CNCF TAG-Runtime and Security TAG (STAG) Collaboration Project
AI Security Whitepaper Meeting Notes

Cloud Native AI Security Whitepaper - Task Force | Meeting Notes 📅

: Runtime TAG Charter
Cloud Native AI WG Charter

: Join CNCF | Add #wg-artificial-intelligence

: CNCF CoC

📅: add CNCF calendar, and check onboarding and meeting “Cloud Native AI Working Group”

🕗: 8 a.m. PT | 11 am. ET (1 hour)

: Zoom (CNCF tag runtime, pw: 77777)

📧📪: wg-artificial-intelligence@lists.cncf.io (Mailing List) We
* Join the Mailing List to receive the calendar meeting invite.

: Past meetings - CN AI WG
Past Meetings (TAG-Runtime channel)

GitHub Tracker - https://github.com/cncf/tag-runtime/issues/177

Meeting Etiquette 👔

Proposing Agenda Items… (Only for CN AI Security Whitepaper)

- **Anyone** can propose discussion points, though only related to CN AI Security Whitepaper, not general AI issues which are to be discussed in AI WG meetings.
- Follow the CNCF Code of Conduct
- *The WG can also schedule meetings on other days per request. Contact the leads or* ***#wg-artificial-intelligence*** *on the CNCF Slack for more information.*

Planned Meetings 📅

Meetings are on the 1st and 3rd Friday of the month.

Meetings

Mar 21, 2025 8:00 AM PDT

Host: Deep Patel

NoteTaker:

- Scribe1 -
- Scribe2 -

Attendance:Please Add yourself

- Josh Halley (Cisco)
- Deep Patel
- Joel Roberts

Agenda:

- Discuss comments -
- AI Agents

Meetings notes:

Editable AI diagram link: https://app.excalidraw.com/l/7CjaPSlCsjJ/8tfeDvhlqEc

https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/

We need to go into CNAPP details whole lot, instead simply focus on what we have and what can be accomplished.
CASB ?? -

Ron - wanted to mention training resources available
Victor - https://training.linuxfoundation.org/certification/certified-kubernetes-security-specialist/

Josh is going to send link to what we have in regard to - AI powered threat detection

Feb. 21, 2025 08:00 GMT-8

Host:

NoteTaker:

- Scribe1 -
- Scribe2 - Josh Halley

Attendance:Please Add yourself

- Josh Halley, Cisco
- Nina Polshakova, Solo.io

Agenda:

- Discuss comments -
- AI Agents

Meetings notes:

Feb 7, 2025 8:00 AM PST

Host: Deep Patel

NoteTaker:

- Scribe1 -
- Scribe2 - Deep Patel

Attendance:Please Add yourself

- Deep Patel, Cisco
- Nimisha, Confluent
- Aonan Guan, Wyze Labs
- Josh Halley, Cisco
- Gerry Seidman, AuriStor
- Thalia Hooker, Red Hat

Agenda:

- Discuss comments -
- AI Agents
- Confidential computing

Meetings notes:

jan. 17, 2025 08:00 GMT-8

Host: Deep Patel

NoteTaker:

- Scribe1 - Pedro Ignácio
- Scribe2 - Deep Patel

Attendance:Please Add yourself

- Pedro Ignácio, Itaú Unibanco
- Andreas Happe, Technical University of Vienna / OWASP
- Deep Patel, Cisco
- Sundar Nadathur
-

Agenda:

- Discuss comments - tooling, model issues, confidential computing sections
- Personas
- Review on comments made to the document

-

Meetings notes:

==========

Jan 3, 2025 8:00 AM PST

Host: Deep Patel

NoteTaker:

- Scribe1 -
- Scribe2 - Deep Patel

Attendance:Please Add yourself

- Deep Patel - Cisco
- Nimisha - Confluent
- Aonan Guan - Wyze Labs

Agenda:

- Hello to new folks (intros)
- Followup comments
- How to proceed to meet Feb end target
- Work distribution/assignment

Meetings notes:

==========

Dec 6, 2024 8:00 AM PST

Host: Deep Patel

NoteTaker:

- Scribe1 - Deep Patel

Attendance:Please Add yourself

- Deep Patel - Cisco
- Boris Kurktchiev - Looking for a Home
- Hubert Siwik
- Nimisha Mehta - Confluent
- Pedro Ignácio - Itaú Unibanco

Agenda:

- Hello to new folks (intros)
- (NM) Are we planning to cover something along the lines of “application architectures” - eg. with retrieval & agents? This probably introduces new attack vectors
- STAG advised phased approach for proceeding on the paper
- Discuss - security issues, effects
- Finally assign issues mentioned in 5 basic questions

Meetings notes:

Topic (scribe -)
Doc links have changed and 3 docs related to this paper are available at: https://drive.google.com/drive/folders/1luVtzJRlkekc2_5ovV0RjvIPpQTcbWot Document where comments are requested is “CN AI Security Whitepaper Research & Resources”
Target audience, goals and scope chapters are discussed and participants are expected to provide comments or edit the section.
Participants need to familiarize themselves with concepts discussed in below papers: 1. Cloud Native Security Whitepaper (Version 2) 2. CNCF Cloud Native AI Whitepaper 3. Building Trust: Foundations of Security, Safety and Transparency in AI
Participants need to go through categorization of personas and provide comments. These could be seen in the context of AI personas we have defined in the document #2 above.
Finally five basic questions mentioned in the doc and highlighted in yellow color, and their subtopics

need to be chosen by participants to add details and/or list findings. After our discussion in the next meeting, these can be included in the final draft.
Next meeting is on Fri, Dec 20th 8 am PT / 5 pm CET.

Nov 15, 2024 8:00 AM PST

Host: Deep Patel

NoteTaker:

- Scribe1 - Deep Patel

Attendance:Please Add yourself

- Deep Patel - Cisco
- Hubert Siwik
-

Agenda:

- Hello to new folks (intros)
- Personas - whether categorization and making it fewer makes sense
- Whether to capture breach cases - as we know today
- Working meeting to collect points on various topics captured in the doc

Meetings notes:

Topic (scribe -)
Attendance - very thin :) because of KubeCon NA. Welcome Hubert to the team! Not much was discussed in this meeting because of lack of quorum.

Nov 1, 2024 8:00 AM PDT

Host: Deep Patel

NoteTaker:

- Scribe1 - Ronald Petty

Attendance:Please Add yourself

- Ronald Petty - RX-M
- Nimisha Mehta - Confluent
- Sudhanshu Prajapati - InfraCloud
- Mehrin Kiani - Protect AI
- Adel Zaalouk - Redhat
- Deep Patel - Cisco

Agenda:

- Hello to new folks (intros)
- Template like ???
- Are legal and AI safety/ethics issues under this whitepaper scope???
- CN AI Security risks - scenario
- CN AI Security - Personas:
    - Engineers - Cloud, AI-Ops, Data
    - Administrators/Maintainers
    - Compliance
- AI for security in scope? [nimisha]
- KubeCon planning [rp]
    - Get others interested in this project, reduce duplicate efforts if other groups are doing similar things, just get the word out, get feedback, etc.

Meetings notes:

Topic (scribe - rp)
(DP) Intro, overview of meeting agenda (DP) Looking for a document formatting process (Google->PDF->Web->...); DP - How to work templates? AZ - use it, it will evolve into paper (No need to create new paper) – move notes to agenda
(DP) Should we add / include legal/ethics/etc. RP - Yes, as a guide but not in detail, this is a technical paper first and foremost; AZ - Agree; DP - Agree

(DP) Reviewing personas to help understand the paper (e.g. usecase-personas); AI world seems to have even more personas (e.g. forensic engineer, etc.); AZ - need to focus, security subset (already defined in TAG-SECURITY link?); there might be new personas that emerge but at this time seems AI Security Personas are same as existing; VL - can persona can be lite touch here because its more on tech; DP - yes; VL - OpenSSF has personas in doc as well; AZ - have dimensions (e.g. persona) then arrange paper around that; AZ - persona can have multiple roles; AZ - I worry about divergence of personas / roles across papers; DP - lots of docs, lots of differences, best effort should be tried though but not perfect; DP - will map TAG-RUNTIME personas to CN AI Security - Personas and maybe split out later
(DP) Add Uses Case below; detail “where the problem” is; AZ - what is the format?; DP - describing use case format; DP - Reviewing Use Case 1 as an example; Issues are general and/or AI specific; DP - which issues to include (AI or not); AZ sharing screen – showing platform tech levels (hw, network, software, workloads); OWASP - top 10 security concerns for LLM; AZ - e.g. guardrails are “new”, how does that work in CN AI systems; AI Gateways help or hurt? (how); app layer, platform layer, etc.; general, semi-specific; very-specific
(DP/AZ) What systems/components/tools are being used related to AI; Can we reuse existing practices?; What is new? RP - What about AI for CN? AZ - E.g. Agents help extract signals, a large ecosystem, do we cover that? DP - get experts in AI/ML to comment (RP - non-expert is ok, expert is better!)
(DP) Summarize personas and cases for next case (help on time); RP/DP let's have meeting on 15th (will represent at KubeCon)

I have mapped the TAG-RUNTIME Personas to suggested Cloud Native AI Security Personas

---

Engineers - Cloud, MLOps, Data

1. **Data Engineers**
    - **Map to: Engineers - Data**
    - **Reason:** Data Engineers work directly on data pipelines, quality, and integrity, which is essential for maintaining secure, reliable data in AI systems.
2. **Data Scientists**
    - **Map to: Engineers - Cloud and MLOps**
    - **Reason:** Define the problem, explore data, select appropriate models/algorithms, train models, and evaluate model performance.
3. **MLOps Engineers**
    - **Map to: Engineers - MLOps**
    - **Reason:** MLOps Engineers handle operational aspects of ML models, ensuring they run securely in production environments, bridging the gap between models and deployment.
4. **AI Engineers**
    - **Map to: Engineers - MLOps**

    - **Reason:** Handle model selection, fine-tuning, and integration of AI components into larger systems.
5. **AI Application Developers (Coders)**
    - **Map to: Engineers - Cloud**
    - **Reason:** They integrate AI into applications, focusing on secure, real-world implementation. Their work directly impacts how AI interacts with cloud-native systems.
6. **Platform Engineers**
    - **Map to: Engineers - Cloud**
    - **Reason:** By managing developer platforms, they ensure secure and compliant pipelines and help AI developers access secure, pre-approved resources.

Administrators/Maintainers

1. **Site Reliability Engineers (SREs)**
    - **Map to: Administrators/Maintainers**
    - **Reason:** SREs are responsible for maintaining uptime and reliability, implementing automated security checks, and monitoring AI systems in production.
2. **Security Architect/Engineers**
    - **Map to: Administrators/Maintainers**
    - **Reason:** They implement protective measures and address vulnerabilities, ensuring AI infrastructure and data remain secure.
3. **Hardware Architects**
    - **Map to: Administrators/Maintainers**
    - **Reason:** Hardware Architects support the infrastructure, ensuring computational efficiency and addressing physical and operational security needs of AI systems.

Compliance

1. **Compliance Officers**
    - **Map to: Compliance**
    - **Reason:** Compliance Officers directly focus on regulatory standards and governance, ensuring AI models meet legal and industry requirements.
2. **AI Ethicists**
    - **Map to: Compliance**
    - **Reason:** AI Ethicists ensure responsible and compliant development practices, advocating for ethical design and deployment across all stages of AI.
3. **AI Safety Researchers**
    - **Map to: Compliance**

    - **Reason:** These researchers work on responsible AI usage, aiming to prevent misuse and unintended consequences, which aligns closely with compliance.

Additional Personas that I couldn't fit in current CN AI Security Personas

Prompt Engineers

- **Role:** Specialize in crafting effective prompts for generative AI models.
- **Quotes/Mottos:**
    - "I speak the language of AI."
    - "I unlock the creative potential of language models."

AI Product Managers

- **Role:** Oversee the development of AI products from conception to launch.
- **Quotes/Mottos:**
    - "I guide AI products from idea to impact."
    - "I ensure our AI solutions meet market needs."

AI Researchers

- **Role:** Explore AI techniques that may not be immediately applicable to products.
- **Quotes/Mottos:**
    - "I push the boundaries of what AI can do."
    - "I explore today what will be mainstream AI tomorrow."

Specific security risks unique to cloud-native AI deployments and the potential impact

Case 1: Data Breach - unauthorized access to data (stored and/or during processing)

- Issues:
    - Training data exposure
    - Generated data exposure
- How do these issues happen?
    - Insufficient protection at rest - lack of encryption, poor at-rest encryption, key compromise
    - Runtime or at-rest - Identity control, RBAC issues, leak (platform, apps, processing), misconfiguration
    - Platforms/Apps - CVEs

- How is it fixed?

Case 2: Data Poisoning - making training data adulterated to induce bias in AI modeling

- Issues:
    - Improper control to training data - RBAC, write access, buffer overflow, supply chain
    - Inadvertent feedback - RAG type of mechanism
    - Lack of data validation - user input can't be a blind source of truth (thin about SQL injection)
- How do these issues happen?
- How is it fixed?

Case 3: Adversarial Attack - AI model manipulation through adversarial input

- Issues:
    - Bigger problem for Predictive AI
    - Lack of data validation
    - Spurious data sources, data supply chain risk
    - Capitalizing upon known/unknown model and/or training data weakness
    - Case #2 weaknesses apply
- How do these issues happen?
- How is it fixed?

Case 4: Insecure APIs - Cloud native platform and AI application APIs' vulnerabilities

- Issues:
    - API weakness K8s - pods, containers, apiserver, etcd, orchestration, underlying OS vulnerabilities
    - Access token, API keys - mismanagement
    - Lack of of proper encryption - transport, data
    - Connection spoofing/hijack - lack of peer auth
    - AuthZ issues
- How do these issues happen?
- How is it fixed?

Case 5: Supply chain attacks

- Issues:
    - Third party libraries, tools - willful insertion, day1 issues
    - Backdoor (Example: https://pentest-tools.com/blog/xz-utils-backdoor-cve-2024-3094)
    - DoS issues
- How do these issues happen?
- How is it fixed?

Case 6: Misconfigurations

- Issues:

        - Cloud services
        - Network settings
        - Access, RBAC issues
        - Lack of guardrails for data at-rest and in transit
        - Lack of and/or insecure crypto
        - Malicious config change
- How do these issues happen?
- How is it fixed?

Case 7: AI Model - Intellectual property misuse/theft

- Issues:
    - AI model exposed/compromised
    - May lead to app/tools/model attack
    - May cause further supply chain attacks
    - Can be used against - bc of enhanced model
- How do these issues happen?
- How is it fixed?

Case 8: Insider malicious actions

- Issues:
    - Improper AuthN, AuthZ protections
    - Privilege escalation
    - Cmd injection
    - Silent data loss
    - Silent intrusions
- How do these issues happen?
- How is it fixed?

Case 9: Infra/platform attacks (AWS, GCP, Azure….)

- Issues:
    - Wider access
    - Lack of app, data protections
    - Platform vulnerabilities
- How do these issues happen?
- How is it fixed?

Case 10: MLOps, LLMOps issues

- Issues:
    - Tooling vulnerabilities
    - Model Integrity -
    - CI/CD pipeline vulnerabilities
    - Model privacy - Leakage of sensitive data
    - Use of LLMs to generate/insert harmful content or poison the outcome
    - Lack of content filtering

- How do these issues happen?
- How is it fixed?

Case 11: Lack of visibility

- Issues:
    - Who, what, when, how - data usage, generation, login
    - Repudiation issues (Non-repudiation challenges)
    - Auditing
    - Phishing and social engineering attacks
- How do these issues happen?
- How is it fixed?

Case 12: AI prompt based attacks

- Issues:
    - Manipulate prompt - during processing pipeline or during prompt feed
    - Induce hallucination
    - Reverse engineering the system prompt
- How do these issues happen?
- How is it fixed?

Case 13: Malicious code creation

- Issues:
    - To be exploited when code runs
    - Backdoor
    - Compromise integrity
- How do these issues happen?
    - Manipulating the system prompt to bypass code execution-related response.
- How is it fixed?

CN AI Security - Personas

- Engineers - Cloud, MLOps, Data
- Administrators/Maintainers
- Compliance

OR

- Security personas needs to be defined in terms of threat surface a user faces, irrespective of above roles. Examples:
    - End user (GUI/API, finished product (image, video), decision) - human or app or machine - music, games, design, control
    - Cloud Native Platforms and Providers - AWS, GCP, Azure
    - Data provider - agency (NASA, Hospital, Youtube)
    - Data consumer/processor - AI apps and platforms

- Data keeper/storage -
- Housekeeper/Maintainer - day to day admin and operator
- App developers - dev engineer
- Build pipeline provider/maintainer
- Deployers
- Security/Policy Compliance/Auditing people
- Legal framework
- Forensic Engineers
- Trust/Identity provider

==========

Oct 18, 2024 8:00 AM PDT

Host: Deep Patel

NoteTaker: Deep Patel

Attendance:Please Add yourself

- Deep Patel
- Ronald Petty
- Nimsha Mehta
- Sudhanshu Prajapati
- Mehrin Kiani
- Cameron McDougle
- Sunil Ravipati
- Jon Zeolla
- Karan Singh

Agenda:

- Hello to new folks (intros)
- Personas - do we need to keep it the same way we have it in the brainstorming template (https://docs.google.com/document/d/1z1150HQ3kxUuixAWV75ZRf_LyHclo9-PKamhDPBuNEk/edit)
- How to make progress - decide a few things collectively or people need to take a swab themselves and propose.
- Useful docs
-

Notes:

Analyze specific security risks unique to cloud-native AI deployments and the potential impact of breaches.

- Data poisoning - at rest
- Data poisoning - during consumption/processing (platform, pods/services, apps, objects)
- Data Leakage
- Data Misuse
- Data unintended access
- Data - PIIs

Personas:

- https://tag-runtime.cncf.io/wgs/cnaiwg/glossary/
    - These personas are useful, given AI impacts is quite unique to end users, provider, maintainer, administrator, and several such silos. However,
    - Security personas needs to be defined in terms of threat surface a user faces, irrespective of above roles. Examples:
        - End user (GUI/API, finished product (image, video), decision) - human or app or machine - music, games, design, control
        - Cloud Native Platforms and Providers - AWS, GCP, Azure
        - Data provider - agency (NASA, Hospital, Youtube)
        - Data consumer/processor - AI apps and platforms
        - Data keeper/storage -
        - Housekeeper/Maintainer - day to day admin and operator
        - App developers - dev engineer
        - Build pipeline provider/maintainer
        - Deployers
        - Security/Policy Compliance/Auditing people
        - Legal framework
        - Forensic Engineers
        - Trust/Identity provider

Threat modeling - How to, what to consider, how to keep pace with changes in AI tech.

- Data flow
- LLM model itself
- User interface/APIs
- Server processing - Infra, compute, storage, resiliency, Visibility
- Data Sources - public, private, large, limited, raw, curated
-
- STRIDE (spoofing, tampering, repudiation, info disclosure, DoS, Elevation of privilege) - what needs to change for AI?
- DREAD (damage, reproducibility, exploitability, affected users, discoverability) - ??
- Consider the TAG-Security Open and Secure model

CN for AI - This is about support/evolution of CN eco system to work for AI

AI for CN - challenges ??

- Efficient utilization - load balancing, scheduling, intelligent spawning
- Policy enforcement
- Heuristics based defense/analysis
- Visibility/pattern
-

Cryptography

- Identity
- Signatures
- Trust
- Encryption
- PQC
-

Draft design considerations for securing AI workloads, data, and infrastructure in cloud-native environments, including Kubernetes security best practices.

LLM:

- RLHF (Reinforcement Learning with Human Feedback)

Supply Chain - in AI context

- Model
- Training data
- Platform
- Trust

[Jon] Reference architecture potentially? Similar to the TAG-Security supply chain reference architecture

- Suggest we have a “know we’re done when” acceptance criteria statement, and an upper bound (regarding length)
    - [Deep] suggested < 50 pages

AI Security Whitepaper Notes

Laundry list of things (Includes what is listed in the template mentioned in the AI issue#177 **[Draft] Security for/with Cloud Native AI)**

Analyze specific security risks unique to cloud-native AI deployments and the potential impact of breaches.

- Data poisoning - at rest
- Data poisoning - during consumption/processing (platform, pods/services, apps, objects)
- Data Leakage
- Data Misuse
- Data unintended access
- Data - PIIs

Personas:

- https://tag-runtime.cncf.io/wgs/cnaiwg/glossary/
    - These personas are useful, given AI impacts is quite unique to end users, provider, maintainer, administrator, and several such silos. However,
    - Security personas needs to be defined in terms of threat surface a user faces, irrespective of above roles. Examples:
        - End user (GUI/API, finished product (image, video), decision) - human or app or machine - music, games, design, control
        - Cloud Native Platforms and Providers - AWS, GCP, Azure
        - Data provider - agency (NASA, Hospital, Youtube)
        - Data consumer/processor - AI apps and platforms
        - Data keeper/storage -
        - Housekeeper/Maintainer - day to day admin and operator
        - App developers - dev engineer
        - Build pipeline provider/maintainer
        - Deployers
        - Security/Policy Compliance/Auditing people
        - Legal framework
        - Forensic Engineers
        - Trust/Identity provider

Threat modeling - How to, what to consider, how to keep pace with changes in AI tech.

- Data flow
- LLM model itself
- User interface/APIs
- Server processing - Infra, compute, storage, resiliency, Visibility
- Data Sources - public, private, large, limited, raw, curated
-
- STRIDE (spoofing, tampering, repudiation, info disclosure, DoS, Elevation of privilege) - what needs to change for AI?

        - DREAD (damage, reproducibility, exploitability, affected users, discoverability) - ??

CN for AI - This is about support/evolution of CN eco system to work for AI

AI for CN - challenges ??

- Efficient utilization - load balancing, scheduling, intelligent spawning
- Policy enforcement
- Heuristics based defense/analysis
- Visibility/pattern
-

Cryptography

- Identity
- Signatures
- Trust
- Encryption
- PQC
-

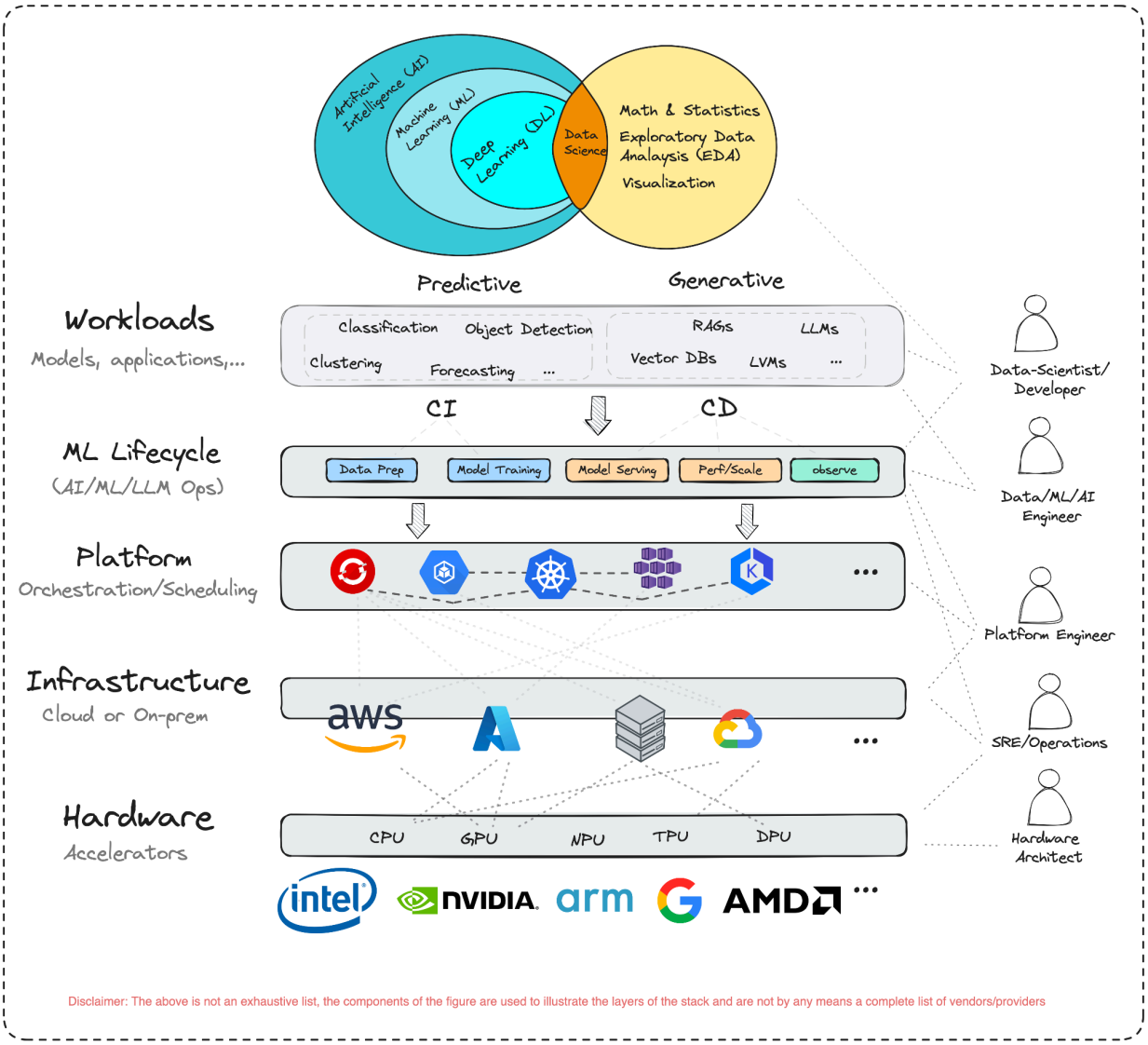
Draft design considerations for securing AI workloads, data, and infrastructure in cloud-native environments, including Kubernetes security best practices.

LLM:

- RLHF (Reinforcement Learning with Human Feedback)

Supply Chain - in AI context

- Model
- Training data
- Platform
- Trust



Solution:

- Basics
  -
  -
- Data protection strategies
  - Encryption at rest and in-transit (training datasets, model artifacts)
  - Federated learning (eg. TensorFlow Federated) to keep sensitive data on-device
  - Detailed, comprehensive Dataset Cards / Data Sheets
  - Reproducible dataset pipelines

    - Content ID-like system to prevent unauthorized and/or malicious injection of copyrighted material
- Infrastructure security
    - Network policies, least-privilege workloads, secure service to service communication (mesh)
- Runtime security
    - Deep visibility into syscalls, network activity and file access patterns
    - Real-time threat detection
    - Secure container runtime
    - WAF-equivalent for AI apps (special token prompt injection)

Explore how cloud-native security tooling/landscape would make AI workloads more secure. If not, explore how to.

- Container & kubernetes security
    - Cilium - advanced network policy, transparent encryption
    - Falco - runtime security monitoring
    - Trivy
- Access control
    - OpenPolicyAgent, Kyverno
- Runtime
    - Tetragon - security observability, runtime enforcement
    - Kata container runtime - enhanced isolation

Provide actionable guidance on securing AI models, data pipelines, and infrastructure, along with recommendations for secure CI/CD pipelines and vulnerability management.

Exploration of emerging trends such as confidential computing, homomorphic encryption, and AI-powered threat detection for cloud-native AI.pop