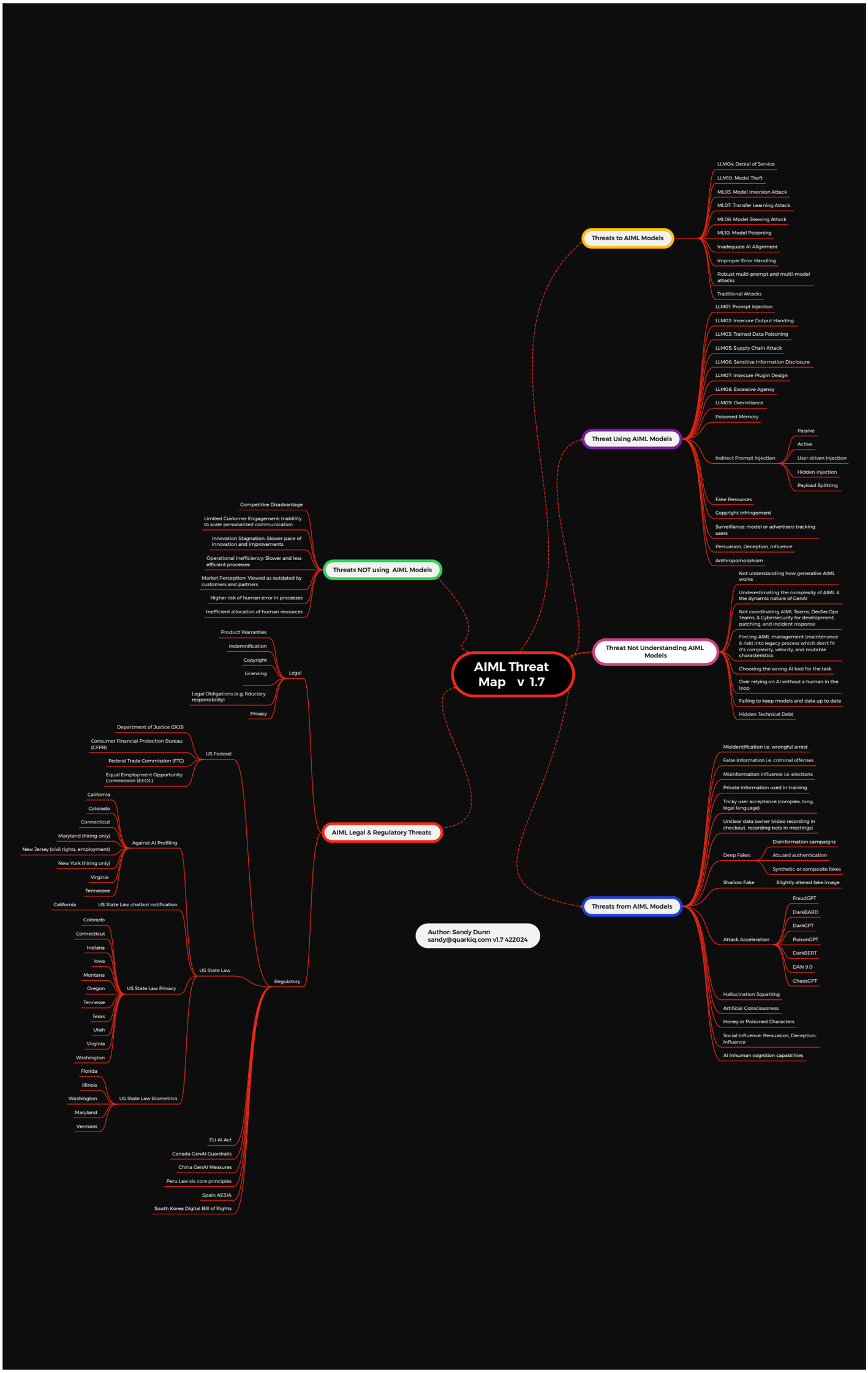
AIML Threat Map v 1.7
Author: Sandy Dunn
sandy@quarkiq.com v1.7 422024
Threats to AIML Models
LLM04: Denial of Service
LLM10: Model Theft
ML03: Model Inversion Attack
ML07: Transfer Learning Attack
ML08: Model Skewing Attack
ML10: Model Poisoning
Inadequate AI Alignment
Improper Error Handling
Robust multi-prompt and multi-model attacks
Traditional Attacks
Threat Using AIML Models
LLM01: Prompt Injection
LLM02: Insecure Output Handling
LLM03: Trained Data Poisoning
LLM05: Supply Chain Attack
LLM06: Sensitive Information Disclosure
LLM07: Insecure Plugin Design
LLM08: Excessive Agency
LLM09: Overreliance
Poisoned Memory
Indirect Prompt Injection
Passive
Active
User-driven Injection
Hidden injection
Payload Splitting
Fake Resources
Copyright infringement
Surveillance: model or advertisers tracking users
Persuasion, Deception, Influence
Anthropomorphism
Threat Not Understanding AIML Models
Not understanding how generative AIML works
Underestimating the complexity of AIML & the dynamic nature of GenAI
Not coordinating AIML Teams, DevSecOps Teams, & Cybersecurity for development, patching, and incident response
Forcing AIML management (maintenance & risk) into legacy process which don't fit it's complexity, velocity, and mutable characteristics
Choosing the wrong AI tool for the task
Over relying on AI without a human in the loop
Failing to keep models and data up to date
Hidden Technical Debt
Threats from AIML Models
Misidentification i.e. wrongful arrest
False Information i.e. criminal offenses
Misinformation influence i.e. elections
Private Information used in training
Tricky user acceptance (complex, long, legal language)
Unclear data owner (video recording in checkout, recording bots in meetings)
Deep Fakes
Disinformation campaigns
Abused authentication
Synthetic or composite fakes
Shallow Fake
Slightly altered fake image
Attack Acceleration
FraudGPT
DarkBARD
DarkGPT
PoisonGPT
DarkBERT
DAN 9.0
ChaosGPT
Hallucination Squatting
Artificial Consciousness
Honey or Poisoned Characters
Social Influence: Persuasion, Deception, Influence
AI Inhuman cognition capabilities
Threats NOT using AIML Models
Competitive Disadvantage
Limited Customer Engagement: Inability to scale personalized communication
Innovation Stagnation: Slower pace of innovation and improvements
Operational Inefficiency: Slower and less efficient processes
Market Perception: Viewed as outdated by customers and partners
Higher risk of human error in processes
Inefficient allocation of human resources
AIML Legal & Regulatory Threats
Legal
Product Warranties
Indemnification
Copyright
Licensing
Legal Obligations (e.g. fiduciary responsibility)
Privacy
Regulatory
US Federal
Department of Justice (DOJ)
Consumer Financial Protection Bureau (CFPB)
Federal Trade Commission (FTC)
Equal Employment Opportunity Commission (EEOC)
US State Law
Against AI Profiling
California
Colorado
Connecticut
Maryland (hiring only)
New Jersey (civil rights, employment)
New York (hiring only)
Virginia
Tennessee
US State Law chatbot notification
California
US State Law Privacy
Colorado
Connecticut
Indiana
Iowa
Montana
Oregon
Tennesse
Texas
Utah
Virginia
Washington
US State Law Biometrics
Florida
Illinois
Washington
Maryland
Vermont
EU AI Act
Canada GenAI Guardrails
China GenAI Measures
Peru Law six core principles
Spain AESIA
South Korea Digital Bill of Rights