-Domain AI Collaboration Systems 264
Monitoring, Containment, and Assurance for Self-Adaptive Models 265
AI Model Protection and PQC Readiness 265
Model Inversion, Prompt Injection, and Data Poisoning Threats 266
Techniques for Model Watermarking, Signing, and Integrity Validation 266
Confidential Computing, Secure Enclaves, and Trusted Execution Environments 267
PQC Readiness and the Transition to Post-Quantum Encryption (FIPS 203, FIPS 204, FIPS 205) 268
Cryptographic Agility and Lifecycle Management for AI Models and Data Pipelines 269

Privacy-Preserving Data Engineering for Next-Generation AI 269
Differential Privacy, Homomorphic Encryption, and Secure Multi-Party Computation 270
Federated Learning and Encryption-in-Use for Distributed AI Training 271
PQC-Based Encryption Methods for AI Inference and Storage Environments 271
The Role of Hardware-Based Isolation and Zero-Knowledge Proofs in Preserving Privacy 272
Regulatory and Compliance Integration for AI Privacy 272
Techniques for Mapping Global Privacy and Security Mandates to AI Systems 273
Alignment with NIST AI RMF, EU AI Act, GDPR, DORA, NIS2, and Other Sectoral Frameworks 274
Industry-Specific Considerations for Financial Services, Healthcare, Energy, and Defense 275
Data Residency, Retention, and Erasure Requirements in AI-Driven Environments 276
Compliance Telemetry and Assurance Reporting for Continuous Verification 276
Zero Trust Strategies for AI Compliance and Enforcement 276
Techniques for Applying Zero Trust Principles Across AI Data Pipelines and Model Lifecycles 277
Identity-Aware Access and Microsegmentation for AI Workloads 278
Enforcement of Contextual Access Policies for Training and Inference Data 278
Continuous Assurance and Compliance-as-Code for AI Infrastructure 279
Integration of PQC Within Zero Trust Data Protection Models 279
Data Governance, Configuration Management, and Lineage 280
AI-Ready Data Dictionary for Zero Trust Configuration Management 281
Integration Guidance 282
AI Bills of Materials for Transparency, Auditability, and Accountability 286
Model Lineage and Dependency Mapping for Explainability and Forensic Readiness 286
Version Control, Rollback, and Governance Across Distributed AI Systems 287
Lifecycle Governance for Agentic and Embedded AI Deployments 287

Business and Operational Impact of AI Privacy Failures 288
Consequences of Data Compromise in Autonomous and Embedded AI Systems 289
Regulatory Penalties, Contractual Risk, and Loss of Market Trust 289
Case Studies: Leakage from Generative AI Platforms and Autonomous Decision Engines 291
Financial and Operational Disruption from Model Exfiltration or Training Data Theft 291
Long-Term Strategic and Reputational Impacts on Global Enterprises 292
Final Recommendations 292
Data as a Continuously Governed Enterprise Asset 293
Agentic and Embedded AI as Critical Expansion Points for Privacy Risk 293
PQC Readiness as a Foundation for Future-Proof AI Data Protection 294
Key Zero Trust Enablers for AI Privacy 294
Strategic Guidance for Maintaining Resilience, Compliance, and Stakeholder Trust 295
Summary 295
Key Terms 296
End-of-Chapter Questions and Answers 296
Questions 296
Answers 299
References 300

Chapter 8 Third-Party AI Risk 303

Third-Party Risk Questionnaire Limitations for Assessing AI Risk 304
What to Know Before Assessing Third-Party AI Risk 305
Data Protection and General Information Security 305
Data Protection and Mobile Security 306
Human Resources Management 306
Asset Management and Media Handling 307
Access Management and User Controls 307
Cryptographic Controls 308
Physical and Environmental Security 308
Operations and Network Security 308
Application Security/Development 309
Supplier Relationships/Vendor Management 309
Incident Management 309

Business Continuity and Disaster Recovery 310
Governance and Compliance 310
Cloud Security 310
Data Center Operations 311
Offshore Delivery Center Controls 311
Continuous Monitoring of Vendor Performance 311
AI Third-Party Risk Assessments 313
Securing the AI Supply Chain Through Zero Trust and Cryptographic Resilience 314
Securing Core AI Supply Chain Components 316
Zero Trust Foundations for the AI Supply Chain 317
Policy and Governance 318
Identity and Access Management 320
Vulnerability Management 321
Enforcement (Policy Enforcement Points) 322
Analytics and Continuous Monitoring 323
How to Use This Questionnaire: A Practical Guide for Executives, Architects, and Engineers 324
Questionnaire Section Summary 327
Strategic Takeaways 328
Continuous Vendor Monitoring: The AI-Specific Playbook 329
Implementation Roadmap, Turning Theory into Practice 330
AI Third-Party Supply Chain—Additional Risks 331
AI Third-Party Dependencies and Risk Amplification 331
Vector Databases in Zero Trust AI Architectures 332
Assessing Third-Party Vector Databases 333
Vector Database Third-Party Risk—Description, Analysis, and Recommendations 334
Third-Party AI Post-Quantum Cryptography Risk 334
Zero Trust for AI PQC Supply Chains 336
Harvest Now, Decrypt Later Vulnerabilities 337
Services as Code, Digitized Delivery, and Network as Code in Third-Party AI Audits 338
Third-Party Services-as-Code Description, Analysis, and Recommendations 339
Third-Party Services-as-Code Key Insights and Business Implications 340

Applying Zero Trust to Third-Party AI Services-as-Code Risks 340
Audit and Monitoring 341
Third-Party Considerations 341
Strategic Implication 341
Real-World Case Studies That Illustrate the Threat Landscape 341
Case Study: Model-Namespace-Reuse on Hugging Face 342
Lessons Learned and Mitigation Framework 342
Case Study: Dual Dependency—CrowdStrike and AWS Outages 343
Impact on AI Operations and Supply Chains 343
Zero Trust Mitigation Strategies 344
Disaster Recovery and Business Continuity for AI Ecosystems 345
The New Definition of Resilience 345
Summary 346
Key Terms 347
End-of-Chapter Questions and Answers 347
Questions 347
Answers 350
References 351

Chapter 9 Build AI-Ready Environments 353
AI-Ready Environments 354
Overview of Network Design 355
Network Design Fundamentals 356
Functional Requirements 357
Technical Requirements 358
Application Requirements 359
Design Use Cases 362
Constraints 364
Identifying Requirements Through Root-Cause Analysis 365
Network Design Principles 366
Architect's Perspective: Design for Enforceability, Not Just Connectivity 366
Implementing Engineer's Perspective: Codify the Topology and Trust Model 367
Operational Engineer's Perspective: Monitor, Respond, and Maintain Policy Integrity 368
Security as Foundational Design Constraint 368

Security-Aware Availability: Designing for Complexity 370
Engineering for Manageable Complexity 371
Architect-Focused Network Design Techniques 372
Failure Isolation 372
Shared Failure States 372
Modularity and Composable Design 373
Hierarchical Design for Scale 373
Technique Synthesis 374
Network Design Pitfalls 375
Avoiding Assumptions 375
The Trap of Overdesign 375
Misapplied Best Practices 376
Preconceived Notions 376
AI Network Design Use Cases 376
Cloud/Hybrid Solutions Based on Business-Critical Operations 377
AI/ML 377
Impacts of AI on Corporate Security Policy 378
Build Policies Using AI/ML 378
Techniques for Designing AI-Ready Environments 379
Key Components Required for an AI-Ready Environment 380
AI Workloads 382
Infrastructure Requirements 383
Compute Infrastructure 384
Storage Systems 385
Networking 386
AI Security 386
AI Use Cases 387
AI in the Cloud 387
Hybrid Cloud 388
Multi-Cloud 388
Edge Computing 388
AI at the Edge 389
AI and IoT 389
Sustainability Intersection with AI-Ready Environments 389
Sustainable Practices and Energy Efficiency with AI Workloads 390
Sourcing Matters 390

Cost Optimization Strategies 390
Cost-Benefit Analysis Framework 391
Ways to Optimize Established or Existing Data Centers to Become AI-Ready 391
Greenfield vs. Brownfield AI Environments 392
Cloud Provider Offerings 392
Organizations with Mature AI Environments 392
Businesses Still Exploring AI Solutions 392
Summary 393
Key Terms 393
End-of-Chapter Questions and Answers 394
Questions 394
Answers 397
References 398

Part III: Defending at Scale—Platform Protection, Monitoring, and Response

Chapter 10 Build and Secure Enterprise-Grade Generative AI Applications: ChatAI, RAG, MCP, Agentic AI, and Embedded AI 401

Enterprise Generative AI Applications 402
Types of Generative AI 402
GenAI in the Enterprise Landscape 406
Techniques to Bridge GenAI to Zero Trust 409
Security Considerations of ChatAI Applications 411
Enterprise Use Cases for ChatAI 411
Retrieval-Augmented Generation 412
Fine-Tuning of Generative AI Models 413
Security Considerations for ChatAI 414
Securing RAG and Fine-Tuned Generative AI Models 416
Securing User Interactions 419
Security Considerations for Agentic AI 422
The Need to Secure MCP 424
Agentic AI: Design, Architecture, and Security Considerations 425
Architectural Design for Enterprise-Grade Agentic AI 426
Identity, State, and Explainability 426
Network Isolation and Segmentation 427
Multi-Agent Environments and the MCP 427
Security Challenges Unique to Agentic AI 427

Additional Considerations for Cyber-Physical Interactions: Embedded AI 428
Case Study: Cyber-Physical Interactions—Musk Robot Army 430
Foundational Zero Trust Security for Enterprise Generative AI 431
Zero Trust Controls Mapped Across the AI System Lifecycle 436
Zero Trust Applied to the AI Lifecycle 438
Case Study: Governance and Shared Responsibility—The Deloitte AI Incident 439
Summary 440
Key Terms 441
End-of-Chapter Questions and Answers 441
Questions 441
Answers 444
References 445

Chapter 11 Monitor and Respond to AI Security Threat Vectors 447

Continuous Monitoring Strategies 448
Monitoring the AI System 448
Using AIOps vs. MLOps 453
Monitoring Changes in External Regulations 455
Monitoring Changes in the Threat Landscape 455
Using Organizational Processes 457
Implementing Real-Time Monitoring Solutions 458
Monitoring Solutions for AI Systems 458
Vulnerability Management 460
Software Bill of Materials 463
Monitoring AI Systems in the SIEM 464
Incident Response Plans 465
Business Continuity and Disaster Recovery for AI-Ready Data Centers 469
Description, Analysis, and Recommendations 469
Executive Imperatives for CIOs and CISOs 470
Foundational Architecture for AI-Ready Data Centers 471
RoCEv2: Driving Secure, Scalable Performance for Modern Infrastructures 471
AI Design Considerations 472
Core Risks Unique to AI Workloads 474
Business Continuity and Disaster Recovery for AI-Enabled Applications 474

Core BC/DR Challenges for AI-Enabled Applications 475
Ways to Build AI-Specific BC/DR Strategies 476
Summary 476
Key Terms 477
End-of-Chapter Questions and Answers 478
Questions 478
Answers 481
References 482

Chapter 12 Case Study 485

Offensive AI: A Deep Dive into How Attackers Can Use AI 486
Offensive AI as a Threat to Individuals 490
Offensive AI as a Threat to Systems 491
Offensive AI as a Threat to Society 492
An Adversary's Perspective: Cost vs. Benefit of Utilizing AI 492
Detailed Case Studies of Offensive AI Attacks 494
LLMs in the Wild 494
AI-Enhanced Malware 496
Dynamic DDOS Attack 498
"Hey Google, Remind Me to Be Phished." 499
Lessons Learned and Best Practices 502
Dual Use 502
Sophistication 502
User Awareness Trainings 503
Advanced Monitoring and Response 503
Zero Trust Principles 503
Summary 505
Key Terms 506
End-of-Chapter Questions and Answers 506
Questions 506
Answers 510
References 510

Chapter 13 Conclusion: Vigilance, Policy, Iteration, and Beyond 515

Part I: Defining Responsible AI and the Evolving AI Landscape (Chapters 1–5) 516
Addressing AI Deployment Models and Zero Trust Principles 516
Overcoming Organizational Barriers and Applying AI/ML Pipelines 517

Ethics, Governance, Zero Trust, and Security in AI Implementation 518
Future of AI: Quantum Computing and the Evolving Threat Landscape 518
Introduction to the AI Threat Landscape and Ethical Foundations 519
AI as a Target: Evasion Attacks, Data Poisoning, and Threat Modeling 519
Backdoor Attacks 520
Model Inversion and Membership Inference Attacks and the Impact on Availability 520
Supply Chain Attacks and Offensive AI 521
Introduction to Zero Trust for AI Security and Historical Context 521
Zero Trust Implementation: The Categories and Foundational Controls 523
The Power of "Know What You Have" 523
Testing, Roadmaps, Zero Trust, and Long-Term Planning 524
The Importance of Data Classification 525
Data Classification, Ethical Considerations, and Governance 526
Threat Modeling for AI, Mitigations, and AI Governance 527
AI Security and the Rundown 528
Part I Summary: Defining Responsible AI and the Evolving AI Landscape 529
Part II: Building Operational Resilience—People, Processes, and Infrastructure (Chapters 6–9) 530
Organizational Readiness and Cultural Maturity 530
Secure Data Practices and Privacy Engineering 530
Third-Party AI Risk and Shared Responsibility 531
Infrastructure for AI: Cloud, Edge, and Hybrid Environments 531
Part II Summary: Aligning Strategy to Execution 531
Part III: Defending at Scale—Platform Protection, Monitoring, and Response (Chapters 10–12) 532
Enterprise-Grade AI Security and GenAI Platforms 532
Secure Integration and API Governance 532
Monitoring and Threat Detection in AI Systems 533
Incident Response and Adaptive Resilience 533
Offensive AI: AI as an Attack Vector 533
Part III Summary: Lifecycle-Driven Defense 533
Conclusion: The Work Ahead—From Readiness to Relentless Execution 534
Continuous Adaptation: Securing the Lifecycle 535

Governance, Culture, and Organizational Readiness 535
Measuring Maturity and Progress 535
Final Guidance for Security Leaders 536
Future Threats: From Visionaries to Reality 536
Conclusion: The Road Ahead 539
References 540
Curated Reading, Film, and TV Series List on AI, Zero Trust, and Future Threats 542

Appendix Study Guide 545
Study Guide Questions 546
Category 1: Policy and Governance 546
Category 1 Answer Key 554
Category 2: Identity 556
Category 2 Answer Key 564
Category 3: Vulnerability Management 565
Category 3 Answer Key 573
Category 4: Enforcement 574
Category 4 Answer Key 583
Category 5: Analytics 584
Category 5 Answer Key 592
Closing 593

Glossary 595

Afterword 609

Final Word 613

Index 615

Icons Used in This Book


Router


Laptop


Hub


Firewall


Network Cloud


Headquarters


Workgroup Switch

Preface

Why Was It Important to Write This Book Now?

For those who have attended my sessions at Cisco Live over the years, the "clock story" may sound familiar. As a small child, I developed an early fascination with technology by disassembling my father's alarm clocks using a butter knife—reimagining them into new, sometimes questionably functional inventions. By the time I reached high school, he had five alarm clocks on his nightstand. My Radio Shack electronics kit had become part of my personal "workbench," and my father had, quite understandably, developed Zero Trust in his timekeeping devices.

What many do not know is that I was raised in the coalfields of West Virginia from the age of five. My great-grandfather, on my mother's side, lost his sight in the Kentucky mines. My grandfather lost his leg in the mines of West Virginia. Outside of my parents, they were the most influential figures in my life. From them, I learned that nothing is impossible.

In Barrett, West Virginia, we had no clean water and almost no access to technology. Even radio signals required climbing to the top of a mountain to get reception. Our only connection to the outside world was cable television—with four channels.

When I moved at age eleven to the nearby county seat of Madison in 1975, I discovered a small public library housed in an old bank building. It had four bookcases of science fiction books. I read every single one. Arthur C. Clarke, Gene Roddenberry, and Robert Heinlein became my early mentors. I also watched every science fiction film released during those formative years, each one fueling my imagination for what technology—and humanity—might someday achieve.

The first computer I ever touched was in 1980 at the Sunrise Museum in Charleston, West Virginia. It was a TRS-80, placed on a pedestal, running the AI program ELIZA—an early natural language processing system that mimicked a psychotherapist by responding to users with pre-scripted prompts. At the time, I didn't realize I was communicating with one of the world's first conversational AI systems. But in hindsight, it's remarkable: The very first computer I encountered—the only one I physically touched at that time—was already running an AI program. From the beginning, artificial intelligence was there.

And now, many decades later, as AI moves from novelty to infrastructure, from niche to necessity, the need to secure it with that same relentless curiosity and passion has become the driving force behind this book.

My early "mentors" told many cautionary and visionary tales of what AI could be—or must never become. I believe deeply in the extraordinary potential of AI to solve the unsolvable, to improve the health of millions, and to bring to life the hopeful future imagined by Gene Roddenberry. Yet the cautions shared by Arthur C. Clarke and Robert Heinlein are no longer fiction—they are here.

It will take all of us, working together, to protect humanity from emerging existential threats while realizing the immense promise of artificial intelligence.

Come on this journey with us—and see how you can be that difference.

—*Cindy Green-Ortiz*

Prologue

The first Microsoft AI red team exercise against a production system ran in late February and early March 2020. At the time, security for machine learning was still mostly academic, with only a few practical demonstrations against models used in cyber defenses. Ram Shankar Siva Kumar and I wanted to know whether compromising a model not built for security could be as consequential as evading an antimalware or spam detection model.

What we found had more impact than we anticipated.

Working with the Azure Red Team, we isolated a small, internal prediction service in Microsoft Azure. It wasn't user-facing—just a simple model that decided which physical hardware should host a user's workload, such as a virtual machine or database. At Azure scale, even small gaps in packing efficiency translate into millions of dollars in extra hardware costs, so this model mattered.

Most of the exercise involved navigating Azure's tight perimeter defenses to isolate the tiny machine learning model itself. Once inside, we reverse-engineered its behavior offline by analyzing how small changes to inputs affected recommendations. That analysis revealed we could craft requests that made the model place heavy workloads on already overburdened servers. The tweaks to normal requests looked meaningless to an ordinary user but could be deliberately generated by an attacker. The result? A "noisy neighbor" attack—one tenant's malicious workload could degrade or disrupt the performance of others. A subtle integrity flaw in a small AI component could cascade into widespread availability failures across the data center.

The lesson was clear: No entity—especially stochastic AI components—should ever have a trust-by-default status, even when operating entirely within what might traditionally be considered a system's normal internal trust boundary.

That experience crystallized a broader path that is outlined in this book: Zero Trust. While well-trodden, Zero Trust deserves a fresh look through the lens of AI. It must guide every stage of AI application development—from data collection to deployment. It must secure the AI supply chain, where fine-tuning, algorithmic backdoors, or tainted datasets introduce new kinds of risk. It must inform application design, where stochastic models will inevitably produce invalid output, but systems should not fail with them. And it must anchor agentic systems, ensuring that an AI agent never inherits the privileges of its designer, or wholly trusts unverified inputs from a third-party agentic function.

Years after that first red team exercise, language and multi-modal models have now become infused in virtually every technology startup and relied on in almost every market vertical. As my colleague Josh Saxe has put it, "AI has become load bearing." These models have been embedded in handheld hardware, launched into space, and have changed business processes and educational practices.

From a security perspective, the productivity and orchestration layers that sit on top of enterprise resources—collaborative copilots, embedded assistants, and automated orchestration services—do not merely *sit atop* existing attack surfaces. They *amplify* them.

Consider a few examples of how powerful automations made possible through agentic AI that sit atop enterprise resources can affect the triad of security failures: confidentiality, integrity, and availability.

Confidentiality. Oracle Autonomous Database delivers AI-powered capabilities with enterprise-grade security designed for confidential data. The service provides multiple layers of protection, including encryption at rest and in transit, integration with external key management systems, and fine-grained access controls.

But even Oracle recognizes that AI-powered database features require additional governance considerations beyond traditional database security. As noted in Oracle's Sovereign AI brief, large language models may inadvertently disclose or infer sensitive information from their training data even when no explicit data is returned. This inference risk means that confidential data could potentially be accessed through indirect queries that bypass conventional access controls. For example, a low-privilege user might be blocked from running a direct query like `SELECT * FROM EMPLOYEE_SALARIES`. However, that user could potentially ask the embedded AI, "What are the typical salary bands for a 'Senior Engineer II' in the 'Finance' department?" The LLM, if it has a privileged service that fetches sensitive data for the LLM, might provide a statistically accurate summary, effectively bypassing the database controls and leaking the very information it was meant to protect. So, Oracle has implemented the security model in its embedded AI so that `SELECT AI` cannot be used to bypass existing database controls or expose data to users without appropriate privileges.

Integrity. Microsoft Copilot for Microsoft 365 is deeply integrated across productivity applications, enabling users to summarize documents, draft emails, generate code, and accelerate workflows. This integration is built on Microsoft's Zero Trust security architecture with multiple layers of protection.

Microsoft acknowledges that AI assistants introduce new security considerations, including prompt injection attacks where adversarial input attempts to manipulate AI behavior. As Microsoft's security research teams have documented, prompt injection represents an evolving threat vector where malicious instructions embedded in emails, documents, or messages could potentially influence Copilot's responses or actions. This isn't just a theoretical threat. Security researchers have demonstrated how an attacker can send a simple calendar invite and, *without any user interaction*, inject malicious prompts to take control of Copilot's behavior. More recent research uncovered EchoLeak, a zero-click prompt injection exploit. A single, specially crafted email was enough to bypass Microsoft's defenses, escalate privileges, and exfiltrate data from a production Copilot system. These integrity failures show that Copilot can be turned from a passive assistant into an attacker-controlled agent, operating with full trust inside the system.

Availability. As our initial red team exercise demonstrated, a subtle integrity failure in a small ML model could cascade to system-wide cloud availability failure. But this dependency is a two-way street, and availability failures cascade in both directions. This is especially true in modern AI software solutions, where we rely more heavily on third-party cloud functions and third-party reasoning models. This creates a systemic dependency risk when the underlying cloud infrastructure fails.

On October 20, 2025, Amazon Web Services (AWS) experienced a major outage in its US-EAST-1 region, rooted in DNS resolution failures within the EC2 internal network. The disruption cascaded: Because many AI applications, backend services, and orchestration platforms depend on AWS, the outage knocked offline not only consumer-facing apps but productivity and AI services as well. Thousands of companies reported errors and downtime, including AI-powered platform Perplexity AI, ML workflows in Robinhood and Coinbase, and communications platforms Discord and Reddit, whose AI-powered mediation platforms became nonresponsive.

For organizations that deeply embed cloud-based AI into their workflows—via agents, copilots, or orchestration services—this outage represented a cascade of failures. When AI layers go offline, not only do services pause; inference pipelines, looped agent workflows, or context-rich automations may degrade, misbehave, or corrupt state. That means AI-mediated business processes are not immune to cascade failure.

So what does all of this have to do with Zero Trust? The security community has always been adept at identifying how things fail and making systems that reduce the likelihood and impact of those failures. With AI, the first job has become dramatically easier in one sense (there are now many practical adversarial techniques to study), and dramatically harder in another (the attack surface has multiplied and become adaptive). The right posture is not to chase perfect models or systems—a fool's errand—it is to design systems that do not depend on perfection. Zero Trust is precisely that posture: Assume components can fail, assume models can be tricked, and design systems that contain faults without causing catastrophe.

That first red team reminded me of how fragile our assumptions were. The work that follows in this book aims to make fewer of those assumptions. It is a practical architecture for a time when AI is both load bearing and everywhere. If you walk away with one message from these pages, let it be this: Secure AI is not primarily a feature you bolt on at deployment. It is a discipline you fold into design, data, supply chains, and day-to-day operations. Adopt it now, and you will have built the infrastructure needed to survive the inevitable failures ahead.

—Hyrum Anderson

Hyrum Anderson is Director of AI and Security at Cisco, via an acquisition of Robust Intelligence, where he served as Chief Technology Officer. Much of his career has been focused on defense and security, having directed research projects at MIT Lincoln Laboratory, Sandia National Laboratories, Mandiant, as Chief Scientist at Endgame (acquired by Elastic), and Principal Architect of Trustworthy Machine Learning at Microsoft. While at Microsoft, he was the architect for Microsoft's AI Red Team inaugural work on production AI systems as chair of the AI Red Team governing board. Hyrum co-founded the Conference on Applied Machine Learning in Information Security (CAMLIS), was appointed to serve on the 2024 National Academies study on Cyber Hard Problems and has been an advisor to the U.S. and UK governments on AI Safety and Security. He has spoken at numerous academic and industry conferences at the intersection of security and machine learning, including RSA, BlackHat, and DEFCON. He has

authored more than 60 peer-reviewed academic publications and co-authored the books *Not with a Bug, but with a Sticker: Attacks on Machine Learning Systems* and *What to Do About Them*. He received his PhD in electrical engineering from University of Washington, with an emphasis on signal processing and machine learning, and BS and MS degrees in electrical engineering from Brigham Young University.

Foreword

Pioneer Foreword

In the rapidly evolving landscape of artificial intelligence, security can no longer be an afterthought. As AI systems become increasingly integrated into our critical systems and infrastructure, financial systems, healthcare networks, and personal devices, the question is no longer *if* adversaries will target these systems, but *when*—and with what consequences.

We stand at a pivotal moment in technological history. AI models now control power grids that sustain entire cities, guide financial algorithms that move trillions of dollars daily, inform medical decisions that determine patient outcomes, and process the most intimate details of our personal lives through the devices we carry everywhere. This unprecedented level of AI penetration across our society's most sensitive domains has created a perfect storm for security vulnerabilities with cascading consequences.

This convergence of AI and cybersecurity presents unique challenges that traditional security models struggle to address. AI systems operate in dynamic environments with complex attack surfaces, making them particularly vulnerable to novel threats. They process vast quantities of sensitive data, interact with multiple systems, and make autonomous decisions that can have far-reaching impacts. In this context, the conventional perimeter-based security approach—where resources inside a network boundary are implicitly trusted—has become dangerously obsolete.

Enter Zero Trust, a security paradigm built on the principle that no entity, whether inside or outside the network perimeter, should be trusted by default. As the authors meticulously demonstrate throughout this book, applying Zero Trust principles to AI security represents not merely a tactical adjustment but a fundamental reimagining of how we build, deploy, and maintain secure AI systems.

What makes this work particularly valuable is its pragmatic approach. The authors have skillfully bridged the gap between theoretical security models and real-world implementation, providing concrete strategies that organizations can adopt regardless of their size or technical maturity. From continuous verification and least-privilege access to microsegmentation and robust authentication, each principle is examined through the specific lens of AI security challenges.

As someone who has witnessed the consequences of inadequate security measures in deployed AI systems, I cannot overstate the timeliness and importance of this contribution. The frameworks, methodologies, and case studies presented here offer a comprehensive roadmap for organizations navigating the complex intersection of AI innovation and security resilience.

Whether you're a CISO developing a security strategy for AI deployment, an AI engineer incorporating security into your design process, or a policymaker considering the broader implications of AI security, this book provides invaluable insights for the journey ahead.

In a world where AI systems increasingly influence critical aspects of our lives, securing these systems is not just a technical challenge; it's an ethical imperative. This book represents an essential step toward meeting that imperative.

—Jeff Hollar

Jeff Hollar is a principal engineer at Lumen Technologies and a pioneering technologist with more than 40 years of experience in AI-driven networking, automation, and cloud-native innovation. A former senior engineer at Cisco Systems, he led initiatives in federated learning, network virtualization, and Zero Trust automation, achieving groundbreaking advances in secure, AI-enhanced infrastructure. His patented innovations and leadership at organizations such as Pixar, Lawrence Livermore National Laboratory, and Silicon Graphics reflect a career dedicated to building resilient, intelligent systems.

Executive Foreword

AI is both the most powerful security tool we've ever had and the most dangerous attack surface we've ever created. If you're learning about AI (and you should be), the next question isn't "What can it do?" but "How do we secure it?" If you've got kids, you probably know the book *If You Give a Mouse a Cookie*, by Laura Numeroff. The concept is that if you give a mouse a cookie, one thing leads to another…well, AI is the cookie. Once you start thinking about its capabilities, you next must think about its risks. And once you think about its risks, you have to ask: "How do we build systems that can adapt as fast as the threats do?"

That's where Zero Trust comes in.

AI can be leveraged to help us see threats faster, lower our mean time to respond, and even potentially repair our defenses when something does happen. On the flip side to that, AI can also help our adversaries attack faster, more accurately, and with more effectiveness than ever before. Or what if AI is our adversary (i.e., an agentic AI) acting on behalf of an adversary? We need to be thinking about both sides of that AI coin, both in how we protect against someone using AI against us and how we protect ourselves by leveraging the power of AI.

One of the biggest challenges with AI is that what we see today is not what we're going to see tomorrow. AI is evolving so rapidly, constantly moving, constantly changing, that we all struggle to keep up with the newest advancements in the technology. Amazing! And also frightening. The fluidity of AI makes it challenging to understand how to think strategically and plan for several years down the road without the risk that what you do today is not going to be useful in the future.

To me, that is the strength of Zero Trust planning and execution. The assumption is that today I do not natively trust anything that interacts with my environment. There are tactical things that we can do now to help with that Zero Trust mantra. We can also assume that strategically, years down the road, our security stance of trusting nothing will stand the test of time. AI gets better, faster, stronger…okay. Do I trust it automatically because of that? Nope. AI will still have to be treated as any other IT asset (people, endpoint, compute, agentic AI, etc.), which is to say, not trusted until verified. Think about it from the perspective of end users. Ostensibly, we have all become more security aware, but do we also still click the sketchy links? Yes, we sure do.

This book is intended to help with Zero Trust theory and execution, regardless of whether you are protecting your environment from typical threats, whether you are using AI to help bring more visibility into your threat landscape, or whether you are shoring up defenses against AI-strengthened adversaries. The authors have spent untold hours bringing the Zero Trust journey to life, to help you leverage Zero Trust in all facets of your security journey. Here's to a good read!

—Jason Penn

Jason Penn is a senior director with Cisco's CX Americas organization, where he leads a large team of engineers and architects across a myriad of technical domains, including security, networking, data center, and collaboration. He spent more than 15 years as a security consultant and director with PricewaterhouseCoopers' Advisory organization before joining Cisco in 2015 as a security principal. Even though his current remit contains more than security, he still holds security topics near and dear to his heart.

Introduction

The goal of this book is to provide you with proven methods to secure artificial intelligence (AI) systems, tools, models, and data centers using the Zero Trust principle throughout an organization. Drawing on many years of security and architectural experience, the authors have designed AI-based solutions and data centers that incorporate Zero Trust architectural processes. This book includes observations on where organizations have been most successful, as well as discussions on common mistakes and false assumptions.

While there is significant debate in the security world regarding the effectiveness of Zero Trust and how its aspects may differ between organizations, this book aims to offer broad recommendations and guidance to assist organizations, architects, and engineers on their journey toward Zero Trust. Considerations when evaluating these assumptions and mistakes typically include an organization's business behavior, industry, and capabilities. Additional considerations include the best ways to mitigate unique organizational risks and analyses that can only be conducted within the organization. This ensures that recommendations within the Zero Trust methodology are properly aligned with an organization's specific needs.

Strategies for Securing AI Solutions, Data Centers, and Models Using Zero Trust Principles

The key to uniting the organization's teams in the Zero Trust journey that is focused on securing AI-based solutions, tools, models and infrastructure is having an executive sponsor with broad oversight across business units and any areas of the organization that may be affected. This executive sponsor should be positioned to influence the participation of the disparate teams required for the project and have direct ownership of outcomes. This role may be filled by an executive at the C-Suite level with mandates from the board of directors, a team of executive managers, or a singular senior manager with broad influence and authority.

Regardless of the person or team, due to the changes in ongoing operations, configurations, and differentiated access, the executive sponsor must have the authority to accept changes to policy and prevent access to individuals while shielding operations staff. At the same time, this executive sponsor must have the influence and connections within the business to socialize and gain buy-in from across the organization. Preparing for and driving toward the implementation of Zero Trust requires broad support and involvement from a wide range of teams within the organization, including those focused on AI solutions, data centers, and models.

With AI being embedded throughout the organization and with new "developers" being added to the development resources, due to the nature of AI, it is imperative that organizations focus on security fundamentals and define what the data is, where the data is, how important the data is, and how transparency is applied to the creation of models based on the data using AI.

In addition, programs should account for key performance indicators (KPIs) of the business, providing metrics for evaluating how the program is working and identifying what improvements will be needed to get the program off the ground. Both aspects are critical to the success of the program.

Goals and Methods

This book aims to articulate the successful strategies we have observed in the hundreds of customers we've worked with over the years who are pursuing similar Zero Trust goals, particularly in the context of AI solutions, data centers, and models. With the continuous changes occurring in the industry related to Zero Trust and its components, this book serves as a current baseline for what we believe is the most practical approach for most customers.

In writing this manuscript to guide organizations, we have applied the 80/20 rule. The goal is that the methodologies within this book will assist 80 percent of customers in working toward their Zero Trust goals with minimal variation on the methods stated here. For the 20 percent of customers who have already advanced significantly in their pursuit of Zero Trust, rendering much of this text confirming and foundational to their goals, the hope is that this book might serve as a reference model for operations and engineering, specifically for how to continue to improve or operate the Zero Trust Architecture.

Many will notice that broader concepts are used throughout the text, with some avoidance of stating singular, definitive technologies that must be present to accomplish a goal or milestone. This approach is intentional. As Zero Trust evolves—and it continues to do so every day—products will change, but their functionality and business-aligned goals will remain the same. This is the pattern we've observed over many years in the industry, and we hope to continue observing it for many more years to come.

Who Should Read This Book?

This book is intended for network and cybersecurity engineers, data scientists, architects, and leaders, as well as professionals involved in AI solutions, data centers, and models. The primary audience includes those responsible for creating a framework based on principles that ensure monitored and managed least-privilege access security controls to remediate and mitigate advanced cybersecurity threats. The secondary audience includes other networking staff members and AI/data center professionals interested in maturing least-privilege cybersecurity access strategies in relation to their specific corporate business environments.

This book is designed for intermediate to advanced readers. Due to the methods explored in the content, industry experts in network security, AI solutions, and data center management could also find this book a valuable reference.

How This Book Is Organized

The initial chapters provide a foundational understanding of AI security, setting the stage for the rest of the book. We discuss the evolution of AI and its security challenges and the historical development of AI and the emerging security challenges that have accompanied its evolution.

Part I: Defining Responsible AI and the Evolving AI Landscape

Chapter 1, "Overview": This chapter establishes foundational concepts of AI security, discussing how Zero Trust principles represent a paradigm shift in securing AI. It highlights the significance of Zero Trust in AI security. This chapter also defines the structure of the book by providing a roadmap of the book's content, outlining its organization to help you navigate the chapters.

Chapter 2, "Responsible AI and Integrated Awareness": This chapter defines responsible AI and outlines its core principles, emphasizing the necessity of ethical AI development and deployment. It underscores that ethical considerations are not only morally imperative but also crucial for understanding the consequences of neglecting them. This chapter delves into ethical considerations and bias mitigation strategies to secure AI systems, ensuring fairness and transparency. It also discusses methods to guarantee that AI systems remain fair and transparent. Additionally, this chapter highlights the importance of building organizational awareness and introduces a maturity model for AI awareness based on industry standards. It integrates security awareness into AI development, stressing the critical role of incorporating security measures in the AI development process.

Chapter 3, "Artificial Intelligence Threat Landscape": This chapter offers a comprehensive overview of the various threats to AI systems, detailing common attack vectors. It identifies and explains these vectors, providing a clear understanding of how AI systems can be targeted. Additionally, the chapter introduces new case studies of AI security events, presenting real-world incidents and the lessons learned from these occurrences.

Chapter 4, "Zero Trust Principles and Methods": This chapter explains the core principles of Zero Trust security and offers a comprehensive review of Zero Trust security capabilities. It delves into the strategies for implementing Zero Trust in AI, presenting practical approaches for applying these principles to AI systems. Additionally, the chapter highlights the benefits of Zero Trust for AI security and outlines potential methods for developing a proof of value (POV).

Chapter 5, "Securing AI from the Start": This chapter emphasizes the importance of classifying data early in the AI development process as crucial for ensuring the accuracy and efficiency of AI models. Early data classification helps in identifying relevant datasets, reducing noise, and enhancing the quality of input data, which in turn leads to more reliable and precise AI outcomes. The chapter discusses the need of securing AI from the start, offering significant business and operational advantages. By integrating security measures early in the AI development lifecycle, organizations can mitigate risks associated with data breaches and ensure compliance with regulatory standards. This proactive approach not only protects sensitive information but also builds trust with stakeholders,

enhances operational efficiency, and supports the seamless deployment of AI solutions, ultimately driving innovation and competitive advantage.

Part II: Building Operational Resilience—People, Processes, and Infrastructure

Chapter 6, "Organizational AI Security Readiness": This chapter provides a comprehensive guide to preparing an organization for AI security. It starts by detailing methods for assessing the organization's readiness to adopt AI security measures. The discussion then shifts to strategies for fostering a security-first culture, emphasizing the integration of security into the company's core values and daily operations. Finally, the chapter underscores the importance of training and awareness, advocating for continuous education and engagement to ensure that all employees are knowledgeable about security protocols and best practices.

Chapter 7, "AI-Ready Data Privacy and Business Impact": This chapter delves into the critical considerations for data privacy in the context of AI. It covers the regulatory compliance requirements such as the NIST AI Risk Management Framework (RMF) and the EU AI Act, emphasizing their importance in maintaining data privacy. Additionally, it explores the business impact of data privacy in AI, highlighting how proper data management and compliance can enhance trust and drive business success.

Chapter 8, "Third-Party AI Risk": This chapter focuses on identifying and managing risks associated with third-party AI solutions. It discusses the importance of thorough procurement and supply chain management to mitigate these risks. The chapter also includes case studies that illustrate real-world examples of third-party AI risks, providing valuable insights into effective risk management strategies.

Chapter 9, "Build AI-Ready Environments": This chapter provides an overview of network design fundamentals and principles essential for creating AI-ready data centers. It covers various network design techniques and the specific requirements for designing data centers that support AI workloads. The chapter also addresses cloud architecture considerations for AI security, the intersection of sustainability with AI-ready data centers, and strategies for optimizing existing data centers to become AI-ready.

Part III: Defending at Scale—Platform Protection, Monitoring, and Response

Chapter 10, "Build and Secure Enterprise-Grade Generative AI Applications: ChatAI, RAG, MCP, Agentic AI, and Embedded AI": This chapter outlines the encryption methods necessary for securing vector databases and the security considerations for implementing retrieval-augmented generation (RAG). It also covers the security aspects of building ChatAI environments and developing GenAI platforms, ensuring that these advanced AI systems are protected against potential threats.

Chapter 11, "Monitor and Respond to AI Security Threat Vectors": This chapter emphasizes the importance of continuous monitoring strategies to detect and respond to AI security threats. It details the development of incident response plans and the necessity of business continuity and disaster recovery plans for both AI-ready data centers and AI-enabled applications, ensuring resilience against disruptions.

Chapter 12, "Case Study": This chapter presents an overview of how AI manifests itself as a dual-use technology. While AI can be used for defensive purposes, attackers can also use AI to drive new forms of attacks. This chapter explores the wide array of AI-driven attacks that may increase the speed, scope, scale, and/or sophistication of attacks compared to traditional ones.

Chapter 13, "Conclusion: Vigilance, Policy, Iteration, and Beyond": This chapter underscores the importance of continuous vigilance, policy iteration, and adaptability in AI security. It highlights the need to stay ahead of emerging threats, regularly review and update security measures, and involve stakeholders in policy development. The chapter also discusses adapting to AI regulatory requirements and new technologies, embracing innovation in security, and preparing a roadmap for future AI security challenges.

Appendix, "Study Guide": This appendix serves as a study guide, offering key concepts and review questions to reinforce the material covered in the book. It includes practice tests with answers to help you assess your understanding and prepare for real-world application of AI security principles.

Afterword, by Christopher Mula: In this afterword, Chris Mula provides closing thoughts and reflections on the themes discussed throughout the book. It serves as a final commentary on the importance of AI security and the ongoing efforts required to maintain it.

Final Word, by Daryl Thomas: In this closing reflection, Daryl Thomas challenges readers to move from reactive security postures to resilient, purpose-driven AI enablement. His final word emphasizes that AI security is not just a technical exercise but a cultural and strategic imperative. As AI becomes both an opportunity and a battlefield, Thomas urges organizations to operationalize trust at scale—because resilience, not speed, will define secure innovation.

Figure Credits

Cover image: Kjetil Kolbjornsrud/Shutterstock

Chapter 6

Organizational AI Security Readiness

The integration of artificial intelligence into organizational structures presents unprecedented opportunities for innovation and efficiency. However, this technological advancement also introduces novel security challenges that demand a proactive and comprehensive approach. Organizational **AI security readiness** is the state of preparedness that enables an organization to securely develop, deploy, and manage AI systems, ensuring their integrity, confidentiality, and availability. This readiness encompasses a multi-faceted strategy that includes assessing the organization's current capabilities, building a security-first culture, implementing tailored training programs, and establishing robust governance frameworks. Without a strong focus on security readiness, organizations risk exposing themselves to a myriad of threats, including data breaches, algorithmic manipulation, and reputational damage. In Chapter 5, "Securing AI from the Start," we covered the conceptual project-level readiness. In this chapter, we take that conceptual aspect into enterprise-level maturity.

The essence of organizational AI security readiness lies in understanding that AI systems are not inherently secure. They are complex entities that require careful planning and execution to protect against potential vulnerabilities. A well-prepared organization recognizes the importance of integrating security into every stage of the AI lifecycle, from initial design and development to deployment and ongoing maintenance. This proactive approach ensures that security considerations are not an afterthought but are integral to the AI system's architecture.

Achieving comprehensive AI security readiness requires a collaborative effort across various departments, including IT, security, data science, and legal. Each department brings unique expertise and perspectives that are essential for identifying and mitigating potential risks. Effective communication and coordination among these teams are crucial for ensuring a unified and coherent security posture.

Assessing Organizational Readiness

Assessing organizational readiness is the foundational step toward ensuring robust AI security. It involves a thorough evaluation of the organization's existing capabilities, policies, and practices to identify strengths, weaknesses, and areas for improvement. This assessment provides a clear understanding of the current state and establishes a baseline against which progress can be measured. The assessment should cover various aspects, including stakeholder engagement, security readiness, current capabilities, risk assessment, gap analysis, compliance, technical infrastructure, and organizational culture.

Stakeholder Engagement and Ownership

Effective AI security requires buy-in and participation from stakeholders across the organization. Identifying key stakeholders from different departments is crucial for ensuring that all perspectives are considered. These stakeholders include executives, IT professionals, data scientists, legal counsel, and end users. Each stakeholder brings unique insights and responsibilities that contribute to a holistic security strategy. For example, executives can provide strategic direction and resources, whereas data scientists can identify potential vulnerabilities in AI models.

Assessing current levels of security awareness and involvement among stakeholders is essential for understanding the existing security culture. This understanding can be achieved through surveys, interviews, and workshops. The goal is to determine how well stakeholders understand the risks associated with AI and their roles in mitigating those risks. Based on this assessment, organizations can develop targeted training and awareness programs to address specific knowledge gaps.

Clarifying roles and responsibilities in AI security readiness ensures that everyone knows what is expected of them. This effort includes defining who is responsible for implementing security controls, monitoring AI systems, and responding to security incidents. Clear roles and responsibilities promote accountability and prevent confusion, leading to a more effective security posture. For instance, the IT department may be responsible for securing the infrastructure, while the data science team is responsible for ensuring the integrity of the data used to train AI models. Please reference Chapter 4, "Zero Trust Principles and Methods," for the role-based accountability model that aligns with this section's governance approach.

Security Readiness Assessments

Conducting a **security readiness assessment** is a critical step in evaluating an organization's ability to securely develop, deploy, and manage AI systems. This assessment involves a comprehensive review of existing security policies, procedures, and controls to identify potential vulnerabilities and gaps. The assessment should cover various aspects, including data security, access controls, incident response, and compliance with relevant regulations.

One effective method for conducting a security readiness assessment is to use a standardized framework, such as the NIST Cybersecurity Framework or the ISO 27001 standard. These frameworks provide a structured approach to evaluating security practices and identifying areas for improvement. The assessment should also include a review of the organization's AI-specific security measures, such as those related to data privacy, algorithmic bias, and model security.

The security readiness assessment should be conducted by a team of experts with diverse backgrounds, including IT security professionals, data scientists, and legal counsel. This team can bring different perspectives and expertise to the assessment, ensuring a comprehensive evaluation. The assessment should also involve interviews with key stakeholders to gather insights into their understanding of security risks and their roles in mitigating those risks.

Baseline of Current Capabilities

Establishing a baseline of current capabilities is essential for understanding an organization's starting point in AI security. This baseline provides a benchmark against which progress can be measured and helps identify areas where resources should be focused. The baseline should include an inventory of existing AI systems and security measures; documentation of existing policies, procedures, and controls; and an evaluation of the organization's maturity level in AI security. An example of a baseline, also known as a *current state*, is shown in Figure 6-1. This example follows the Department of Defense (DoD) Zero Trust Architecture framework.

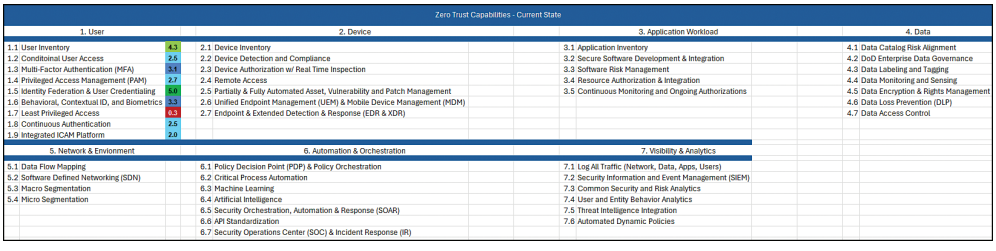


Figure 6-1 *Baseline and Current State Example Using the DoD ZTA Framework*

An inventory of existing AI systems and security measures provides a clear picture of the AI landscape within the organization. This inventory should include details about each AI system, such as its purpose, architecture, data sources, and security controls. It should also identify any existing security measures that are in place, such as access controls, encryption, and monitoring systems. This inventory helps identify potential gaps in security coverage and prioritize areas for improvement. AI-specific metrics, such as model inventory completeness, dataset lineage tracking, and ML pipeline coverage, should be included as part of the baseline criteria.

Documentation of existing policies, procedures, and controls is crucial for ensuring that security practices are well defined and consistently applied. This documentation should

include details about the organization's security policies, incident response procedures, and data privacy controls. It should also include any specific procedures for securing AI systems, such as those related to data validation, model testing, and vulnerability management. This documentation helps ensure that everyone understands their roles and responsibilities in maintaining AI security.

Evaluating the organization's maturity level in AI security provides a high-level assessment of its overall security posture. This evaluation should consider the readiness of *people, process,* and *technology*. For example, the organization's security culture (people), governance framework (process), and technical capabilities (technology) should all be part of the evaluation of the organization's readiness and maturity level of its AI security landscape. A maturity model can be used to assess the organization's progress along a continuum from ad hoc security practices to a mature, well-defined security program. This evaluation helps identify areas where the organization needs to improve its security practices and prioritize resources accordingly.

Risk Assessment and Prioritization

Risk assessment and prioritization are essential for identifying and mitigating potential threats to AI systems. This process involves identifying AI-specific risks, assessing their potential impact and likelihood, and developing risk mitigation strategies. The goal is to focus resources on the most critical risks and ensure that appropriate security controls are in place to protect AI systems.

Conducting AI-specific risk assessments requires a deep understanding of the unique vulnerabilities and threats associated with AI systems. These threats include risks related to data privacy, algorithmic bias, model security, and adversarial attacks. The risk assessment should consider the potential impact of each risk, such as financial loss, reputational damage, or regulatory penalties. It should also consider the likelihood of each risk occurring, based on factors such as the organization's security controls and the threat landscape.

Prioritizing risks based on their potential impact and likelihood ensures that resources are focused on the most critical threats. Prioritization can be achieved using a **risk matrix**, which plots risks based on their impact and likelihood. Risks with high impact and high likelihood should be prioritized for immediate mitigation, while risks with low impact and low likelihood can be addressed later. This prioritization helps ensure that the organization's security efforts are aligned with its overall risk tolerance. Figure 6-2 shows a risk matrix that can be leveraged to help prioritize what an organization should take care of first.

Developing risk mitigation strategies involves implementing security controls to reduce the likelihood or impact of identified risks. This effort can include implementing access controls, encryption, monitoring systems, and incident response procedures. It can also include developing AI-specific security measures, such as those related to data validation, model testing, and vulnerability management. The mitigation strategies should be tailored to the specific risks and the organization's overall security posture.

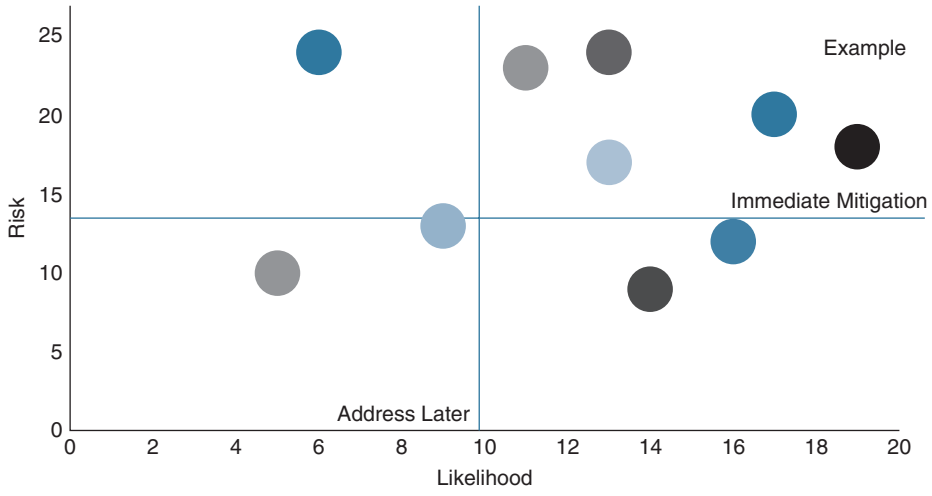


Figure 6-2 *Sample Risk Matrix*

Gap Analysis and Areas for Improvement

Identifying gaps and areas for improvement is a critical step in enhancing an organization's AI security posture. This process involves comparing the organization's current security practices against industry best practices, regulatory requirements, and internal policies. The goal is to identify any areas where the organization is falling short and develop a plan to address those gaps. It is recommended to integrate automated compliance and risk tracking capabilities (governance, risk, and compliance, or GRC platforms) to assist with the operationalization of the corresponding gaps that have been found.

A **gap analysis** should cover various aspects of AI security, including data security, access controls, incident response, and compliance with relevant regulations. It should also consider AI-specific security measures, such as those related to data privacy, algorithmic bias, and model security. The gap analysis should involve a thorough review of existing policies, procedures, and controls, as well as interviews with key stakeholders.

Areas for improvement should be prioritized based on their potential impact on AI security. This prioritization can be achieved using a risk-based approach, which focuses on addressing the most critical vulnerabilities and threats first. The improvement plan should include specific, measurable, achievable, relevant, and time-bound (SMART) goals. It should also assign responsibility for implementing the improvements and track progress against the goals.

Regularly reviewing and updating the gap analysis and improvement plan are essential for maintaining a strong AI security posture. This effort ensures that the organization stays abreast of emerging threats and vulnerabilities and continuously improves its security practices. The review process should involve key stakeholders and should consider feedback from security audits, incident response exercises, and industry best practices.

Compliance and Regulatory Readiness

Compliance and **regulatory readiness** are critical aspects of organizational AI security. Ensuring adherence to relevant laws, standards, and industry guidelines is essential for avoiding legal penalties, maintaining customer trust, and protecting the organization's reputation. Ensuring adherence requires a thorough understanding of the regulatory landscape and the implementation of appropriate security controls.

Adhering to these laws and standards involves identifying the specific regulations that apply to the organization's AI systems. These laws and standards can include data privacy laws such as GDPR and CCPA; industry standards such as HIPAA and PCI DSS; and AI-specific regulations such as the EU AI Act, DORA, and NIS2. The organization should develop a compliance framework that outlines the steps necessary to meet these requirements. A sample framework was highlighted in Chapter 5.

Assessing compliance maturity and gaps involves evaluating the organization's current compliance practices against the requirements of the relevant laws, standards, and guidelines. This evaluation can be achieved through self-assessments, internal audits, and external audits. The assessment should identify any areas where the organization is not in compliance and develop a plan to address those gaps.

Preparing for audits and certifications is essential for demonstrating compliance to regulators, customers, and other stakeholders. This effort involves gathering the necessary documentation, implementing appropriate security controls, and conducting internal audits to ensure that the organization is ready for an external audit. The organization should also consider obtaining relevant certifications, such as ISO 27001 or SOC 2, to demonstrate its commitment to security.

The Cybersecurity Maturity Model Certification (CMMC) is a framework developed by the U.S. Department of Defense to ensure that its contractors and subcontractors adequately protect sensitive unclassified information, specifically controlled unclassified information (CUI). Unlike previous self-certification models, CMMC requires independent third-party assessments to verify a contractor's cybersecurity maturity level. This rigorous certification process aims to reduce the risk of data breaches and cyberattacks that could compromise national security. CMMC covers a range of cybersecurity practices and processes, spanning from basic cyber hygiene to advanced security measures, providing a comprehensive baseline for organizations to secure their information systems.

CMMC 1.0 categorizes cybersecurity maturity into five distinct levels, each representing an increasing degree of sophistication and robustness in an organization's security posture. Although we're referencing CMMC 1.0 here, CMMC 2.0 is the current active version.

- Level 1 focuses on basic cyber hygiene, requiring organizations to implement a minimum set of security controls to protect federal contract information (FCI). As organizations progress through the levels, they must demonstrate increasingly sophisticated practices and processes.

- Level 5, the highest level, requires organizations to implement advanced security practices and optimize their security processes for continuous improvement. The specific level required for certification depends on the type of information an organization handles and the sensitivity of that information.

Figure 6-3 shows how to take a framework, like CMMC, and develop a maturity scorecard. Once created, this scorecard can then be leveraged to measure the current state to retrieve a baseline, a future state to understand where an organization needs to be, and a gap analysis between the two.



Figure 6-3 *CMMC Scorecard Levels for Grading Example*

Although CMMC was initially designed to protect CUI within the defense industrial base, its framework and principles can be effectively leveraged to enhance an organization's readiness for AI and ensure the security of its AI systems. AI systems often rely on vast amounts of data, some of which may be considered CUI or other sensitive information. By implementing CMMC-aligned security controls, organizations can protect this data from unauthorized access, modification, or disclosure. This capability is particularly important for AI models that are trained on sensitive data because a data breach could compromise the model's integrity and potentially expose sensitive information.

Furthermore, CMMC's emphasis on process maturity can help organizations establish robust security practices for AI system development and deployment. By incorporating security considerations into every stage of the AI lifecycle, organizations can proactively identify and mitigate potential vulnerabilities. This effort includes implementing secure coding practices, conducting regular security assessments, and establishing incident response plans specifically tailored to AI systems. CMMC's requirement for continuous monitoring and improvement also ensures that organizations stay abreast of emerging threats and adapt their security measures accordingly.

By aligning their AI security practices with the CMMC framework, organizations can demonstrate a commitment to protecting sensitive information and ensuring the security of their AI systems. This commitment can not only enhance their competitive advantage but also build trust with customers, partners, and stakeholders. Moreover, a CMMC-aligned approach to AI security can help organizations comply with relevant regulations and industry standards, further strengthening their overall security posture. As AI becomes increasingly integrated into various sectors, leveraging CMMC principles can provide a solid foundation for building secure and resilient AI systems.

Technical Infrastructure and Tooling Evaluation

The security of AI systems heavily relies on the underlying technical infrastructure and the tooling used to manage and protect them. A thorough evaluation of these components is essential for identifying vulnerabilities and ensuring that appropriate security measures are in place. This evaluation should cover various aspects, including security tools, monitoring systems, incident response capabilities, AI model security measures, and data security and privacy controls.

Assessing security tools, monitoring systems, and incident response capabilities involves evaluating the effectiveness of the organization's existing security tools, such as firewalls, intrusion detection systems, and antivirus software. The assessment should also evaluate the organization's monitoring systems, which are used to detect and respond to security incidents. The organization should review the incident response capabilities to ensure that it can effectively respond to security incidents involving AI systems.

Evaluating AI model security measures involves assessing the security of the AI models themselves. This task includes evaluating the models' resilience to adversarial attacks, their explainability, and their potential for bias. *Adversarial robustness* refers to the ability of a model to resist attacks that are designed to fool it. *Model explainability*

refers to the ability to understand how a model makes decisions. *Bias* refers to the potential for a model to discriminate against certain groups of people.

Reviewing data security and privacy controls involves evaluating the security of the data used to train and operate AI models. This review includes assessing the organization's data encryption practices, access controls, and data loss prevention measures. The review should also consider the organization's compliance with data privacy regulations, such as GDPR and CCPA.

Culture and Awareness Readiness

A security-conscious culture is a critical component of organizational AI security. This component involves fostering a culture where security is valued and everyone understands their role in protecting AI systems. Assessing culture and awareness readiness involves measuring organizational culture regarding security and AI ethics, identifying training and awareness gaps, and planning for cultural change initiatives.

Measuring organizational culture regarding security and AI ethics involves assessing the extent to which security is valued within the organization. This assessment can be achieved through surveys, interviews, and focus groups. The assessment should also evaluate the organization's commitment to AI ethics, which includes issues such as fairness, transparency, and accountability. The results of this assessment can be used to identify areas where the organization needs to improve its security culture. In Chapter 11, "Monitor and Respond to AI Security Threat Vectors," we will illustrate how awareness (which is covered in this chapter) translates into incident responsiveness.

Identifying training and awareness gaps involves assessing the knowledge and skills of employees regarding AI security. This effort can be achieved through training assessments, quizzes, and hands-on exercises. The assessment should identify any gaps in knowledge or skills and develop a plan to address those gaps.

Planning for cultural change initiatives involves developing a strategy to foster a security-conscious culture within the organization. This effort can include providing security awareness training, promoting security best practices, and recognizing employees who demonstrate a commitment to security. The cultural change initiatives should be tailored to the specific needs of the organization and should be integrated into the organization's overall security strategy.

Incident Response and Recovery Preparedness

Incident response and recovery preparedness are crucial for minimizing the impact of security incidents involving AI systems. This process involves reviewing existing incident response plans, testing response procedures, and evaluating disaster recovery plans. The goal is to ensure that the organization can effectively respond to and recover from security incidents involving AI systems.

Reviewing existing incident response plans involves evaluating the organization's current incident response plans to ensure that they address the unique challenges of AI security

incidents. This task includes considering the potential impact of AI security incidents on data privacy, algorithmic bias, and model security. The review should also ensure that the incident response plans are aligned with the organization's overall security strategy.

Testing incident response procedures involves conducting simulated security incidents to test their effectiveness. This testing can include tabletop exercises, where incident response teams walk through the steps they would take to respond to a security incident. It can also include live exercises, where incident response teams respond to a simulated security incident in real time.

Evaluating the organization's disaster recovery and business continuity plans involves ensuring that they address the potential impact of AI security incidents. This evaluation includes considering the potential for AI systems to be unavailable due to a security incident and developing plans to restore those systems as quickly as possible. The evaluation should also consider the potential impact of a security incident on data privacy and the need to develop plans to protect sensitive data during a disaster recovery. Simulation exercises should be conducted regularly; MITRE ATLAS and OWASP GenAI have scenario templates that can be referenced and followed if needed.

Actionability and Roadmap Development

The assessment process culminates in the development of an actionable roadmap that outlines the steps necessary to enhance organizational AI security readiness. Developing a structured roadmap for advancing security maturity involves creating a detailed plan that outlines the steps necessary to achieve the desired level of security maturity. This roadmap should be based on the findings of the assessment and should prioritize the most critical areas for improvement. The roadmap should include SMART goals and should assign responsibility for implementing the improvements. Additionally, key performance indicators (KPIs) should be determined for each goal to measure its corresponding success.

Defining short-term and long-term goals involves setting specific goals for AI security readiness. Short-term goals should focus on addressing the most critical vulnerabilities and threats, while long-term goals should focus on building a robust and sustainable AI security program. The goals should be aligned with the organization's overall security strategy and should be regularly reviewed and updated.

Establishing milestones, success criteria, and KPIs enables the organization to track progress toward the goals. The milestones should be measurable and should be tied to specific outcomes. Success criteria should be defined for each milestone to provide a clear indication of whether the milestone has been achieved.

Building a Security-First Culture

Building a **security-first culture** is paramount to creating a robust and resilient AI security posture. A security-first culture is one in which security is valued at all levels of the organization, and everyone understands their role in protecting AI systems.

Building this culture requires a shift in mindset, from viewing security as a compliance requirement to seeing it as an integral part of the organization's mission.

Creating a security-first mindset involves fostering an environment where security is a shared responsibility and everyone is committed to protecting AI systems. This effort requires a cultural transformation that starts at the top and permeates throughout the organization. The importance of a security-first mindset in AI organizations cannot be overstated, because it is essential for mitigating the risks associated with AI systems.

The impact of organizational culture on AI security posture is significant. A strong security culture can help prevent security incidents by promoting security awareness, encouraging employees to report vulnerabilities, and fostering a sense of shared responsibility. Conversely, a weak security culture can increase the risk of security incidents by leading to complacency, neglect, and a lack of accountability.

A security-first mindset involves promoting security awareness, encouraging employees to report vulnerabilities, and fostering a sense of shared responsibility. It also involves creating a culture of continuous improvement, where security practices are regularly reviewed and updated to address emerging threats. Creating this culture requires a commitment from leadership to prioritize security and provide the resources necessary to support security initiatives.

Leadership and Commitment

Leadership and commitment are critical for building a security-first culture. The role of executive sponsorship in security initiatives cannot be overstated, because it provides the necessary resources, support, and direction for security programs. When executives champion security, their support sends a clear message that security is a priority and that everyone is expected to take it seriously. This leadership accountability is reinforced in Chapter 5 in greater detail.

Setting security as a core value across all levels involves integrating security into the organization's mission, vision, and values. This means that security considerations are considered in all decision-making processes, from the design of AI systems to the hiring of employees. It also means that security performance is measured and rewarded, and that security incidents are treated as learning opportunities.

Executive sponsorship involves actively participating in security initiatives, communicating the importance of security to employees, and providing the resources necessary to support security programs. This sponsorship can include funding security training, investing in security technologies, and establishing security metrics and goals. When executives are actively involved in security, their support sends a clear message that security is a priority for the overall organization.

Security Policies and Governance

Security policies and governance provide the framework for managing AI security risks. Establishing clear AI security policies is essential for defining the organization's security

expectations and requirements. These policies should cover various aspects of AI security, including data privacy, algorithmic bias, model security, and incident response.

Aligning policies with compliance frameworks and regulations ensures that the organization is meeting its legal and regulatory obligations. This effort involves identifying the relevant laws, standards, and guidelines and incorporating them into the organization's security policies. The organization should also establish a process for monitoring changes in the regulatory landscape and updating its policies accordingly.

Defining roles and responsibilities ensures that everyone knows their role in protecting AI systems. This task includes identifying who is responsible for implementing security controls, monitoring AI systems, and responding to security incidents. Clear roles and responsibilities promote accountability and prevent confusion, leading to a more effective security posture.

Risk Management and Accountability

Risk management and accountability are essential for mitigating AI security risks. Incorporating AI-specific security risk assessments into the organization's risk management program is crucial for identifying and addressing potential vulnerabilities. This task involves conducting regular risk assessments to identify, assess, and prioritize AI-specific risks.

Creating accountability structures involves assigning responsibility for managing AI security risks to specific individuals or teams. Having this responsibility ensures that there is clear ownership of security risks and that someone is accountable for mitigating those risks. Accountability structures can include security committees, risk management teams, and security champions. An example of an accountability structure includes AI-specific risk governance committees that leverage frameworks and structures found in ISO 27005 and NIST AI RMF.

Promoting transparency and reporting involves encouraging employees to report security incidents and vulnerabilities. This task requires creating a culture of trust and openness, where employees feel comfortable reporting security concerns without fear of reprisal. The organization should also establish a process for investigating and responding to security incidents and communicating the results to stakeholders.

Integration of Security in AI Lifecycle

Integrating security into the AI lifecycle is critical for ensuring that security is considered throughout the development, deployment, and maintenance of AI systems. Embedding security from design (DevSecOps) involves incorporating security practices into the software development lifecycle from the beginning. This effort includes conducting security reviews, performing penetration testing, and implementing security controls during the design and development phases.

Continuous monitoring and improvement involve regularly monitoring AI systems for security vulnerabilities and implementing improvements to address those vulnerabilities.

This task requires establishing a process for monitoring AI systems, analyzing security logs, and conducting regular security audits. The organization should also establish a process for updating and patching AI systems to address emerging threats.

Encouraging cross-department collaboration involves fostering communication and collaboration between different departments, such as IT, security, data science, and legal. This collaboration ensures that security is considered from multiple perspectives and that everyone is working together to protect AI systems. Collaboration can include regular meetings, joint training exercises, and shared security goals.

Cultural Change Strategies

Cultural change strategies are essential for building a security-first culture. Promoting open communication about security issues involves creating an environment where employees feel comfortable discussing security concerns and sharing security information. This effort requires establishing clear communication channels, providing training on security topics, and encouraging employees to ask questions and raise concerns.

Recognizing and rewarding security-conscious behavior involve acknowledging and rewarding employees who demonstrate a commitment to security. These rewards can include bonuses, awards, or promotions to employees who go above and beyond to protect AI systems. They can also include recognizing employees who report security incidents or vulnerabilities.

Fostering a proactive security mindset involves encouraging employees to constantly look for ways to improve security and be willing to take action to prevent security incidents. This approach can be achieved through training, awareness campaigns, and by empowering employees to make security decisions.

Training and Awareness Programs

Training and awareness programs are essential for ensuring that employees have the knowledge and skills necessary to protect AI systems. These programs should be tailored to the specific needs of the organization and should cover various aspects of AI security, including data privacy, algorithmic bias, model security, and incident response. Referencing the NIST NICE Cybersecurity Workforce Framework and its associated categories applicable to AI like Data Analyst and Machine Learning Engineer is a great starting point for organizations that may not know where to begin with their training and awareness programs.

The role of training in strengthening AI security cannot be overstated. Training programs provide employees with the knowledge and skills necessary to protect AI systems and to respond to security incidents. Training should be tailored to the specific roles and responsibilities of employees and should cover various aspects of AI security.

The evolving threat landscape and the need for updated knowledge require that training programs are regularly updated to address emerging threats and vulnerabilities.

This means that organizations must stay abreast of the latest security trends and incorporate them into their training programs. It also means that training programs must be flexible and adaptable to address changing threats and vulnerabilities.

Training programs should be designed to be engaging and interactive. This approach can include using real-world examples, case studies, and hands-on exercises to help employees understand and apply security concepts. Training should also be delivered in a variety of formats, such as online courses, in-person workshops, and webinars.

Tailored Training Programs

Tailored training programs are essential for ensuring that employees receive the training they need to protect AI systems. Role-based security training involves providing training that is specific to the roles and responsibilities of employees. This means that developers, data scientists, managers, and other employees receive training that is relevant to their job functions.

Organizations must balance the need for technical skills training with the need for policy compliance training. Technical skills training focuses on providing employees with the technical knowledge and skills necessary to protect AI systems. Policy compliance training focuses on ensuring that employees understand and comply with the organization's security policies.

Training programs should be designed to be relevant to the specific needs of the organization. This means that the training should be tailored to the specific AI systems that the organization uses and the specific threats that the organization faces. Training should also be designed to be practical and applicable to the employees' daily work. Suggested KPIs to measure training effectiveness include completion rates and knowledge-retention scores.

Awareness Campaigns

Awareness campaigns are essential for raising awareness of AI security issues among employees. Regular security alerts and updates provide employees with timely information about emerging threats and vulnerabilities. Organizations can provide this information by sending out security alerts, publishing security newsletters, and hosting security webinars.

Case studies of AI security breaches provide real-world examples for employees. These case studies help employees understand the potential impact of security breaches and the importance of taking security seriously. Case studies should be relevant to the organization's AI systems and should highlight the lessons learned from the breaches.

Best practices for secure AI development and deployment can include providing guidance on secure coding practices, data privacy, algorithmic bias, and model security. Best practices should be tailored to the specific needs of the organization and should be integrated into the organization's security policies.

Hands-On Exercises

Hands-on exercises are essential for reinforcing security concepts and providing employees with practical experience in protecting AI systems. By creating simulated security incidents, an organization can help test the effectiveness of its incident response procedures. This effort can include tabletop exercises, where incident response teams walk through the steps they would take to respond to a security incident.

The organization can use penetration testing and vulnerability assessments to identify security vulnerabilities in its AI systems. Penetration testing involves simulating real-world attacks to test the security of AI systems. Vulnerability assessments involve scanning AI systems for known vulnerabilities.

The purpose of conducting red teaming and blue teaming exercises is to improve the organization's security posture. Red teaming involves simulating real-world attacks to test the organization's security defenses. Blue teaming involves defending the organization's AI systems against simulated attacks. It is recommended to reference NIST SP 1270 (focusing on bias management in AI) and MITRE ATLAS frameworks and include the AI red teaming exercises that these frameworks highlight in detail.

Continuous Learning

Continuous learning is essential for keeping employees up to date on the latest AI security threats and vulnerabilities. To keep pace with emerging threats and technologies, employees can attend security conferences, read security blogs, and participate in security communities.

An organization can provide employees access to webinars, workshops, and conferences on AI security topics. These programs allow employees to learn from experts and to network with other security professionals.

The organization also can encourage employees to pursue certifications and professional development opportunities in AI security. Examples for such certificates are, for example, the OffSec OSAI+ certificate, the ISACA Advanced AI Security Management (AAISM) certificate, or the CompTIA SecAI+ certificate. This continuous learning demonstrates a commitment to security and provides employees with valuable skills and knowledge that ultimately affect the organization's security posture.

The Reasons to Measure Effectiveness

Measuring effectiveness is essential for ensuring that training and awareness programs are achieving their goals. By establishing metrics and KPIs, an organization can measure the impact of its training programs. This effort can include measuring the number of employees who have completed training, the results of training assessments, and the reduction in security incidents. Leveraging the analytics and dashboards that a learning management system (LMS) provides can help automate training progress metrics.

The organization can establish feedback loops to gather feedback from employees about training programs. This feedback then can be used to improve the content, delivery, and effectiveness of training programs.

By creating a culture of continuous learning, the organization can encourage employees to learn about AI security and to share their knowledge with others. This effort can include establishing a security knowledge base, hosting security brown bag lunches, and rewarding employees who share their security knowledge.

Organizational AI Security Readiness: Challenges and Considerations

An organization will face numerous challenges when it starts to adopt and leverage AI. In the following sections, we cover controls and capabilities that would be graded in an AI readiness assessment. Each item would be given an overall score, and together they would create an AI readiness scorecard. Additionally, we include considerations an organization should take as it prepares, implements, and operationalizes its organizational AI security.

AI Model Security Readiness

Organizations that are creating, building, and training their own AI model should leverage the following AI model security topics with the corresponding considerations to help define their security readiness when it comes to AI model security.

AI model security focuses on protecting the integrity, confidentiality, and availability of AI models themselves. Unlike traditional software, AI models possess unique vulnerabilities stemming from their data-driven nature and complex architectures. Securing these models requires a multi-faceted approach that addresses threats ranging from adversarial attacks and model inversion to data poisoning and membership inference. A robust AI model security strategy is critical for preventing AI systems from being manipulated to produce incorrect or harmful outputs, exposing sensitive data, or being used for malicious purposes.

One of the most prominent threats to AI models is adversarial attacks. These attacks involve carefully crafting input data designed to fool the model into making incorrect predictions. Adversarial examples can be subtly modified images, audio clips, or text strings that are indistinguishable from benign inputs to humans but cause the AI model to misclassify them. These attacks can have severe consequences in safety-critical applications such as autonomous vehicles, where an adversarial example could cause the car to misinterpret a stop sign and lead to an accident. Defending against adversarial attacks requires employing techniques such as adversarial training, where the model is trained on both benign and adversarial examples, and input sanitization, where potentially adversarial inputs are preprocessed to remove or mitigate their harmful effects. This issue was covered in Chapter 3, “Artificial Intelligence Threat Landscape.”

Another significant concern is **model inversion**, which involves attempting to reconstruct sensitive training data from the AI model itself. By querying the model

with carefully chosen inputs and analyzing its outputs, an attacker may be able to infer information about the data used to train the model, such as individual identities or confidential attributes. This threat can be particularly problematic for AI models trained on sensitive personal data, where model inversion attacks could lead to privacy breaches. Defending against model inversion requires techniques such as differential privacy, which adds noise to the training data or model parameters to limit the amount of information that can be inferred from the model, and output sanitization, which limits the information revealed by the model's outputs.

Data poisoning attacks involve injecting malicious data into the training dataset to corrupt the AI model. This type of attack can cause the model to learn incorrect patterns or exhibit biased behavior, leading to inaccurate predictions or unfair outcomes. Data poisoning attacks can be particularly difficult to detect because the malicious data may be subtly embedded within the larger dataset. Defending against data poisoning requires implementing robust data validation techniques, such as anomaly detection and outlier removal, and verifying the provenance and integrity of the training data. It also involves monitoring the model's performance over time to detect any signs of degradation or unexpected behavior.

Membership inference attacks aim to determine whether a specific data point was used to train the AI model. This can be problematic if the data point contains sensitive information, because it could reveal that the individual was included in the training dataset, potentially exposing that person to privacy risks. Defending against membership inference requires techniques such as limiting the model's accuracy on known data points and employing regularization methods to prevent the model from overfitting to the training data. It also involves carefully controlling the information revealed about the model's training data and performance.

Data Governance and Privacy for AI Readiness

Data governance and privacy are paramount considerations in the age of AI because these systems heavily rely on data to function and often process sensitive information. Effective data governance ensures that data is handled responsibly, ethically, and in compliance with relevant regulations, while privacy protections safeguard individual rights and prevent the misuse of personal information. In the context of AI, data governance and privacy require a holistic approach that addresses the unique challenges posed by these systems, such as the potential for algorithmic bias, the difficulty of ensuring data anonymization, and the need for transparency and accountability. Figure 6-4 highlights the AI Data Governance Loop, which we explain in detail.

One of the most significant challenges in data governance for AI is ensuring data quality and integrity. AI models are only as good as the data they are trained on, so it is crucial to ensure that the data is accurate, complete, and representative of the population it is intended to serve. Poor data quality can lead to biased models, inaccurate predictions, and unfair outcomes. To address this challenge, organizations need to implement robust data validation techniques, such as data profiling, cleansing, and validation rules. They also need to establish clear data governance policies and procedures that define data ownership, data quality standards, and data access controls.

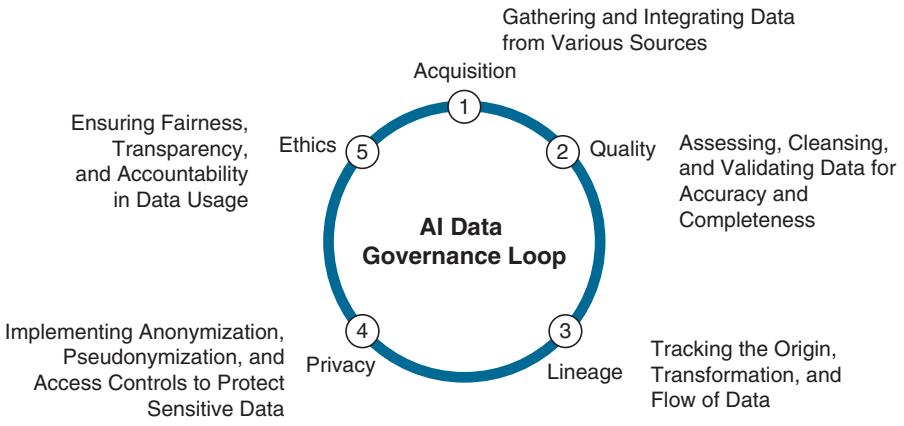


Figure 6-4 *AI Data Governance Loop*

Another critical aspect of data governance for AI is managing data lineage and provenance, which is essential to track the origin and transformation of data as it flows through the AI system to understand how it affects the model's behavior and ensure that the data is used in accordance with its intended purpose. Data lineage tracking can help identify potential data quality issues, detect unauthorized data modifications, and ensure compliance with data privacy regulations. To implement data lineage tracking effectively, organizations need to use metadata management tools, data catalogs, and data governance platforms.

Privacy is a fundamental right, and AI systems must be designed and deployed in a way that respects individual privacy. However, ensuring privacy in AI systems can be challenging due to the complex interactions between data, algorithms, and models. One of the most significant challenges is ensuring data anonymization, which involves removing or modifying personal identifiers from the data to prevent it from being linked back to individuals. Traditional anonymization techniques, such as de-identification and pseudonymization, may not be sufficient to protect privacy in AI systems because AI models can often infer sensitive information from seemingly anonymized data.

To address the challenge of data anonymization, organizations need to employ advanced privacy-enhancing technologies (PETs), such as differential privacy, federated learning, and **homomorphic encryption**. Differential privacy adds noise to limit the amount of information that can be inferred from the model, while federated learning allows AI models to be trained on decentralized data without sharing the data itself. Homomorphic encryption allows computations to be performed on encrypted data without decrypting it, enabling AI models to process sensitive data without compromising its privacy.

Transparency and accountability are also essential for data governance and privacy in AI. Organizations need to be transparent about how their AI systems work, how they use data, and what potential impacts they may have on individuals. They also need to

be accountable for ensuring that their AI systems are used responsibly, ethically, and in compliance with relevant regulations. To promote transparency and accountability, organizations need to provide clear and accessible explanations of their AI systems, establish mechanisms for addressing complaints and concerns, and conduct regular audits to ensure that their AI systems are used in accordance with their stated policies and procedures. A global governance baseline for AI management systems has recently started to emerge with ISO/IEC 42001.

AI-Specific Incident Response Readiness

AI-specific incident response procedures are critical for effectively addressing security incidents that target or involve AI systems. Traditional incident response plans often fail to account for the unique characteristics of AI systems, such as their data-driven nature, complex architectures, and potential for algorithmic bias. AI-specific incident response procedures should be tailored to address these unique challenges, enabling organizations to detect, contain, eradicate, and recover from AI-related security incidents effectively. If organizations do not know where to start with AI incidents, it is recommended that they reference NIST SP 800-61 Rev. 3 Incident Handling Guide as a starting point.

One of the most important aspects of AI-specific incident response is the ability to detect AI-related security incidents. Traditional security monitoring tools may not be sufficient to detect attacks that target AI systems because these attacks often involve subtle manipulations of data or model parameters that are difficult to identify. To detect AI-related security incidents, organizations need to implement specialized monitoring tools that can analyze AI system behavior, identify anomalies, and detect signs of adversarial attacks or data poisoning. These tools should be integrated with the organization's existing security information and event management (SIEM) system to provide a comprehensive view of the security landscape.

Another critical aspect of AI-specific incident response is the ability to contain AI-related security incidents. Traditional incident response procedures may not be effective for containing attacks that target AI systems because these attacks can often spread rapidly through the system and affect multiple components. To contain AI-related security incidents, organizations need to implement containment strategies that are specifically tailored to AI systems, such as isolating affected models, shutting down compromised data pipelines, and revoking access to sensitive data. These strategies should be implemented in a timely and coordinated manner to prevent the incident from spreading further.

Eradicating AI-related security incidents requires identifying the root cause of the incident and removing the malicious code, data, or model parameters that caused it. This process can be complex and time-consuming because AI systems often involve multiple interconnected components and dependencies. To eradicate AI-related security incidents, organizations need to have access to specialized expertise and tools that can analyze AI system behavior, identify the source of the incident, and remove the malicious components. This process may involve reverse-engineering AI models, analyzing data samples, and debugging complex code.

Recovering from AI-related security incidents involves restoring the AI system to its pre-incident state and ensuring that it is functioning correctly. This process can be challenging because AI systems often require significant time and resources to retrain and redeploy. To recover from AI-related security incidents, organizations need to have backup copies of their AI models, data, and code, as well as procedures for restoring these components in a timely and efficient manner. They also need to test the restored AI system to ensure that it is functioning correctly and that it has not been compromised by the incident.

In addition to these technical aspects, AI-specific incident response procedures should also address the ethical and legal considerations associated with AI security incidents. This effort includes notifying affected individuals, disclosing the incident to regulatory authorities, and providing remediation to individuals who have been harmed by the incident. Organizations need to have policies and procedures in place for addressing these ethical and legal considerations, as well as training their incident response teams on how to handle these issues.

Explainable AI (XAI) Readiness

Explainable AI (XAI) is a field of artificial intelligence that focuses on developing methods and techniques to make AI systems more transparent and understandable to humans. Traditional AI models, particularly deep learning models, are often considered "black boxes" due to their complex and opaque internal workings. XAI aims to overcome this limitation by providing insights into how AI models arrive at their decisions and outcomes, enabling humans to understand, trust, and effectively manage these systems. Explainability is essential for ensuring that AI systems are used responsibly, ethically, and in compliance with relevant regulations, as well as for identifying and mitigating potential biases and errors. With that said, explainability techniques may expose additional attack surfaces. A full risk assessment should be completed and reviewed before any explainability techniques are leveraged.

One of the key goals of XAI is to provide explanations that are understandable to humans. This requires translating complex black-box AI models into simpler, more interpretable representations that can be easily grasped by nonexperts. There are several different approaches to achieving this goal, including using approximations with rule-based systems, decision trees, and linear models. These interpretable methods offer a more intuitive mapping between inputs and outputs and approximate the understanding of how the complex AI system is making decisions. However, these explanatory techniques, which are in essence a surrogate model of the original black-box model, often sacrifice accuracy or performance compared to more complex models.

Another approach to XAI involves developing techniques that can explain the behavior of complex AI models without an interpretable surrogate model. These techniques typically involve analyzing the model's internal workings to identify the most important features or factors that influence its predictions. One popular technique is feature importance analysis, which assigns a score to each input feature indicating its relative contribution to the model's output. This analysis can help identify potential biases or errors in the model, as well as provide insights into the underlying relationships between the inputs and outputs.

Some well-known XAI techniques are LIME and SHAP. Local interpretable model-agnostic explanations (LIME) is an XAI technique that approximates the behavior of a complex AI model locally around a specific data point using a simpler, interpretable model. LIME generates a set of perturbed data points around the data point being explained and then trains a linear model to predict the output of the AI model on these perturbed data points. The coefficients of the linear model provide insights into the importance of each feature in the vicinity of the data point, allowing humans to understand how the AI model is making decisions in that specific context. LIME is model agnostic, meaning it can be applied to any AI model, regardless of its complexity or architecture.

SHapley Additive exPlanations (SHAP) is another XAI technique that provides a unified framework for explaining the output of any AI model based on game-theoretic principles. SHAP assigns a value to each input feature indicating its contribution to the difference between the model's output and the expected output. SHAP values are based on the Shapley value from game theory, which provides a fair and consistent way to allocate credit among the players in a cooperative game. SHAP values can be used to explain the overall behavior of the AI model, as well as to explain the predictions for individual data points.

Counterfactual explanations are an XAI technique that provides explanations by identifying the smallest changes to the input data that would result in a different outcome from the AI model. Counterfactual explanations can help humans understand why the AI model made a particular decision and what they would need to do to change the outcome. Counterfactual explanations can be generated using optimization algorithms that search for the smallest changes to the input data that satisfy the desired outcome.

Zero Trust Principles Applied to AI Security Readiness

Zero Trust is a security framework that assumes that no user or device, whether inside or outside the organization's network, should be trusted by default. Instead, every access request must be verified and authorized based on multiple factors, such as user identity, device posture, and application context. Zero Trust principles can be effectively applied to securing AI systems, providing a robust defense against a wide range of threats, including unauthorized access, data breaches, and adversarial attacks. Implementing Zero Trust controls can help organizations ensure that their AI systems are used responsibly, ethically, and in compliance with relevant regulations.

One of the core principles of Zero Trust is verifying user identities. This process involves implementing strong authentication mechanisms, such as multifactor authentication (MFA), to ensure that only authorized users can access AI systems. MFA requires users to provide multiple forms of authentication, such as a password, a biometric scan, or a one-time code, making it more difficult for attackers to compromise user accounts. In the context of AI systems, verifying user identities is particularly important for preventing unauthorized access to sensitive data, code, and models.

Another key principle of Zero Trust is limiting access to data and resources based on the principle of least privilege. This principle involves granting users only the minimum level

of access necessary to perform their job functions, preventing them from accessing data or resources that they do not need. In the context of AI systems, least-privilege access can help prevent data breaches and model theft. For example, data scientists should be granted access only to the data required to train and evaluate AI models, and operational personnel should be granted access only to the models and code required to deploy and monitor the AI systems.

Continuous monitoring and validation are also essential components of Zero Trust, as shown in detail in Chapter 4. This process involves continuously monitoring the security posture of AI systems and validating that they are functioning correctly and in compliance with security policies. Monitoring can include logging system activity, analyzing network traffic, and detecting anomalies in AI model behavior. Validation can include conducting regular security assessments, penetration testing, and red teaming exercises. In the context of AI systems, continuous monitoring and validation can help detect adversarial attacks, data poisoning attempts, and other security incidents.

Segmentation is another important Zero Trust principle that involves dividing the network into smaller, isolated segments to limit the impact of security breaches. In the context of AI systems, segmentation can be used to isolate different components of the AI system, such as the data storage, model training, and deployment environments. This can help prevent attackers from gaining access to the entire AI system if they compromise one component.

Data encryption is another crucial Zero Trust control that protects sensitive data from unauthorized access. Encryption involves converting data into an unreadable format that can be decrypted only by authorized users. In the context of AI systems, data encryption should be applied to all sensitive data at rest and in transit, including training data, model parameters, and output data.

AI Supply Chain Readiness

The supply chain in AI focuses on identifying, assessing, and mitigating the security risks associated with third-party components, datasets, and services used in AI systems. Organizations increasingly rely on external providers for various aspects of their AI systems, such as data acquisition, model development, and deployment infrastructure. However, these dependencies can introduce new security vulnerabilities that organizations must manage effectively. A robust supply chain risk management program is essential for ensuring that AI systems are secure, reliable, and compliant with relevant regulations.

One of the first steps in supply chain risk management is identifying all of the third-party components, datasets, and services used in AI systems. This step involves creating an inventory of all external dependencies, including open-source libraries, cloud-based services, and proprietary datasets. For each dependency, organizations should identify the provider, the purpose of the dependency, and the potential security risks associated with it. This inventory should be regularly updated to reflect changes in the AI system and the threat landscape.

Once the dependencies have been identified, organizations need to assess the security risks associated with each one. This assessment involves evaluating the security practices of the providers, such as their data protection policies, incident response procedures, and vulnerability management programs. Organizations should also assess the potential impact of a security breach involving the dependency, such as data loss, model corruption, or service disruption. Risk assessments should be conducted regularly to identify emerging threats and vulnerabilities.

Based on the risk assessments, organizations need to develop mitigation strategies for addressing the identified risks. This process may involve implementing contractual requirements for security, such as requiring providers to comply with industry standards or undergo security audits. It may also involve implementing technical controls, such as verifying the integrity and provenance of AI components or isolating AI systems from untrusted networks. Mitigation strategies should be tailored to the specific risks and the organization's overall security posture.

Verifying the integrity and provenance of AI components is another crucial aspect of supply chain risk management. This step involves ensuring that the components have not been tampered with or altered in any way, and that they originate from a trusted source. Organizations can use cryptographic techniques, such as digital signatures and checksums, to verify the integrity of AI components. They can also use blockchain technology to track the provenance of AI components and ensure that they have not been modified or replaced.

Establishing contractual requirements for security is essential for ensuring that providers are held accountable for protecting the security of AI systems. These requirements should include provisions for data protection, incident response, and vulnerability management. They should also include provisions for auditing the provider's security practices and terminating the contract if the provider fails to meet the security requirements.

Summary

The chapter addresses the critical need for organizations to proactively secure their artificial intelligence systems. It begins by emphasizing that AI, while offering tremendous potential, introduces unique security challenges. True readiness is defined as enabling an organization to securely develop, deploy, and manage AI, ensuring integrity, confidentiality, and availability. This process involves assessing capabilities, building a security-first culture, and implementing targeted training and strong governance frameworks.

The chapter explores AI readiness assessments in multiple contexts to include highlighting stakeholder engagement, security readiness assessments, baseline establishment, risk assessment, gap analysis, and compliance. The overall emphasis is on identifying strengths, weaknesses, and areas for improvement to establish a clear understanding of the current security posture, to then build out a roadmap to address any gaps an organization may have.

Building a security-first culture requires executive sponsorship, setting security as a core value, establishing clear security policies, and integrating security into the AI lifecycle. Promoting open communication, recognizing security-conscious behavior, and fostering a proactive mindset are identified as essential strategies for cultural change.

Training and awareness programs are deemed essential, tailored to specific organizational needs. Role-based training, awareness campaigns, hands-on exercises, and continuous learning are detailed as vital components to empower employees with the necessary knowledge and skills. Metrics and feedback loops also are necessary to ensure the training programs are effective.

Finally, this chapter emphasizes the need for ongoing adaptation, continuous improvement, and a proactive approach to stay ahead of evolving threats in the dynamic landscape of AI security.

Key Terms

AI security readiness, AI-specific incident response, awareness campaigns, continuous learning, cultural change strategies, data poisoning, gap analysis, hands-on exercises, homomorphic encryption, model inversion, regulatory readiness, risk matrix, security readiness assessment, security-first culture, tailored training programs, training and awareness programs

End-of-Chapter Questions and Answers

Questions

1. What is the primary objective of organizational AI security readiness in a Zero Trust–aligned architecture?

   A. Maximizing parallelization efficiency in GPU clusters

   B. Ensuring AI systems maintain verifiable integrity, confidentiality, and governance controls across the entire lifecycle

   C. Reducing the operational cost of AI pipelines through automation

   D. Accelerating model deployment frequency without additional controls

2. Which activity is LEAST relevant to a formal AI security readiness baseline assessment?

   A. Mapping security controls to NIST AI RMF and ISO/IEC 27001 requirements

   B. Evaluating attestation capabilities for model-signing and artifact verification

   C. Reviewing organizational marketing segmentation strategies

   D. Identifying gaps in identity, telemetry, and data governance supporting AI workflows

3. What is the primary role of executive sponsorship in establishing enterprise AI security governance?

   A. Assigning all AI security tasks to DevOps engineers

   B. Authorizing policy, resource allocation, and escalation paths for AI risk management

   C. Minimizing security instrumentation to reduce cloud consumption

   D. Ensuring data scientists can bypass change control for rapid prototyping

4. In a mature DevSecOps pipeline for AI, what distinguishes continuous AI risk validation from traditional software CI/CD?

   A. Removal of model registries to reduce operational overhead

   B. Integration of model provenance checks, dataset integrity scanning, and adversarial evaluation gates

   C. Isolation of security controls to post-deployment monitoring only

   D. Exclusive reliance on unit-testing frameworks for model quality checks

5. Why must training programs be role-tailored in AI security readiness?

   A. To reduce learning curve complexity for nontechnical staff

   B. To ensure personnel understand threat models specific to their function, such as gradient poisoning for ML engineers or attestation validation for platform teams

   C. To prevent engineers from accessing advanced security content

   D. To meet HR compliance minimums

6. Which requirement is fundamental to an AI-specific incident response plan?

   A. Avoiding telemetry collection to minimize log retention

   B. Relying solely on conventional IOC-based detection

   C. Having the capability to analyze and respond to AI model drift, manipulation of feature pipelines, or adversarial perturbation vectors

   D. Excluding model versioning from evidence preservation

7. Which maturity framework is most applicable for evaluating contractor readiness in AI-sensitive DoD environments?

   A. ISO 26000

   B. SOX

   C. CMMC

   D. ITIL

8. What is a core operational requirement of explainable AI (XAI) readiness?
   - **A.** Ensuring model feature weights remain encrypted
   - **B.** Enabling regulated stakeholders to audit model outputs through interpretable mechanisms such as SHAP or counterfactual reasoning
   - **C.** Hiding the internal inferencing process to maintain IP secrecy
   - **D.** Replacing human review with automated decisioning
9. Adversarial training primarily mitigates which threat vector?
   - **A.** Unauthorized inference API enumeration
   - **B.** Input perturbations engineered to exploit model decision boundaries
   - **C.** Training data exfiltration from object storage
   - **D.** Poisoned labels introduced during dataset creation
10. Differential privacy provides protection against which form of model exploitation?
   - **A.** Gradient-shaping attacks that corrupt convergence
   - **B.** Membership inference and reconstruction attacks targeting training records
   - **C.** Time-based side-channel leakage in TPM-backed key storage
   - **D.** GPU-coalescing contention in high-density clusters
11. What is the primary function of data validation pipelines in preventing AI model corruption?
   - **A.** Reducing electronic medical record (EMR) data dimensionality
   - **B.** Providing efficient feature normalization
   - **C.** Detecting anomalous or malicious training inputs via statistical, semantic, or cryptographic verification
   - **D.** Increasing training throughput for large-batch optimizers
12. Why is least-privilege enforcement essential in Zero Trust AI systems?
   - **A.** To simplify IAM roles for system administrators
   - **B.** To ensure users, services, and pipelines cannot access model artifacts or data beyond their authorized trust zone
   - **C.** To enable broader data-sharing across federated teams
   - **D.** To reduce audit frequency
13. Homomorphic encryption is relevant for which AI security requirement?
   - **A.** Federated orchestration scheduling
   - **B.** Confidential inference on untrusted execution platforms
   - **C.** High-throughput checkpointing
   - **D.** GPU memory optimization

14. Under Zero Trust, continuous telemetry and log analysis enable what outcome?

A. Blocking all east–west traffic by default

B. Simplifying quarterly compliance reporting

C. Identifying anomalous model behavior, policy violations, or unauthorized pipeline access in real time

D. Accelerating incident recovery by skipping root-cause analysis

15. AI supply chain security focuses primarily on which requirement?

A. Minimizing dependency on proprietary accelerators

B. Ensuring transparency and integrity across upstream data sources, pretrained models, libraries, and toolchains

C. Avoiding adoption of open-source components entirely

D. Increasing AI build frequency

16. Which mechanism most effectively verifies integrity in AI supply chain components?

A. Accepting vendor claims without validation

B. Using cryptographically signed artifacts, reproducible builds, and SBoM-driven dependency verification

C. Avoiding all third-party ML frameworks

D. Storing all model files on a shared network drive

17. What is the overarching objective of AI supply chain risk management?

A. Reducing procurement delays

B. Eliminating all external dependencies

C. Ensuring the AI ecosystem remains secure and trustworthy even when third-party components are integrated

D. Minimizing cloud compute consumption

18. Within the DoD Zero Trust Architecture (ZTA) model, which execution enabler completes the framework by enabling advanced Zero Trust operations?

A. Training

B. Personnel

C. Organization

D. Facilities

19. Which of the following actions demonstrates measurable organizational commitment to AI security governance?

A. Standardizing exclusively on proprietary development languages

B. Reducing audit scope for efficiency

C. Achieving certifications such as ISO/IEC 27001 or ISO/IEC 42001 for AI governance

D. Restricting infrastructure access to a single administrator

20. When applying Zero Trust to AI workloads, which of the following is NOT required?

A. Enforcing identity-aware segmentation for models, pipelines, and data flows

B. Continuous verification of all access attempts

C. Immutable logging for inference and training operations

D. Restricting innovation to prevent risk

Answers

1. B

2. C

3. B

4. B

5. B

6. C

7. C

8. B

9. B

10. B

11. C

12. B

13. B

14. C

15. B

16. B

17. C

18. D

19. C

20. D

References

Department of Defense. "Cybersecurity Maturity Model Certification (CMMC)." (2021). https://dodcio.defense.gov/CMMC/.

European Union. "EU Artificial Intelligence Act" (2021). https://artificialintelligenceact.eu.

European Union. "Digital Operational Resilience Act (DORA)." (2022). https://finance.ec.europa.eu/digital-finance/digital-operational-resilience-act-dora_en.

European Union. "General Data Protection Regulation (GDPR)." Accessed January 15, 2025. https://gdpr.eu.

European Union Agency for Cybersecurity (ENISA). "NIS2 Directive." (2022). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022L2555.

European Union Agency for Cybersecurity (ENISA). "NIS Directive 2." 2023. https://www.enisa.europa.eu/topics/state-of-cybersecurity-in-the-eu/cybersecurity-policies/nis-directive-2.

International Organization for Standardization. "ISO/IEC 27001:2022 Information Security Management Systems—Requirements." Accessed January 15, 2025. https://www.iso.org/standard/27001.

International Organization for Standardization and International Electrotechnical Commission. "ISO/IEC 27002:2022 Information security, cybersecurity and privacy protection—Information security controls." Accessed January 15, 2025. https://www.iso.org/standard/75652.html.

Molnar, Christoph. Interpretable Machine Learning: LIME and SHAP Techniques for Explainable AI. Accessed January 15, 2025. https://christophm.github.io/interpretable-ml-book/lime.html.

Nadeem, Azqa, Daniël Vos, Clinton Cao, Luca Pajola, Simon Dieck, Robert Baumgartner, and Sicco Verwer. "SoK: Explainable Machine Learning for Computer Security Applications." In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, 221–240. IEEE, 2023. https://doi.org/10.1109/EuroSP57164.2023.00023.

NIST. "NICE Workforce Framework for Cybersecurity." Accessed June 3, 2026. https://www.nist.gov/itl/applied-cybersecurity/nice/nice-framework-resource-center.

NIST. "NIST Artificial Intelligence Risk Management Framework (AI RMF)." National Institute of Standards and Technology. (2023). https://www.nist.gov/itl/ai-risk-management-framework.

NIST. "NIST Cybersecurity Framework (CSF)." National Institute of Standards and Technology. (2024). https://www.nist.gov/cyberframework.

OWASP. "GenAI Incident Response Guide 1.0" (2025). https://genai.owasp.org/resource/genai-incident-response-guide-1-0/.

State of California Department of Justice. "California Consumer Privacy Act (CCPA)." Accessed January 15, 2025. https://oag.ca.gov/privacy/ccpa.

Index

A

AAA (authentication, authorization, and accounting), 122–123, 597

access control, 37, 264

accountability, 29, 30, 196–197, 236, 242–243, 595

adaptable AI systems, 198–199

adaptive governance, 595

adaptive misuse detection, 11

adaptive resilience, 533

adaptive threat response, 35

adversarial attacks, 206–207, 213, 240, 595

adversarial examples, 70, 73–74, 595

adversarial input, 595

adversarial machine learning, 9, 486, 595

adversarial robustness, 232–233

adversarial training, 87, 213

agentic AI, 33, 255, 257, 259, 260, 262, 405, 408–409, 595

- architectural design, 426
- autonomous data processing and contextual inference, 263
- chain of custody for data, 263
- consequences of data compromise, 289
- identity, state, and explainability, 426
- influence of visionary fiction, 259
- lifecycle governance, 287–288
- multi-agent environments and the MCP, 427
- network isolation and segmentation, 427
- privacy and security challenges, 262, 293
- risk management, 264
- securing, 207–209, 422–423
- security challenges, 427–428
- software development, 160–161
- Zero Trust guardrails, 55
  - *allow known good, block everything else, 55–56*
  - *fail closed, not open, 57*
  - *log everything, 56–57*
  - *make readable policies, 56*
  - *use multiple layers, 57–58*

agents, 33, 206

AI, 1, 31. *See also* GenAI; ML (machine learning); models; responsible AI; XAI
agentic, 33, 255, 257, 259, 260, 262, 405, 408–409. *See also* agentic AI
influence of visionary fiction, 259
lifecycle governance, 287–288
privacy and security challenges, 293
securing, 422–423
agents, 55–58
application development
securing, 200–201
transitioning from software development, 201–203
as an attack vector, 67, 69, 93–97
autonomous agents, 33
barriers to adoption, 36–37
siloed decision-making and political gridlock, 38
undeveloped use cases and unrealized potential, 38
BoM (bill-of materials), 286, 463–464, 596
continuous monitoring, 448
CoT (chain-of-thought) reasoning models, 33–34
logic exploits, 35
monitoring, 34
Zero Trust principles, 35
cybersecurity, 8–10, 11
data classification, 184
data governance, 258–259
deployment models, 516–517
autonomous agent AI, 55
cloud, 387–388
distributed AI, 54
edge computing, 388–389
enterprise-controlled AI, 53–54
externally managed AI, 54
hybrid cloud, 388
multi-cloud, 388
differential privacy, 79
distributed, 271
distributed intelligence, 260
DL (deep learning), 31
-driven automation, 258
dual use, 502
edge, 48
embedded, 259–260. *See also* embedded AI
enforcing access control and policy verification, 263–264
lifecycle governance, 287–288
embodied, 155
collaborative, ethical, and societal risks, 156
securing perception, planning and action, 155–156
simulation, noise, and real-world deployment, 156
threat landscape, 157–158
trust gap, 155
-enhanced malware, 496–498
essential skills, 44–45
ethical, 518
ethics, data-driven bias, 45–46
evolution of, 404–405
explainable, 41–42, 466
future-proofing, 197–198
generative, 8, 33, 93
governance, 38, 42, 274–275, 280–281
hallucinations, 110, 453
impacts on corporate security policy, 378

inference, 214, 271–272
IoT and, 20–21, 389
KPIs (key performance indicators), 450–451
literacy, 36
LLMs (large language models), 33
low trust, 37
maturity model, 51–53
ML (machine learning), 31
models. *See* model/s
monitoring, 447, 448–453, 458–460, 464–465, 533
-native software development, 160–161
offensive, 9, 69, 95–97
organizational security readiness. *See* organizational AI security readiness
in organizations, 43
origins, 8–9
pilots, 38
predictive, 33
privacy. *See* privacy
regulations, 182–183
reinforcement learning, 32
risk management, 36
secure by default, 87
secure by design, 86–87
security, 528–529, 539–540, 609–611
security monitoring, 11
security posture, 468–469
security readiness. *See* organizational AI security readiness
self-auditing ecosystem, 265
-specific incident response, 243–244
supervised learning, 32
supply chain attacks, 80–81
as a target, 67, 69
third-party risk, 2
threat landscape, 206–207, 519
threat model, 71–73
threats, 67–68, 107
tools, 36, 39
training, 17, 40–41
adversarial, 87, 93–94
bias, 45
considerations for data used for, 184–185
employee, 233, 237–238
exposure risks and mitigation, 188–189
privacy risks, 185–187
removing data points, 185
trust dependencies, 109
trust restoration, 469, 596
trustworthy, 40, 44–45, 175
unsupervised learning, 32
user awareness training, 503
vendors, securing, 314–316
core components, 316–317
Zero Trust foundations, 317–318
workloads, 14–15, 596
Zero Trust, 260–262
Zero Trust workshops, 150–151
capability alignment, 151–152
implementation, 151
AI adoption framework, 43, 595
agentic, privacy and security challenges, 262
change management, continuous learning, and ethical oversight, 44
data preparation and infrastructure development, 43–44

model development and deployment, 44
strategic assessment, 43
AI development lifecycle
applying Zero Trust, 438–439
data classification, 174, 191
estimated cost ratios for remediation, 194–195
governance, 431–439
integration of security, 189–191, 236–237
mapping Zero Trust controls, 436–438
AI governance board, 439–440, 596
AI maturity model, 596
AI RMF (AI Risk Management Framework), 6–7, 79, 596
AI security readiness, 596
AI supply chain assurance, 431, 596. *See also* **supply chain**
AI-enhanced malware, 596
AIID (AI Incident Database), 91–92, 596
AI/ML pipeline, 39, 517–518
data collection, 40
data security and integrity, 39–40
data validation, 40–41
LLMs, 40–41
AIOps, 453–454, 596
Air Canada, 196–197, 453
AI-ready data dictionary, 281–285
AI-ready environments, 13–14, 354, 379
AI security, 386–387
AI workloads, 382–383
AI/ML workloads, 377–378
architect-focused network design techniques, 372
failure isolation, 372
hierarchical design for scale, 373–374
modularity and composable design, 373
shared failure states, 372–373
technique synthesis, 374
businesses still exploring AI solutions, 392
cloud, 18
cloud provider offerings, 392
compute infrastructure, 15–16, 384–385
cost optimization strategies, 390–391
edge computing, 19–20
established or existing data centers, 391–392
hybrid cloud, 18–19, 377
infrastructure requirements, 15, 383
key components, 380–382
multi-cloud, 18–19
network design, 355–356
application requirements, 359–361
brownfield, 363
constraints, 364–365
design use cases, 362–364
deterministic throughput and latency, 356
functional requirements, 357–358
greenfield, 362
identifying requirements through root-cause analysis, 365–366
identity-based routing and access control, 356–357

infrastructure-agnostic security enforcement, 357
mergers and divestitures, 363
scaling a network, 364
segmentation aligned to AI workloads, 356
technical requirements, 358–359
technology or workload introduction, 363
technology replacement, 363
unified policy and telemetry fabric, 357
network design pitfalls
avoiding assumptions, 375
misapplied best practices, 376
overdesign, 375
preconceived notions, 376
network design principles
codify the topology and trust model, 367–368
design for enforceability, not just connectivity, 366–367
engineering for manageable complexity, 371
monitor, respond, and maintain policy integrity, 368
security as foundational design constraint, 368–369
security-aware availability, 370–371
networking, 17, 386
organizations with mature AI environments, 392
storage systems, 385
sustainable practices and energy efficiency, 390
AI-specific incident response readiness, 596
algorithm/s
bias, 46
"criminality from face prediction", 467
differential privacy, 186–187
model-based generative, 404
PQC (post-quantum cryptography), 162
shallow learning, 31
AML (adversarial machine learning), 9, 69–70, 88–89, 110. ***See also*** **attack/s**
adversarial examples, 70, 73–74
threat model, 71–73
analytics, 5, 323–324, 597
Analytics domain, 144
APM (Application Performance Monitoring), 144–145
asset monitoring and discovery, 150
auditing, logging, and monitoring, 145–146
change detection, 146
network threat behavior analytics, 147–148
SIEM (security information and event management), 148
threat intelligence, 148–149
traffic visibility, 149–150
anomaly-based detection, 2, 597
APIs, governance, 532
APM (Application Performance Monitoring), 144–145
application development
securing, 200–201
third-party risk, 309
transitioning from software development, 201–203
application requirements, AI-ready environments, 359–361
application security/development, 597

architect-focused network design techniques
failure isolation, 372
hierarchical design for scale, 373–374
modularity and composable design, 373
shared failure states, 372–373
technique synthesis, 374
architecture. *See also* ZTA (Zero Trust Architecture)
edge computing, 19–20
enterprise-grade agentic AI, 426
storage, 17
Zero Trust, NIST SP 800–207, 3
ART (Adversarial Robustness Toolbox), 88
Asimov, Isaac, 259, 261–262
asset management, 127–128, 307, 597
asset monitoring and discovery, 150
"assume breach" mindset, 216
assurance, 597
assurance reporting, 276
attack/s
adversarial, 206–207, 213
AI-driven, 486
attribute inference, 175, 187–188, 597
availability, 597
backdoor, 75–76, 520–521
C&W (Carlini and Wagner), 74
data poisoning, 74–75, 90, 110–111, 241, 457, 519–520
DDoS (distributed denial-of-service), 136–137
energy-latency, 79–80
evasion, 73–74, 90, 519–520
feature space, 72
generative AI, 83–84
GPT-2 replication, 188
label poisoning, 179
logic exploits, 35
membership inference, 78–79, 185–186, 241
model extraction, 78, 90–91, 185–186, 456
model inversion, 77–78, 240–241, 266
namespace reuse, 343
problem space, 72
Rules File Backdoor, 462
side-channel, 491, 492–493, 494
social engineering, 9
sponge, 90
state-sponsored, 158–160
supply chain, 80–81, 111, 521
zero-width, 202
attestation, 159, 268
attribute inference attack, 175, 187–188, 597
auditing, logging, and monitoring, 145–146. *See also* monitoring
authentication, multifactor, 127, 206
automation, AI-driven, 258
automation tools, 40–41
autonomous agents, 33, 55
availability, security-aware, 370–371
availability attacks, 79–80, 597
awareness campaigns, 238, 597

B

backdoor attacks, 75–76, 520–521, 597
barriers to AI adoption, 36–37, 38
siloed decision-making and political gridlock, 38
undeveloped use cases and unrealized potential, 38

baseline, current capabilities, 227–228
behavioral trust, 261
benchmarking, 538
Bengio, Yoshua, 539
bias, 597
algorithmic, 46
data-driven, 45–46
human, 45
inductive, 41
model, 232–233
Biggio, Battista, 9
black box model, 184, 244, 597
black-box scenario, 72
BlackMamba, 497, 597
blast radius, 597
blockchain, 49
BoM (bill-of materials), 81–83, 286
BoM methodologies, 111
brownfield, AI deployment, 363
business continuity, 118–119, 597. ***See also*** **data centers**
for AI ecosystems, 345
for AI-enabled applications, 474–475
building AI-specific strategies, 476
core challenges for AI-enabled applications, 475
data centers, 469
recovery preparedness, 233–234
third-party risk, 310

C

C&W (Carlini and Wagner) attack, 74
CA (Certificate Authority), 123–124, 597
Cambridge Analytica, 175
capabilities, Zero Trust, 111–112
aligning with organizational dynamics, 153
aligning with roadmap, 154
aligning with workshops, 151–152
CASB (cloud access security broker), 136, 598
chain of custody, data, 263
chain-of-thought prompting, 34
change control, 598
change detection, 146
change management, 44, 113–114
chatbots, 409. ***See also*** **agentic AI; GenAI**
Air Canada, 196–197, 453
enterprise use cases, 411–412
fine-tuning of GenAI models, 413
retrieval-augmented generation, 412–413
LLMs in the wild, 494–496
securing, 206
security considerations, 411, 414–416
ChatGPT, 8, 403, 406–407, 496–497
Checkmarx, 461–462
chiplets, 385, 598
choosing the right model, 41
CI/CD (continuous integration/continuous deployment), 40–41, 191, 203, 204–205, 319, 461–462, 598
CIO/CISO, security considerations for data centers, 470–471
CISA (U.S. Cybersecurity and Infrastructure Security Agency), 86, 201
Cisco, responsible AI, 29–30
Cisco AI Defense, 420–421, 459–460

Cisco Secure Access, 459
Clarke, Arthur C, 259, 261–262, 515, 537
clean-label poisoning, 74–75
cloud
-based AI, 18, 387–388
hybrid, 18–19, 388
IoT devices, 49
MLaaS (Machine Learning as a Service), 78
multi-, 19, 388
security, third-party risk, 310–311
CMDB (configuration management database), 128–130, 281–285, 307
CMMC (Cybersecurity Maturity Model Certification), 230–232
CMMI (Capability Maturity Model Integration), 51–53
cognitive threat agent, 598
cognitive trust, 261
commitment, leadership, 216
compliance, 196–197, 295
AI privacy, 272–273
-as-code, 279
data classification, 181–183
GDPR (General Data Protection Regulation), 185
regulatory, 214–215, 230–232
telemetry, 273, 276
third-party risk, 310
Zero Trust, security, 108
composable design, 373
compute infrastructure, 15–16, 384–385
concept drift, 450, 598
confidential computing, 267–268, 598
confidentiality, 76–77, 159–160
membership inference attacks, 78–79
model extraction attacks, 78, 90–91
model inversion attacks, 77–78
confidentiality attack, 598
constraints, AI-ready environments, 364–365
containment, self-adaptive AI models, 265
contextual access policies, 207–208, 278–279
contextual inference, 263
continuous assurance, 279, 598
continuous learning, AI security, 239
continuous monitoring, 246, 323–324
AI systems, 448
vendors, 329–331
continuous verification, 598
core principles
ethical AI development, 30
responsible AI, 29
cost-benefit analysis
AI infrastructure, 391
offensive AI, 492–494
CoT (chain-of-thought) reasoning models, 33–34, 598
logic exploits, 35
monitoring, 34
Zero Trust principles, 35
counterfactual explanations, 245
criminality from face prediction algorithm, 467
crosswalks, 113, 319. *See also* questions with crosswalks
CrowdStrike outages, 343–345
crypto-agility, 197–198, 266, 269, 538, 598
cryptographic controls, risk assessment, 308

CRYSTALS-Dilithium, 51
CRYSTALS-Kyber, 51
cultural buy-in, 36
cultural change strategies, 598
culture, security-aware, 216
current state. *See* baseline
CVE (Common Vulnerabilities and Exposures), 4, 159, 598
cyberattacks, 9
cybercrime, 69, 494–496. *See also* deepfakes; fraud; theft
cyber-espionage, 161
cybersecurity, 2
AI and, 8–10
core capabilities, 524
ground truth labeling, 10
intersections of AI and, 486
Morris Worm, 4
proactive measures, 2, 10
reactive measures, 2, 10
SIEM (security information and event management), 11
signature-based detection, 10

D

DarkBERT, 495, 598
Darknet, 494–496
data
chain of custody, 263
encryption. *See* encryption
as an enterprise and national security asset, 256–257
governance, 258
protecting, 257
security, 39–40, 530
strategic, 256–257
synthetic, 93–94
validation, 40
data anonymization, 187, 242
data breaches, penalties, 289–291
data centers
AI design considerations, 472–474
AI workloads, 14–15
AI-ready environment
compute infrastructure, 15–16
infrastructure requirements, 15
networking, 17
storage systems, 16–17
business continuity, 469
core risks to AI workloads, 474
DR (disaster recovery), 469
FPGAs (field-programmable gate arrays), 16
GPUs (graphics processing units), 16
security considerations, 470–471
third-party risk, 311
TPUs (tensor processing units), 16
Zero Trust integration, 469–470
data classification, 177, 525, 599
AI, 184
AI development lifecycle, 191
considerations for publicly released data, 187–189
data labeling, 177, 178, 179
failures, 178–179
lifecycle, 178
metadata, 179
sensitivity labels, 178, 180–181
early, 174
criteria, 175–176
embedding in data governance and AI governance frameworks, 176

securing intellectual property, 175
ethical considerations, 526
governance and legal requirements, 181–183
tools, 176–177
for trustworthy AI, 175
data collection, AI/ML pipeline, 40
data drift, 450, 599
data governance, 241–243, 599
data labeling, 599
data labeling lifecycle, 178
data leakage from generative AI platforms, 291
data lineage, 599
data poisoning attacks, 74–75, 90, 110–111, 157, 241, 457, 519–520, 599
data privacy, 255–256. ***See also*** **privacy**
data provenance tracking, 40, 242, 599
data residency, 258, 276
data retention, 115–116, 276
data retention policies, 192
data sovereignty, 599
database change control, 135
DDoS (distributed denial-of-service) attacks, 136–137, 343, 498–499
decision-making
AI-driven automation, 258
models, 46
siloed, 38
Deep Locker, 496–497, 599
deepfakes, 92–93, 304, 341
defense in depth, 57
defense industry, AI privacy and compliance requirements, 275
Deloitte AI incident, 439–440
dependencies
AI third-party, 331–332
model, 286–287
trust, 109
de-perimeterization, 4, 108–109
deployment models
autonomous agent AI, 55
cloud, 387–388
distributed AI, 54
edge computing, 388–389
enterprise-controlled AI, 53–54
externally managed AI, 54
hybrid cloud, 388
multi-cloud, 388
design use cases, AI-ready environments, 362–364
brownfield, 363
greenfield, 362
mergers and divestitures, 363
scaling a network, 364
technology or workload introduction, 363
technology replacement, 363
devices, IoT (Internet of Things), 48
DevOps, 30, 191, 203, 599
CI/CD (continuous integration/continuous deployment), 204–205
vs. MLOps, 203–204
DevSecOps, 30, 203, 205–206, 215–216, 599
Dick, Philip K., ***Do Androids Dream of Electric Sheep?*****, 539**
differential privacy, 79, 185, 186–187, 270–271, 599
digital sovereignty, 258–259
direct prompt injection, 84

distributed AI, 54, 271, 287
distributed intelligence, 260
DL (deep learning), 31, 44, 88–89, 404, 496–497, 599
LLMs, 33
models, 93
DLP (data loss prevention), 37, 137–139, 459, 599
domains, Zero Trust, 5
Analytics, 144
APM, 144–145
asset monitoring and discovery, 150
auditing, logging, and monitoring, 145–146
change detection, 146
network threat behavior analytics, 147–148
SIEM, 148
threat intelligence, 148–149
traffic visibility, 149–150
Enforcement, 135–136
CASB, 136
DDoS attacks, 136–137
DLP, 137–139
file integrity monitoring, 142–143
firewall, 139
IPS, 139–140
proxy, 140–141
segmentation, 143–144
SOAR, 141–142
VPN/RA, 141
Identity, 122
AAA, 122–123
asset management, 127–128
CA (Certificate Authority), 123–124
CMDB (configuration management database), 128–130
IP schemas and addressing, 130–131
MAF (multifactor authentication), 127
NAC, 124–125
PAM (privileged access management), 126–127
provisioning, 125–126
Policy and Governance, 112–113
business continuity, 118–119
change control, 113–172
data governance and encryption, 114–115
data retention, 115–116
disaster recovery, 119–120
quality of service, 116–117
redundancy and replication, 117–118
risk classification policy, 120–121
segmentation, 121–122
Vulnerability Management, 131, 133–134
database change control, 135
endpoint protection and encryption, 132–133
malware prevention and detection, 133
vulnerability scanning and threat intelligence, 134–135
DORA (Digital Operational Resilience Act), 310, 318, 334, 339–340, 600
DR (disaster recovery), 119–120, 599. *See also* data centers
for AI ecosystems, 345
for AI-enabled applications, 474–475

building AI-specific strategies, 476
core challenges for AI-enabled applications, 475
data centers, 469
recovery preparedness, 233–234
third-party risk, 310
dual use, 502, 600
dynamic DDoS attack, 498–499

E

early data classification, 174. ***See also*** **data classification**
criteria, 175–176
embedding in data governance and AI governance frameworks, 176
securing intellectual property, 175
east-west traffic inspection, 17
edge AI, 48
edge computing, 19–20, 388–389, 600
EDR (endpoint detection and response) systems, 497
email
phishing, 499–502
security, 138–139
spam, 32
embedded AI, 259–260, 428–430, 600
consequences of data compromise, 289
enforcing access control and policy verification, 263–264
lifecycle governance, 287–288
Musk Robot Army, 430–431
privacy and security challenges, 262, 293
embodied AI, 155, 600
collaborative, ethical, and societal risks, 156
securing perception, planning and action, 155–156
simulation, noise, and real-world deployment, 156
threat landscape, 157–158
trust gap, 155
emerging technologies
blockchain, 49
IoT (Internet of Things), 48
edge AI, 48
federated learning, 48
PQC (post-quantum cryptography), 50–51
quantum computing, 49–50
employee, training, 233, 237–238
encryption, 50, 114–115, 132–133, 246
homomorphic, 242, 270–271
post-quantum, 268
PQC-based, 271–272
quantum-safe, 161–162
-in-use, 271
endpoint protection, 132–133
energy sector, AI privacy and compliance requirements, 275
energy-latency attacks, 79–80
enforcement, 600
designing networks for, 366–367
policy, 322–323
Zero Trust, 5
Enforcement domain, 135–136
CASB (cloud access security broker), 136
DDoS (distributed denial-of-service) attacks, 136–137
DLP (data loss prevention), 137–138

email security, 138–139
FIM (file integrity monitoring), 142–143
firewall, 139
IPS (intrusion prevention system), 139–140
proxies, 140–141
segmentation, 143–144
SOAR (security orchestration, automation, and response), 141–142
VPN/RA, 141
entanglement, 49
enterprise-controlled AI, 53–54
essential skills for the AI era, 44–45
ethical AI, 46, 518, 600. ***See also*** **responsible AI**
algorithmic bias, 46
core principles, 30
data-driven bias, 45–46
frameworks, 46
actionable recommendations, 47
broad principles, 47
regulations, 47
human bias, 45
ethical requirements, AI, 214–215
ethics
human bias, 45
PPDE (privacy-preserving data engineering), 269–270
EU (European Union)
AI Act, 3, 7–8, 274, 290, 600
AI governance, 274–275
GDPR (General Data Protection Regulation), 184, 185, 274
NIS2, 310, 318, 324, 334, 604
evasion attacks, 73–74, 90, 519–520, 600
events, 11
evolution of AI, 404–405
expanded capability set, Zero Trust, 4–5
explainability, model, 46
externally managed AI, 54
EyeSpy, 497–498, 600

F

failure isolation, 372, 600
fairness, 29, 600
feature importance analysis, 244
feature space, 72
federated learning, 48, 79, 260, 271, 600
FIM (file integrity monitoring), 142–143
financial services, AI privacy and compliance requirements, 275
fine-pruning, 76
fine-tuned models, security, 416–418
fine-tuning, 413, 452, 600
FIPS 203–205, 197, 268, 271, 537, 600
firewalls, 139
Forrester Research, 4
foundational model, 83
FPGAs (field-programmable gate arrays), 15, 16, 600
frameworks
AI governance, 42
CMMC (Cybersecurity Maturity Model Certification), 230–232
crosswalks, 319
ethical AI, 46
actionable recommendations, 47
broad principles, 47
regulations, 47
MITRE ATLAS, 67, 68, 89, 91, 92

Ray, 80
regulatory, 175
"Responsible Practices for Synthetic Media", 92–93
Secure Software Development, 87
fraud, 92–93, 493–494. ***See also*** **deepfakes; theft**
FraudGPT, 495, 608
functional requirements, AI-ready environments, 357–358
future-proofing AI systems, 197–198, 293–294

G

GAI (general artificial intelligence), 407
GANs (generative adversarial networks), 93–97, 601
gap analysis, 229, 600–601
GDPR (General Data Protection Regulation), 184, 185, 274
GenAI, 8, 33, 93, 401, 402, 601, 609. ***See also*** **agentic AI**
adoption, 406
attacks, 83
ChatAI applications
enterprise use cases, 411–412
fine-tuning of GenAI models, 413
retrieval-augmented generation, 412–413
security considerations, 411, 414–416
ChatGPT, 8, 403, 406–407
criminal use of, 493
data leakage, 291
enterprise applications, 406–409
fine-tuned models, security, 416–418
foundational Zero Trust security, 431–439
LLMs (large language models), 33
model-based generative algorithms, 404
models, 83, 403
multi-modal models, 402
prompt injection attacks, 83–84
direct, 84
indirect, 84–86
mitigation, 86
proof-of-concept initiatives, 407
rule-based generative systems, 403–404
securing user interactions, 419–422
use cases, 408
Zero Trust, 410
global privacy, mapping to AI systems, 273
Goertzel, Ben, 539
golden model, 475, 601
Google Assistant
leveraging for social engineering attacks, 500–502
vulnerabilities, 500
govern function, AI RMF, 36
governance, 318–320, 601. ***See also*** **Policy and Governance domain**
adaptive, 595
AI, 38, 42, 211–212, 280–281
AI data, 258–259
AI security, 235–236
API, 532
data, 37, 114–115, 241–243
data classification, 181–183
Deloitte AI incident, 439–440
distributed AI, 287

lifecycle, 269, 287–288
third-party risk, 310
GPT-2 replication attack, 188
GPU (graphics processing unit), 15–16
gray-box scenario, 72
greenfield, AI deployment, 362
Green-Ortiz, Cindy, *Zero Trust Architecture*, 4, 6, 522
ground truth labeling, 10
guardrails, 601
AI agent, 55
allow known good, block everything else, 55–56
fail closed, not open, 57
log everything, 56–57
make readable policies, 56
use multiple layers, 57
chatbots, 206
guidelines, ethical AI, 47

H

Hacker News, The, **304**
hallucinations, 110, 453
hands-on exercises, 239, 601
hardware-based isolation, 272
Harvard Business Review, **33**
healthcare, AI privacy and compliance requirements, 275
Heinlein, Robert, 539
hierarchical network design, 373–374
HIMSS (Healthcare Information and Management Systems Society), 51–53
Hinton, Geoffrey, 515, 536–537
HIPAA (Health Insurance Portability and Accountability Act of 1996), 181, 187, 601
HNDL ("harvest now, decrypt later"), 161, 162, 337–338, 601
homomorphic encryption, 242, 270–271, 601
HQC, 51
Hugging Face, 342–343, 460–461
human bias, 45
human resources management, third-party risk, 306–307
hybrid cloud, 18–19, 377, 388, 601

I

IAM (Identity and Access Management), 35
IBM, 2025 Cost of a Data Breach Report, 291
identity and access management, 601
third-party risk, 320–321
Zero Trust, 5
Identity domain, 122
AAA (authentication, authorization, and accounting), 122–123
asset management, 127–128
CA (Certificate Authority), 123–124
CMDB (configuration management database), 128–130
IP schemas and addressing, 130–131
MAF (multifactor authentication), 127
NAC (network access control), 124–125
PAM (privileged access management), 126–127
provisioning, 125–126
identity-aware access, 278
IDS (intrusion detection system), 10, 602
IEEE P7001 Standard for Transparency, 41–42

immutable storage, 470, 601
implementation
organizational dynamics, 152–153
PQC (post-quantum cryptography), 162–163
roadmap for Zero Trust, 154
Zero Trust workshops, 151
implicit trust, 1, 107
incident management, third-party risk, 309–310
incident response, 192, 214, 243–244, 533
incident response and recovery preparedness, 233–234
incident response planning, 11, 465–469
indirect prompt injection, 84–86
inductive bias, 41, 601
inference, 214, 601–602
contextual, 263
PQC-based encryption, 271–272
validation, 266–267
inference endpoints, 456, 602
inference logging, 450
infrastructure, 531. ***See also*** **AI-ready environments**
-agnostic security enforcement, 357
AI-ready environments, 383, 384–385
security-resilient, 215–216
input logging, 449, 602
input validation, 203
integrity, 159–160
backdoor attacks, 75–76
data poisoning attacks, 74–75, 90, 110–111
evasion attacks, 73–74
model, 212–214
validation, 267
integrity attack, 602
integrity check, 602
intellectual property, 187–188, 257
securing, 175
theft, 186, 188
interpretability, XAI, 41–42
inventory management, CMDB (configuration management database), 281–285
IoT (Internet of Things), 20–21, 48, 602
AI and, 389
edge AI, 48
federated learning, 48
IP schemas and addressing, 130–131
IPS (intrusion prevention system), 10, 139–140, 602
ISO/IEC 23894, 30

J-K

Jericho Forum, 4, 108–109
JIT (just-in-time) provisioning, 320, 323, 602
key provisions
AI RMF, 7
EU AI Act, 8
Kindervag, John, 4
kinetic risk, 430, 602
KMS (key management system), 322, 327, 602
KPIs (key performance indicators), 160
AI system, 450–451
enforcement effectiveness, 322

L

label poisoning, 179, 602
labeled data, 32, 174, 177, 179. *See also* **data classification**
data labeling lifecycle, 178
failures, 178–179
sensitivity, 180–181
sensitivity labels, 178
leadership
commitment, 216
creating a security-first culture, 235
least-privilege access, 35, 245–246, 277, 602
LeCun, Yann, 539
legal frameworks, data classification, 181–183
lifecycle governance, 269, 287–288
LIME (local interpretable model-agnostic explanations), 41–42, 245
lineage tracking, 286–287
literacy, AI, 36
Llama Drama, 81
LLMOps, 448, 454, 602
LLMs (large language models), 33, 183, 602
adversarial attacks, 83
AI/ML pipeline, 40–41
changing in an AI system, 451–452
output handling, 83
reasoning models, 34
risk management, 33
security considerations, 452
"in the wild", 494–496
workflows, 33
logging, 214, 449–450
inference, 450
input, 449
logic exploits, 35
lossless Ethernet, 472, 602

M

machine learning, 9
adversarial, 9, 486
supervised, 10
unsupervised, 10
machine unlearning, 185
macrosegmentation, 121
MAF (multifactor authentication), 127
MalGAN, 94–95
malware
AI-enhanced, 496–498
polymorphic, 497, 505–506
malware prevention and detection, 133
Marsh, Stephen Paul, "Formalising Trust as a Computational Concept", 1, 3–4
MathGPT, 90
maturity model, AI, 51–53
MBoM (model bill of materials), 81–83
MCP (Model Context Protocol), 424–425, 461, 603
MDLC (model development life cycle), 39
media handling, third-party risk, 307
membership inference attacks, 78–79, 185–186, 241, 603
mergers and divestitures, AI environment, 363
metadata, 179, 281–285. *See also* **labeled data**
metamorphism, 10
metrics, proactive security, 193
MFA (multifactor authentication), 206, 604

microsegmentation, 17, 108, 121, 278, 603
Microsoft 365, sensitivity labels, 178
Mikhail, Mike, *Quantum Advantage*, 198
misinformation campaigns, 489–490
mitigation
AI security risk, 228–229
AI-specific threats and attack vectors, 214
backdoor attacks, 75–76
DDoS attacks, 343
deepfakes, 341
exposure, 188–189
prompt injection, 86
MITRE ATLAS, 67, 68, 89, 91, 92, 188, 604
MITRE ATT&CK, 603
ML (machine learning), 31, 44. *See also* AML (adversarial machine language)
ML Privacy Meter, 79
MLaaS (Machine Learning as a Service), 78
MLOps, 30, 191, 203, 603
vs. AIOps, 453–454
CI/CD (continuous integration/continuous deployment), 204–205
vs. DevOps, 203–204
MLSecOps, 30, 205–206, 603
mobile security, third-party risk, 306
model drift, 37, 603
model exfiltration, 291–292
model extraction attack, 78, 90–91, 185–186, 456, 603
model inversion attack, 77–78, 240–241, 266, 603
model poisoning attack, 603
model reconstruction, 77
model watermarking, 266–267, 603
modeling, trust, 3–4
model/s. *See also* threat model
versus AI systems, 88–89
attacking, 68
availability attacks, 79–80
backdooring, 467
-based generative algorithms, 404
bias, 232–233
choosing, 41
confidentiality
membership inference attacks, 78–79
model extraction attacks, 78
model inversion attacks, 77–78
decision-making, 46
dependency mapping, 286–287
deserialization, 461–462
development and deployment, 44
DL (deep learning), 93
explainability, 46, 232–233
fine-tuned, security, 416–418
fine-tuning, 413, 452
foundational, 83
GenAI, 403
generative, 83
inductive bias, 41
integrity, 212–214
backdoor attacks, 75–76
data poisoning attacks, 74–75
evasion attacks, 73–74
libraries for testing, 88
lineage, 449, 603
lineage tracking, 286–287
monitoring, 213

multi-modal, 402
NFL (No Free Lunch) theorem, 41
reasoning, 34
removing data points, 185
replication, 188, 603
securing, 265–266
security, 461–462
security assessment, 232–233
security readiness, 240–241
self-adaptive
containment, 265
monitoring, 265
signing, 266–267
theft, 207, 214
modularity, 373, 374, 391, 603
monitoring
AI systems, 447, 464–465, 533
application performance, 144–145
changes in regulations, 455
continuous, 246, 323–324
CoT (chain-of-thought) reasoning models, 34
file integrity, 142–143
models, 213
real-time, 458
self-adaptive AI models, 265
solutions for AI systems, 458–460
telemetry, 35
threat landscape, 455–457
vendor performance, 311–313, 329–331
Morris Worm, 1, 4, 108–109
Mula, Christopher, 611
multicloud, 18–19, 388, 604
multi-layer defense, 57, 604
multi-modal models, 402
Musk Robot Army, 430–431

N

NAC (network access control), 124–125, 604
namespace reuse attack, 343
Nature, **76–77**
NetFlow, 11, 42, 604
network threat behavior analytics, 147–148
networks. ***See also*** **cloud; data centers**
AI-ready, 17, 355–356, 379, 386
AI security, 386–387
AI workloads, 382–383
AI/ML workloads, 377–378
application requirements, 359–361
architect-focused network design techniques, 372–374
avoiding assumptions, 375
brownfield, 363
businesses still exploring AI solutions, 392
cloud provider offerings, 392
codify the topology and trust model, 367–368
compute infrastructure, 15–16, 384–385
constraints, 364–365
cost optimization strategies, 390–391
design for enforceability, not just connectivity, 366–367
design use cases, 362–364
deterministic throughput and latency, 356
engineering for manageable complexity, 371
established or existing data centers, 391–392

functional requirements, 357–358
greenfield, 362
hybrid cloud, 377
identifying requirements through root-cause analysis, 365–366
identity-based routing and access control, 356–357
infrastructure requirements, 15, 383
infrastructure-agnostic security enforcement, 357
key components, 380–382
mergers and divestitures, 363
misapplied best practices, 376
monitor, respond, and maintain policy integrity, 368
organizations with mature AI environments, 392
overdesign, 375
preconceived notions, 376
scaling, 364
security as foundational design constraint, 368–369
security-aware availability, 370–371
segmentation aligned to AI workloads, 356
storage systems, 16–17, 385
sustainable practices and energy efficiency, 390
technical requirements, 358–359
technology or workload introduction, 363
technology replacement, 363
unified policy and telemetry fabric, 357
workloads, 14–15
de-perimeterization, 4, 108–109
generative adversarial, offensive AI, 93–97
RoCEv2 (Remote Direct Memory Access over Converged Ethernet version 2), 471–472
security, third-party risk, 308–309
segmentation, 121–122, 143–144, 246
neural cleanse, 75–76
neural networks, 31, 93–94
"never trust, always verify", 503
NFL (No Free Lunch) theorem, 41, 604
NIS2, 310, 318, 324, 334, 604
NIST AI RMF (AI Risk Management Framework), 3, 36, 79, 191–192, 274, 604. *See also* AI RMF (AI Risk Management Framework)
NIST Post-Quantum Standardization Project, 51
NIST SP 800–53 Rev. 5, 37
NIST SP 800–207, 3, 4, 6, 12
NIST SP 800–218, 87, 190–191
NIST SP 800–226 (Draft), 79
NIST-AI-600–1, 7
NLP (natural language processing), 11, 33–34
NVMe-oF (NVMe over Fabrics), 471, 604

O

ODCs (offshore delivery centers), third-party risk, 311
offensive AI, 9, 69, 485–486, 521, 538–539, 604
advanced monitoring and response, 503
AI-enhanced malware, 496–498
cost-benefit analysis, 492–494
defending against, 488

dynamic DDoS attack, 498–499
in GANs, 93–97
intersections of AI and cybersecurity, 486
LLMs in the wild, 494–496
scale, 488
scope, 487, 488–489
sophistication, 487, 502–503
speed, 487
targets, 489
as threat to individuals, 490
as threat to society, 492
as threat to systems, 491
online probing, 450–451, 604
OPA (Open Policy Agent), 322, 604
OPA/Rego, 604
operational trust, 261
operations, third-party risk, 308–309
organizational AI security readiness, 209–211, 225, 528–529, 530
actionability and roadmap development, 234
assessing, 226
baseline of current capabilities, 227–228
building a security-first culture, 234–235
building a security-resilient infrastructure, 215–216
challenges and considerations, 240
AI model security readiness, 240–241
AI supply chain readiness, 246–247
AI-specific incident response readiness, 243–244
data governance and privacy, 241–243
XAI readiness, 244–245
compliance and regulatory readiness, 214–215, 230–232
continuous learning, 239
cultivating a security-aware culture, 216
cultural change strategies, 237
culture and awareness readiness, 233
embedding security in AI governance frameworks, 211–212
ensuring model integrity and robustness, 212–214
gap analysis and areas for improvement, 229
hands-on exercises, 239
incident response and recovery preparedness, 233–234
integration of security in AI development lifecycle, 236–237
leadership and commitment, 235
measuring effectiveness, 239–240
mitigating AI-specific threats and attack vectors, 214
risk assessment and prioritization, 228
risk management and accountability, 236
security policies and governance, 235–236
security readiness assessment, 226–227
stakeholder engagement and ownership, 226
strengthening data security and privacy, 212
technical infrastructure and tooling evaluation, 232–233
training and awareness programs, 237–238
Zero Trust principles, 245–246

organizational dynamics, 152
capability alignment, 153
implementation, 152–153
risk and regulation, 152
organizations
AI, 43
AI Adoption Framework, 43
change management, continuous learning, and ethical oversight, 44
data preparation and infrastructure development, 43–44
model development and deployment, 44
strategic assessment, 43
AI governance, 38
AI pilots, 38
barriers to AI adoption, 36–37
investment in immature technologies, 39
siloed decision-making and political gridlock, 38
undeveloped use cases and unrealized potential, 38
cultural buy-in, 36
internal processes, 457–458
investment in immature technologies, 39
low trust, 37
origins of Zero Trust, 3–4, 108–110
OSINT (open-source intelligence), 490, 604
outages, CrowdStrike, 343–345
output handling, 83
overdesign, network, 375
OWASP Foundation
Top 10 for LLM Applications, 200
Top 10 for Machine Learning, 266
Top 10 risks of adversarial attacks against LLMs, 83

P

packers, 10
PAI (Partnership on AI), "Responsible Practices for Synthetic Media" framework, 92–93
PAM (privileged access management), 126–127, 605
parallel file systems, 16–17, 604
PassGAN, 94–95
patch and pray, 604
PCA (principal component analysis), 76
PCI DSS (Payment Card Industry Data Security Standard), 181, 604
penalties for AI failures, 289–291
PEP (policy enforcement point), 303–304, 315, 327, 605
performance
benchmarking, 538
vendor, continuous monitoring, 311–313, 329–331
PGD (projected gradient descent), 74
phishing attacks, 499–502
phishing kits, 9
physical and environmental security, third-party risk, 308
PKI (Public Key Infrastructure), CA (Certificate Authority), 123–124
Policy and Governance domain, 112–113. ***See also*** **policy/ies**
business continuity, 118–119
change control, 113–114
data governance and encryption, 114–115
data retention, 115–116
DR (disaster recovery), 119–120
QoS (quality of service), 116–117
redundancy and replication, 117–118

risk classification policy, 120–121
segmentation, 121–122
policy/ies, 32, 318–320
building, 378–379
-as-code, 318
contextual access, 207–208, 278–279
corporate security, impacts of AI on, 378
crosswalks, 319
data retention, 192, 276
enforcement, 322–323
readable, 56
responsible AI, 30
risk classification, 120–121
security, 235–236
verification, 263–264
Zero Trust, 5
polymorphic malware, 497, 505–506, 605
polymorphism, 10
post-quantum signed tokens, 605
PoV (proof of value), Zero Trust, 108
PPDE (privacy-preserving data engineering), 269–270
PQC (post-quantum cryptography), 50–51, 115, 158, 161, 256, 262, 274, 305, 537–538, 605
-based encryption, 271–272
crypto-agility, 197–198, 269
implementation, 162–163
insights and business implications, 163–164
integration within Zero Trust data protection models, 279
readiness, 265–266, 268, 293–294
third-party risk, 334–338
predict and prevent posture, 468–469, 605
predictive AI, 33
privacy, 76–77, 110, 293, 530. ***See also*** **confidentiality**
AI security readiness, 241–243
challenges in agentic and embedded AI, 262
compliance telemetry, 276
data, 255–256
differential, 79, 185, 186–187, 270–271
failures, business and operational impact, 288–289, 291–292
global, mapping to AI systems, 273
hardware-based isolation, 272
industry-specific considerations, 275
key Zero Trust enablers, 294
regulatory frameworks, 272–273
“right to be forgotten”, 276
risks of trained models, 185–187
ZKPs (zero-knowledge proofs), 272
proactive security measures, 2, 191–192. ***See also*** **organizational AI security readiness**
business and operational benefits, 193
challenges and considerations, 199–200
cost savings, 194–196
cybersecurity, 10
future-proofing AI systems, 197–198
improving trust and compliance, 196–197
moving from software development to AI application development, 201–203
quantitative and qualitative metrics, 193
scalability and adaptability of secure AI, 198–199

securing AI application development, 200–201
value propositions, 193–194
problem space, 72
prompt injection, 83–84, 605
direct, 84
indirect, 84–86
mitigation, 86
prompts, chain-of-thought, 34
protecting data, 257
provenance tracking, 605
provisioning, 125–126, 605
proxies, 140–141
Purdy, Mark, on agentic AI, 33
Python, 460–461
libraries for testing AI models, 88
Llama Drama, 81
sponge attack, 90

Q

QoS (quality of service), 116–117, 605
quantum computing, 49–50, 256, 518–519
quantum-resistant cryptography, 537
quantum-safe encryption, 161–162
questionnaire
analytics and continuous monitoring, 324
how to use, 324–326
identity and access management, 320–321
policy and governance, 319–320
policy enforcement, 323
section summary, 327–328
vulnerability management, 321–322
questions with crosswalks
AAA (authentication, authorization, and accounting), 123
APM (Application Performance Monitoring), 145
asset management, 128
asset monitoring and discovery, 150
auditing, logging, and monitoring, 145–146
business continuity, 119
CA (Certificate Authority), 124
CASB (cloud access security broker), 136
change control, 113–114
change detection, 146
CMDB (configuration management database), 129
data governance and encryption, 114–115
data retention, 116
database change control, 135
DDoS (distributed denial-of-service) attacks, 137
DLP (data loss prevention), 137–138
DR (disaster recovery), 120
email security, 138–139
endpoint protection and encryption, 132–133
file integrity monitoring, 142–143
firewall, 139
IP schemas and addressing, 131
IPS (intrusion prevention system), 140
MAF (multifactor authentication), 127
malware prevention and detection, 133
NAC (network access control), 125

network threat behavior analytics, 147–148
PAM (privileged access management), 126–127
provisioning, 125–126
proxies, 140–141
QoS (quality of service), 118
redundancy and replication, 117–118
risk classification policy, 120–121
segmentation, 120–121, 143–144
SIEM (security information and event management), 148
SOAR (security orchestration, automation, and response), 142
threat intelligence, 149
traffic visibility, 149–150
VPN/RA, 141
vulnerability management, 134
vulnerability scanning and threat intelligence, 134–135

R

RA (remote access), 141
RAG (retrieval-augmented generation), 207–209, 412–413, 416–418, 451, 606
rate limiting, 456, 605
Ray framework, 80
RBAC (role-based access control), 264, 606
RDMA (remote direct memory access), 17
reactive security measures, 2, 10, 192
real-time monitoring, 458
reasoning models, 34. ***See also*** **CoT (chain-of-thought) reasoning models**
recommender systems, data poisoning attacks, 75
redundancy and replication, 117–118, 605
regulations, 214–215
AI, 182–183
AI privacy, 272–273
AI security, compliance, 230–232
compliance telemetry, 273
ethical AI, 47
monitoring, 455
organizational dynamics, 152
penalty frameworks, 289–291
regulatory readiness, 605
reinforcement learning, 32, 452, 605
removing data points from trained models, 185
resilience, 295, 613–614
adaptive, 533
post-quantum, 197–198
responsible AI, 29, 191–192, 196, 516, 529–530, 606
core principles, 29
frameworks, 46
actionable recommendations, 47
broad principles, 47
regulations, 47
policies, 30
Responsible AI Champions Network, 192
Responsible AI Committee, 29–30
Responsible AI Framework, 29–30
Reversing Labs, "Software Supply Chain Security Report", 460–461
"right to be forgotten", 276

risk
kinetic, 430
third-party. *See* third-party risk
risk assessment, 211
AI system, 228
third-party, 305
access management and user controls, 307
AI vendors, 313–314
application security/ development, 309
asset management and media handling, 307
business continuity, 310
cloud security, 310–311
cryptographic controls, 308
data center operations, 311
data protection and general information security, 305–306
data protection and mobile security, 306
disaster recovery, 310
governance and compliance, 310
human resources management, 306–307
identity and access management, 320–321
incident management, 309–310
ODCs (offshore delivery centers), 311
operations and network security, 308–309
physical and environmental security, 308
policy and governance, 319–320
policy enforcement, 323
supplier relationships/vendor management, 309
vulnerability management, 321–322
risk classification policy, 120–121, 606
risk management, 187–188, 236. ***See also*** **threats**
agentic AI, 264
AI, 36
ISO/IEC 23894, 30
LLMs, 33
supply chain, 246–247
third-party, 303
risk matrix, 228–229, 606
RLHF (reinforcement learning from human feedback), 45
roadmap
AI security readiness, 234
Zero Trust, 153, 234
capability alignment, 154
implementation, 154
risk and regulation, 153–154
robotics, 259
robustness, model, 212–214
RoCEv2 (Remote Direct Memory Access over Converged Ethernet version 2), 469–470, 471–472, 606
Roddenberry, Gene, 259, 261–262
rollback, distributed AI, 287
rule-based generative systems, 403–404
Rules File Backdoor, 462
runtime enforcement, 606

S

Salt Typhoon, 157–158
Santos, Omar, 270
SBoM (software bill of materials), 39, 81–83, 111, 286, 318, 463–464, 607. ***See also*** **supply chain**
scalable AI systems, 198–199
scale, offensive AI, 488
scaling a network, 364
scope, offensive AI, 487, 488–489
scores, trust, 208, 264
SDLC (software development lifecycle), 54, 87, 453
SDN (software-defined networking), 17
SecML, 88
secure by default, 87
secure by design, 86–87
secure enclaves, 268, 606
Secure Software Development Framework, 87, 190–191
security. ***See also*** **encryption; organizational AI security readiness; Zero Trust**
agentic AI, 207–209, 427–428
agents, 206
AI, 528–529, 609–611
AI model, 265–266, 461–462
AI-ready environments, 386–387
-aware culture, 216
building into AI development lifecycle, 189–191, 236–237
challenges in agentic and embedded AI, 262
ChatAI, 414–416
chatbots, 206
comparing traditional architectures with Zero Trust, 109–110
compliance, 108
data, 39–40, 305–306
data classification, 174
defense in depth, 57
email, 138–139
fine-tuned GenAI models, 416–418
-first culture, 234–235, 606
as foundational network design constraint, 368–369
gap analysis, 229
hands-on exercises, 239
integrating in AI governance and strategy, 211–212
LLMs (large language models), 452
MCP (Model Context Protocol), 424–425
networks, third-party risk, 308–309
policies, 235–236
proactive measures, 191–192
business and operational benefits, 193
challenges and considerations, 199–200
cost savings, 194–196
future-proofing AI systems, 197–198
improving trust and compliance, 196–197
moving from software development to AI application development, 201–203
quantitative and qualitative metrics, 193
scalability and adaptability of secure AI, 198–199
securing AI application development, 200–201
value propositions, 193–194

RAG (retrieval-augmented generation), 416–418
reactive measures, 192
-resilient infrastructure, 215–216
software, 201–202
security monitoring, AI, 11
security posture, 468–469
security readiness assessment, 226–227, 606
security-by-design, 606
segmentation, 121–122, 143–144, 246, 606
AI-aligned, 356
macro-, 121
micro-, 17, 108, 121, 278
Zero Trust, 264
self-adaptive AI models
containment, 265
monitoring, 265
self-auditing AI ecosystem, 265, 606
sensitivity labels, 178, 180–181
sensors, as an attack vector, 157–158
Services as Code
applying Zero Trust, 340
description, analysis, and recommendations, 339–340
key insights and business implications, 340
service-to-service identities, 606
shadow AI, 606
shallow learning, 31
SHAP (SHapley Additive exPlanations), 41–42, 245
shared responsibility, 439–440, 531
short-lived tokens, 332–333, 607
side-channel attacks, 491, 492–493, 494, 607
SIEM (security information and event management), 11, 148, 464–465, 606
signature-based detection, 2, 10
skills for the AI era, 44–45
SLA (service-level agreement), 601
SMART goals, 229
smishing, 607
SMPC (secure multi-party computation), 271
Snyk, 461–462
SOAR (security orchestration, automation, and response), 141–142
SOC (security operations center), 458
social engineering, 9, 499–502
software development, 30
AI-native, 160–161
CI/CD (continuous integration/continuous deployment), 40–41, 191, 203, 204–205, 461–462
DevOps, 191, 203–204
DevSecOps, 205–206
lifecycle, 30, 87, 453
Python, 460–461
transitioning to AI application development, 201–203
software security, 201–202
"Software Supply Chain Security Report", 460–461
SolarWinds incident, 80, 161, 462
sophistication, offensive AI, 487, 502–503
sovereignty-aware Zero Trust, 258
SP 800–207. *See* NIST SP 800–207
spam, 32
spam filtering, 9
spear phishing, 607

spectral signature analysis, 76
speed, offensive AI, 487
sponge attack, 90, 607
Spotify, 187–188
SSDs (solid-state drives), 16
stakeholder/s
engagement and ownership, 226
trust, 295
state-sponsored network compromise campaigns, 158–160, 161
steganography, 457
storage systems
AI-ready environments, 16–17, 385
immutable storage, 470
PQC-based encryption, 271–272
strategic data, 256–257
STRIP, 76
supervised learning, 10, 32
supply chain. ***See also*** **third-party risk**
as an attack vector, 157
BoM methodologies, 81–83
namespace reuse attack, 343
risk management, 246–247
vendor-supplied AI, securing, 314–316
core components, 316–317
Zero Trust foundations, 317–318
vulnerabilities, 80–81
vulnerability management, 460–462
supply chain attacks, 80–81, 111, 521, 607
sustainability, AI-ready environments, 390, 607
Sweden, welfare system, 290
sycophancy, 452
synthetic data, 92–94
SynthID, 267

T

tailored training programs, 607
targeted poisoning attack, 74
targets, offensive AI, 489
technical debt, 362, 607
technical infrastructure and tooling evaluation, 232–233
technical requirements, AI-ready environments, 358–359
TEEs (trusted execution environments), 267–268, 272, 608
telemetry, 35, 607
compliance, 273, 276
as trust, 278
TensorFlow, 461–462
testing, 88, 200, 524–525
theft
intellectual property, 186, 188
model, 207, 214
training data, 291–292
third-party risk, 2, 303–304, 305, 531
access management and user controls, 307
AI supply chain, securing, 314–316
core components, 316–317
Zero Trust foundations, 317–318
AI third-party dependencies and risk amplification, 331–332
AI vendors, 313–314
application security/development, 309
asset management and media handling, 307
business continuity, 310
cloud security, 310–311

cryptographic controls, 308
data center operations, 311
data protection and general information security, 305–306
data protection and mobile security, 306
disaster recovery, 310
governance and compliance, 310
human resources management, 306–307
incident management, 309–310
ODCs (offshore delivery centers), 311
operations and network security, 308–309
physical and environmental security, 308
PQC (post-quantum cryptography), 334–338
questionnaire
analytics and continuous monitoring, 323–324
how to use, 324–326
identity and access management, 320–321
policy and governance, 319–320
policy enforcement, 323
section summary, 327–328
vulnerability management, 321–322
supplier relationships/vendor management, 309
vector databases in Zero Trust AI architectures, 332–334
Thomas, Daryl, 614
threat hunting, 11
threat intelligence, 11, 134–135, 148–149, 607
threat landscape
AI, 206–207, 519
monitoring changes in, 455–457
threat modeling, 71–73, 519–520, 527–528
threats
AI, 107
embodied AI, 157–158
Three Laws of Robotics, 259
TLS 1.3, 334, 343, 607
tokenization, 449, 607
tools
Adversarial Robustness, 88
AI, 36, 39
assessment, 232–233
automation, 40–41
data classification, 176–177
TPM (Trusted Platform Module), 268
TPUs (tensor processing units), 15, 16, 607
traffic visibility, 149–150, 607
training, 40–41
adversarial, 87, 93–94, 213
AI, 17
bias, 45
considerations for data used for, 184–185
data theft, 291–292
employee, 233, 237–238
exposure risks and mitigation, 188–189
privacy risks, 185–187
removing data points, 185
tailored, 607
user awareness, 503
training and awareness programs, 608

transparency, 608
AI, 242–243
blockchain, 49
organizational security, 236
XAI, 41–42
Trojan attack, 75–76
trust, 1, 110, 196–197, 260
behavioral, 261
cognitive, 261
dependencies, 109
in embodied AI, 155
implicit, 1, 107
low, 37
modeling, 3–4
operational, 261
scores, 208, 264
stakeholder, 295
telemetry as, 278
trust gap, 608
trustworthy AI, 40, 44–45, 175, 608
TTPs (tactics, techniques, and procedures), 67, 89
Turing test, 8, 407

U

unsupervised learning, 10, 32
untargeted poisoning attack, 74
user awareness training, 503

V

value propositions, 193–194
vector databases
risk assessment, 332–334
Zero Trust, 341
vendors, 430. ***See also*** **supply chain**
AI supply chain, securing, 314–316
core components, 316–317
Zero Trust foundations, 317–318
continuous monitoring of performance, 311–313, 329–331
risk assessment, 305
access management and user controls, 307
AI vendors, 313–314
application security/ development, 309
asset management and media handling, 307
business continuity, 310
cloud security, 310–311
cryptographic controls, 308
data center operations, 311
data protection and general information security, 305–306
data protection and mobile security, 306
disaster recovery, 310
governance and compliance, 310
human resources management, 306–307
identity and access management, 320–321
incident management, 309–310
ODCs (offshore delivery centers), 311
operations and network security, 308–309
physical and environmental security, 308
policy and governance, 319–320
policy enforcement, 323

supplier relationships/vendor management, 309
vulnerability management, 321–322
verifying user identities, 245
version control, 214, 287
vishing, 488, 608
VPN (virtual private networks), 141
vulnerabilities, supply chain, 80–81. ***See also*** **CVE (Common Vulnerabilities and Exposures)**
vulnerability management, 460–462, 608
third-party risk, 321–322
Zero Trust, 5
Vulnerability Management domain, 131, 133–134
database change control, 135
endpoint protection and encryption, 132–133
malware prevention and detection, 133
vulnerability scanning and threat intelligence, 134–135
vulnerability scanning, 134–135

W

watermarks, model, 266–267
welfare system, Sweden, 290
white-box scenario, 72
wild patterns, 73–74. ***See also*** **adversarial examples**
workflows, 33, 323–324
workloads
AI, 14–15, 382–383, 596
AI/ML, 377–378
microsegmentation, 278
workshops, Zero Trust, 150–151
capability alignment, 151–152
implementation, 151
WormGTP, 495, 608

X

XAI
counterfactual explanations, 245
feature importance analysis, 244
interpretability, 41–42
LIME (local interpretable model-agnostic explanations), 41–42, 245
readiness, 244–245
SHAP (SHapley Additive exPlanations), 245
XAI (explainable AI), 41–42, 466, 600

Y-Z

Yudkowsky, Eliezer, 539
Zero Trust, 1, 2–3, 21–22, 107, 521–523, 608
for agents and reasoning models, 35
for AI, benefits, 107–108
AI agent guardrails, 55
allow known good, block everything else, 55–56
fail closed, not open, 57
log everything, 56–57
make readable policies, 56
use multiple layers, 57–58
for AI systems, 260–262
Analytics domain, 144
APM, 144
asset monitoring and discovery, 150
auditing, logging, and monitoring, 145–146
change detection, 146

network threat behavior analytics, 147–148
SIEM, 148
threat intelligence, 148–149
traffic visibility, 149–150
applying across AI data pipelines, 277–278
applying to Services as Code, 340
architect checklist, 329
capabilities, 111–112
compliance, 108
core principles, 410
for data centers, 469–471
decision-maker checklist, 328–329
de-perimeterization, 4
domains, 5
edge computing, 20
Enforcement domain, 135–136
CASB, 136
DDoS attacks, 136–137
DLP, 137–139
file integrity monitoring, 142–143
firewall, 139
IPS, 139–140
proxy, 140–141
segmentation, 143–144
SOAR, 141–142
VPN/RA, 141
expanded capability set, 4–5
for GenAI, 410, 431–439
Identity domain, 122
AAA (authentication, authorization, and accounting), 122–123
asset management, 127–128
CA (Certificate Authority), 123–124
CMDB (configuration management database), 128–130
IP schemas and addressing, 130–131
MAF (multifactor authentication), 127
NAC (network access control), 124–125
PAM (privileged access management), 126–127
provisioning, 125–126
integration of PQC within data protection models, 279
microsegmentation, 608
NIST SP 800–207, 3
organizational dynamics, 152
capability alignment, 153
implementation, 152–153
risk and regulation, 152
origins, 3–4, 108–110
Policy and Governance domain, 112–113
business continuity, 118–119
change control, 113–114
data governance and encryption, 114–115
data retention, 115–116
disaster recovery, 119–120
QoS, 116–117
redundancy and replication, 117–118
risk classification policy, 120–121
segmentation, 121–122
PoV (proof of value), 108
principles applied to AI security readiness, 245–246, 503–505

roadmap, 153
capability alignment, 154
implementation, 154
risk and regulation, 153–154
segmentation, 264
sovereignty-aware, 258
strategies for AI compliance and enforcement, 276–277
technical practitioners checklist, 329
third-party risk questionnaire, 317–318
versus traditional security architecture principles, 109–110
for vector databases, 341
Vulnerability Management domain, 131, 133–134
database change control, 135
endpoint protection and encryption, 132–133
malware prevention and detection, 133
vulnerability scanning and threat intelligence, 134–135
workshops, 150–151
capability alignment, 151–152
implementation, 151
Zero Trust Architecture, **317–318**
zero-shot, 34
zero-width attacks, 202
ZKPs (zero-knowledge proofs), 272, 608
ZTA (Zero Trust Architecture), 6, 608