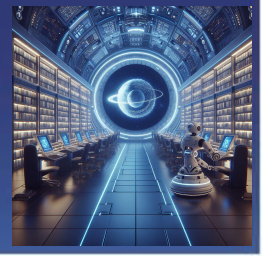
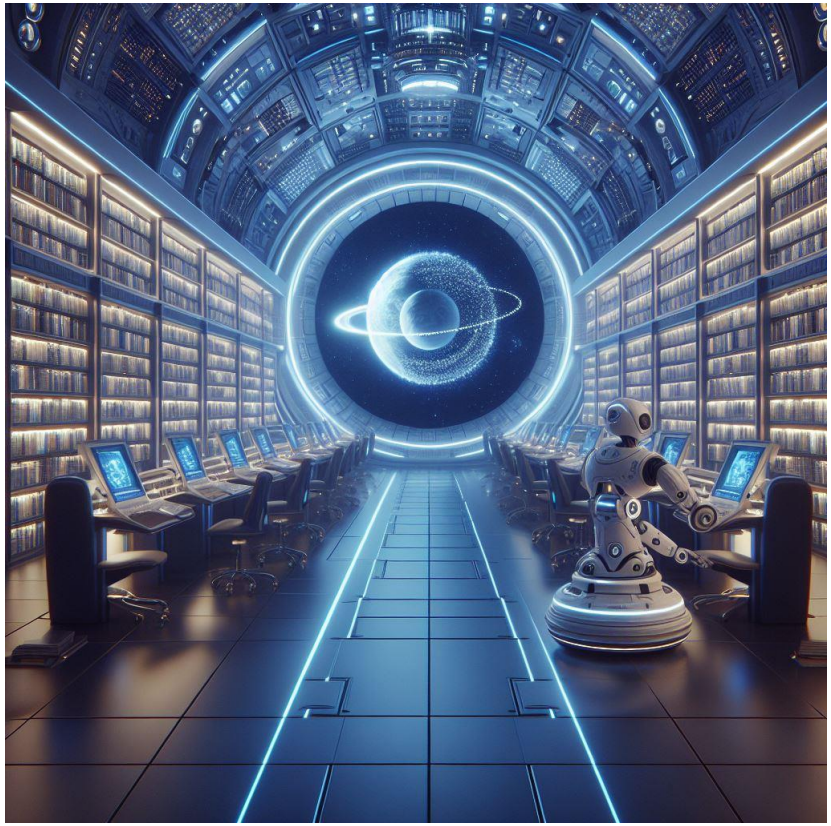


ISVV for AI Systems

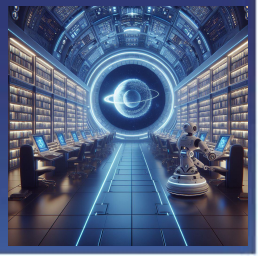


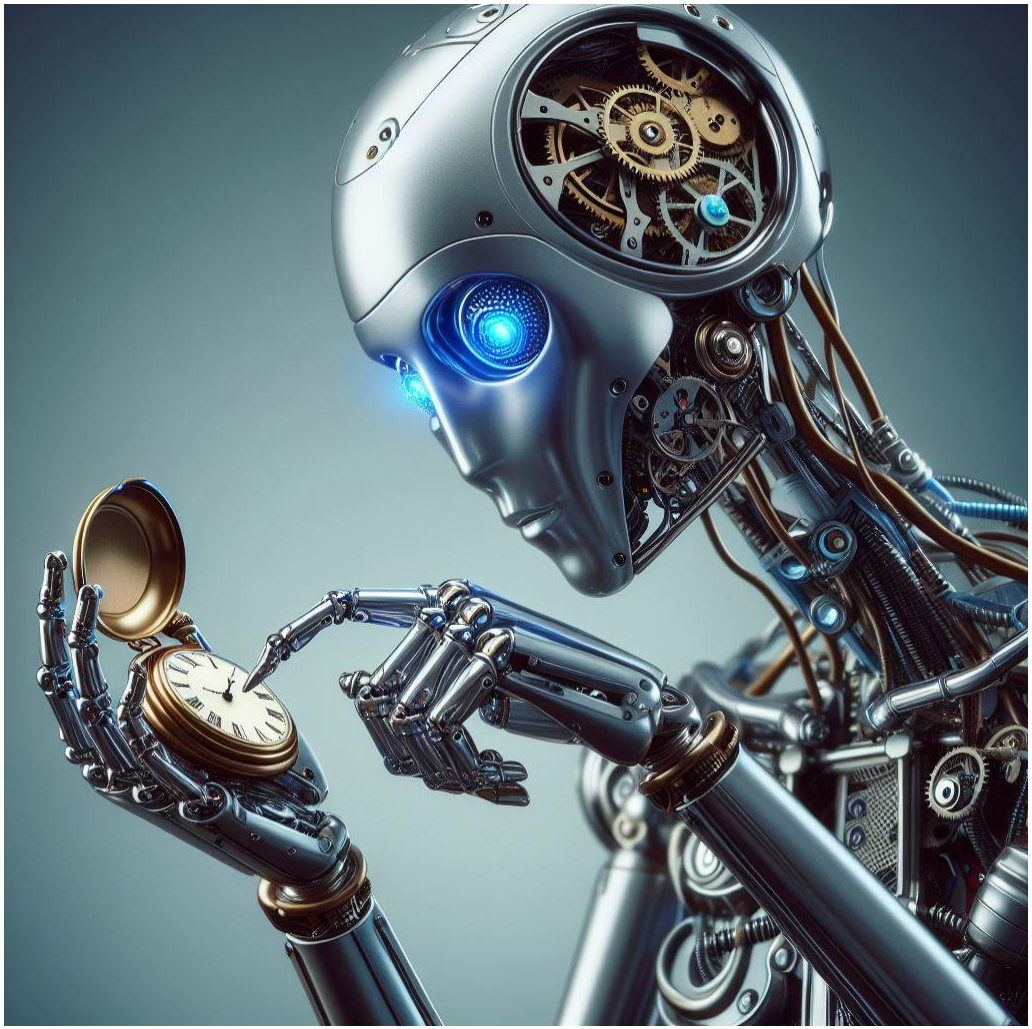
NOTE: most of the images in this presentation have been generated with Microsoft's Image Creator

June 2025

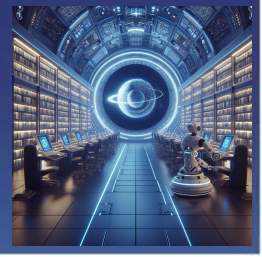
Maurizio Martignano
Spazio IT – Soluzioni Informatiche s.a.s
Via Manzoni 40
46051 San Giorgio Bigarello, Mantova
https://spazioit.com

Agenda



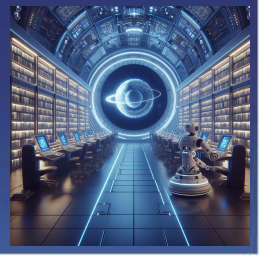


Agenda



- Who am I?
- ISVV ←→ AI
  - Types of AI
  - ISVV Core Principles
  - Common ISVV Phases (Customized per AI Type)
  - Key Differences & Adjustments by AI Type
  - Summary Table: ISVV by AI Type
- Conclusions

Who am I?



Maurizio Martignano – Spazio IT
https://spazioit.com
https://www.linkedin.com/in/mauriziomartignano

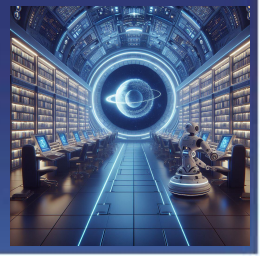
Active since the 1990s (sigh) in Avionics Software Development and Verification
https://spazioit.com/pages_en/sol_inf_en/code_quality_en/

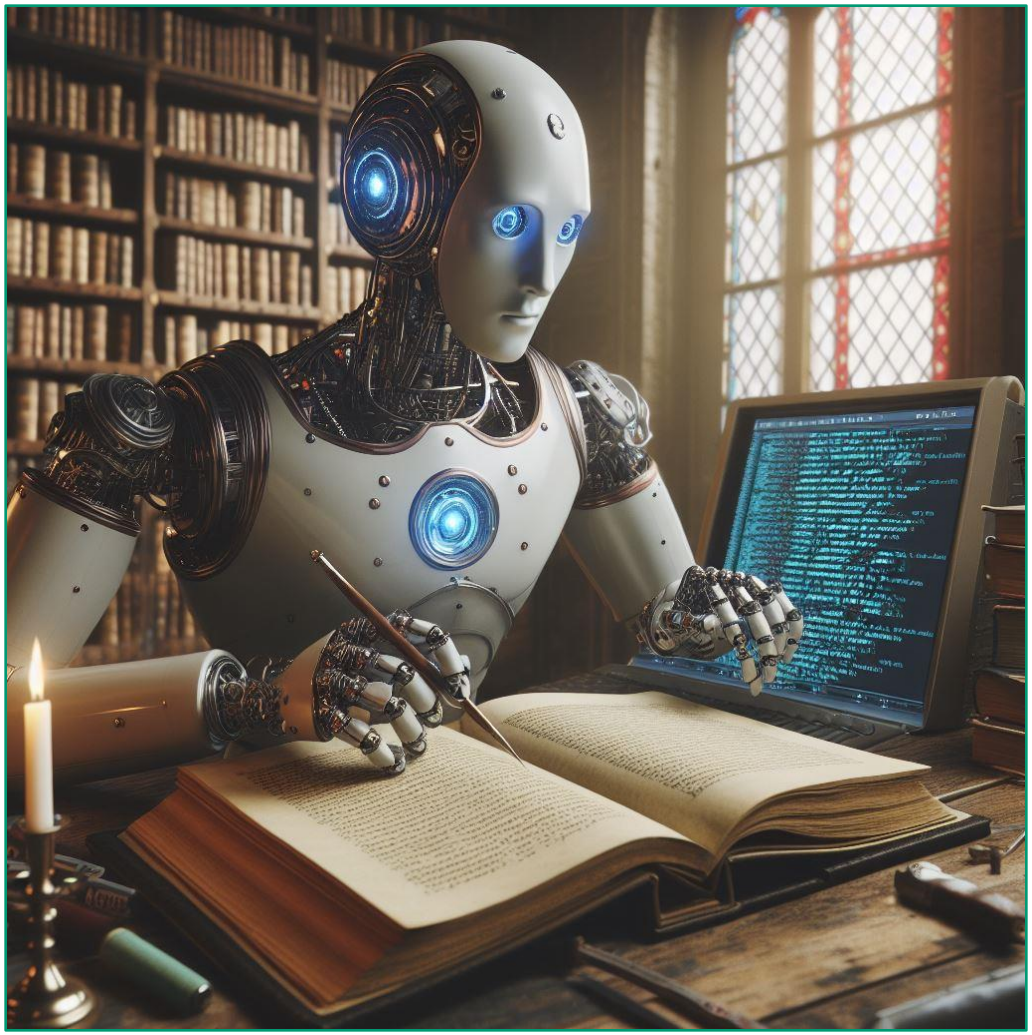
Providing software consultancies in various application domains, e.g. Healthcare, Cybersecurity, Data Protection,…

Not a developer but a **user** of AI Technologies (especially in Healthcare and Software Verification Applications)

AI → ISVV (not covered here)

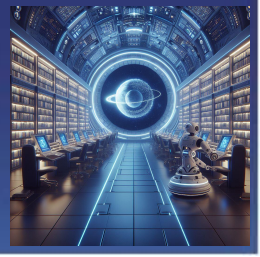
How to use AI to improve the ISVV of traditional non-AI software



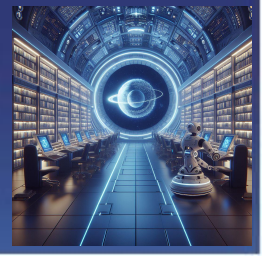


ISVV → AI (covered here)

How to perform ISVV on AI software



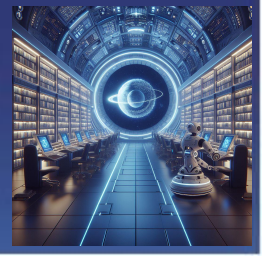
Types of AI



- Applying **Independent Software Verification and Validation (ISVV)** to different types of AI systems
  - **classic AI (e.g., classifiers)**
  - **generative AI (e.g., LLMs)**, and
  - **AI agents (both AI agents and agentic AI, e.g., autonomous agents)**

  requires adapting core ISVV principles to the unique properties of each system.

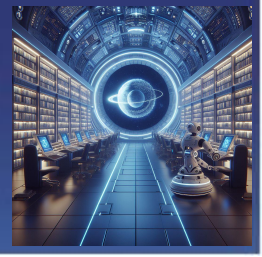
AI Agents vs. Agentic AI



- **AI Agents**
  - AI agents are **task-specific** systems designed to execute predefined actions.
  - They follow **rules, workflows, and instructions** to complete tasks efficiently.
  - Examples include **chatbots, recommendation engines, and automation tools** that respond to user input but do not independently set goals or adapt dynamically.
  - They operate within **fixed boundaries** and require human intervention for adjustments.

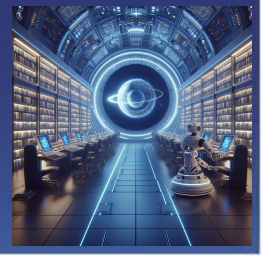
- **Agentic AI**
  - Agentic AI is a **more advanced and autonomous** form of AI that can **sense, decide, and act** without constant human oversight.
  - It **adapts to changing environments**, makes decisions proactively, and refines strategies over time.
  - Unlike AI agents, **Agentic AI can break down complex objectives into smaller tasks**, self-correct, and iterate toward long-term goals.
  - It is **goal-driven**, meaning it can **plan, execute, and optimize** workflows dynamically.

ISVV Core Principles



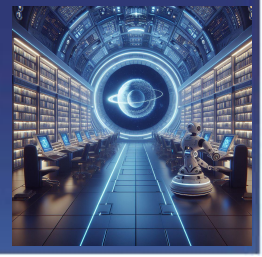
- ISVV is a systematic and **independent** process to ensure a software system:
  - Complies with **requirements**
  - Is **fit for purpose**
  - Operates **safely and reliably**
  - Is **traceable, testable, and auditable**

Common ISVV Phases (Customized per AI Type)



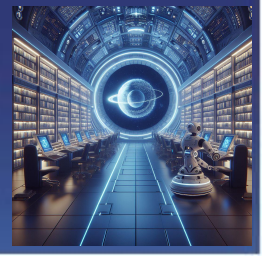
ISVV Phase	Classic AI (e.g., classifiers)	Generative AI (e.g., LLMs)	AI Agents (e.g., AutoGPT)
1. Requirements Analysis	Verify data scope, accuracy, performance metrics	Validate prompt-output behaviors, safety, and factuality	Define goals, constraints, ethics, autonomy boundaries
2. Data V&V	Data bias, representativeness, labeling correctness	Same as classifiers + prompt examples & training data audit	Multi-modal memory input, data poisoning risks, evolving state data
3. Model Evaluation	Confusion matrix, AUC, F1, accuracy	BLEU, ROUGE, hallucination, toxicity, truthfulness	Task success rate, plan reliability, safe completion
4. Behavior Validation	Classify under edge cases, adversarial testing	Prompt injection, prompt drift, robustness to paraphrasing	Goal misalignment, plan deviation, unintended tool use
5. Integration Testing	Model in pipeline context (e.g., image → diagnosis)	End-to-end prompt → system → action (e.g., response → UI)	Multi-step reasoning, tool usage, rollback behavior
6. Safety & Security	Model robustness, fairness	Toxicity filters, output filters, bias mitigation	Sandbox execution, override mechanisms, plan kill switches
7. Audit & Traceability	Feature → model → prediction trace	Prompt → token path → output trace	Goal → memory → plan → action traceability tree

Key Differences & Adjustments by AI Type (Classic AI)



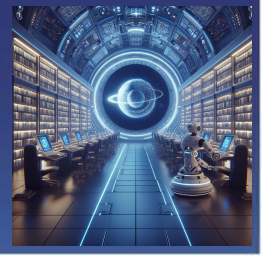
- **Classic AI (e.g., Classifiers, Rule-Based Systems)**
- **Deterministic** → Easier to test exhaustively
- Focus on:
  - Input space coverage
  - Labeling accuracy
  - Model fairness and performance
- **Tools**: sklearn, SHAP, adversarial robustness libraries
- **Status**: Defined, Standardized, e.g. ECSS-E-HB-40-02A – Machine learning handbook (15 November 2024)

Key Differences & Adjustments by AI Type (Generative AI)



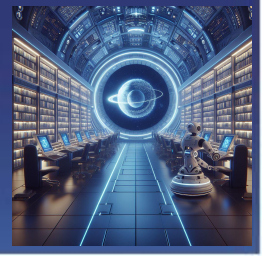
- **Generative AI (e.g., GPT, image generation)**
- **Stochastic outputs** → Must test statistically and at scale
- Focus on:
  - Hallucination rate
  - Prompt sensitivity
  - Safety filters and moderation
- **Challenges**:
  - No "ground truth" for many outputs
  - Emergent behaviors
- **Tools**: OpenAI Evals, PromptLayer, Detoxify, TruthfulQA, RAGAS
- **Status**: Spazio IT is currently verifying and validating AI applications on ISVV itself and Wound Care

Key Differences & Adjustments by AI Type (AI Agents)



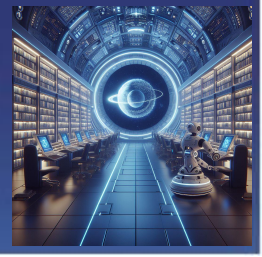
- **AI Agents (e.g., AutoGPT, hospital agents)**
- **Autonomous**, multi-step, stateful → Must validate across **sequences** and **goals**
- Focus on:
  - Goal alignment
  - Safety of external actions
  - Planning reliability
- **Emergent risks**: Loops, subversive behaviors, deceptive reasoning
- **Tools**: LangSmith, Trulens, EvalAgent, simulation environments, XAI modules
- **Status**: Spazio IT is currently experimenting with the Mistral AI agentic LLM Devstral, integrating it with static analysis tools

Key ISVV Artifacts per AI Type



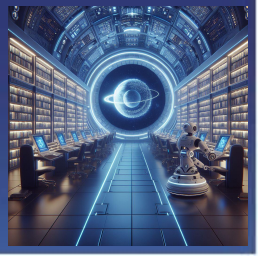
Artifact	Classic AI	Generative AI	AI Agents
Test Suite	Unit + edge case inputs	Prompt variation + scenario-based tests	Scenario simulations + goal-chaining tests
Audit Logs	Input/output/weights	Prompts, outputs, moderation decisions	Plans, actions, memory states
Compliance Reports	Accuracy, bias, fairness	Safety, hallucination, explainability	Ethical behavior, constraint violations, incident logs
Traceability Matrix	Features → Labels → Predictions	Prompts → Model → Outputs	Goals → Plans → Actions →

Summary Table: ISVV by AI Type

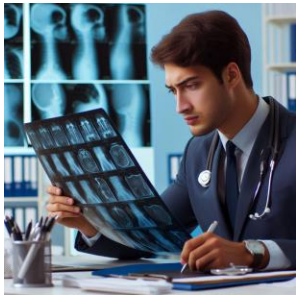


Aspect	Classic AI	Generative AI	AI Agents
Behavior	Predictive	Generative (stochastic)	Goal-directed & autonomous
V&V Focus	Accuracy, bias, robustness	Prompt safety, hallucination, coherence	Planning integrity, safety, goal alignment
Test Methods	Dataset slicing, adversarial examples	Prompt engineering, statistical testing	Simulation, red teaming, scenario fuzzing
Traceability Unit	Feature-label link	Prompt-output chain	Goal-plan-action tree
Risk Areas	Model overfitting, bias	Hallucination, toxicity	Misalignment, unsafe autonomy, tool misuse

AI Role by AI Type



Classic AI



Expert

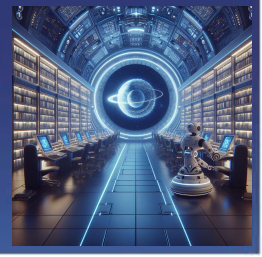
Generative AI

Assistant

AI Agents

Agent

Conclusions



- Each AI type requires a **tailored ISVV process**, but the **core ISVV pillars remain: independence, traceability, safety, and reproducibility**.

- If you're developing a **software system for ISVV**, you could build it with modular plug-ins to support:
  - Classifier evaluation and audit
  - Generative model prompt testing and moderation
  - Agentic behavior simulation and trace analysis