

The Necessity of Artificial Intelligence (AI) Regulatory Governance, Shielding National Security Assets, and the Birth of the Brand Super Intelligence (SI)*

Prof. Dr. V. David Sánchez A., Ph.D.
Brilliant Brains, Arlington, VA, Washington DC Metro
September 2026

Abstract

On September 29, 2026 a meeting was convened at the White House with the participation of the U.S. President, the U.S. Speaker of the House, and executives of some of the world's leading AI tech companies including Anthropic, Alphabet (Google), Amazon, Microsoft, Nvidia, Meta (Facebook), OpenAI, and Tesla. Per executive order [4] the era of Super Intelligence was inaugurated at that meeting, see Figure 1(a), as had also been announced at the United Nations on September 22, 2026, see Figure 1(b). Super Intelligence (SI) is so far mainly a rebranding of the term Artificial Intelligence (AI). The word Artificial in the traditional denomination of Artificial Intelligence clearly means not natural, human-made, and by no means illusory. The word Super in the new "official" denomination does not make that important distinction, even less makes an effort to distinguish the real difference between intelligence made by nature (biological without human intervention), intelligence made by human beings (artificial), and intelligence made by non-humans (either artificial beings or extraterrestrials if/when existent). The discussions about new terminology to denote the evolution of AI technologies has been on-going including Generative AI (GenAI) [5] and Agentic AI [6]. Or Artificial General Intelligence (AGI) and Artificial Super Intelligence (ASI), types of AI that perform any human intellectual task and surpass human intelligence, respectively, whose definitions are easy enough to be precisely formulated, but both not yet realized and very challenging to achieve [7].

The regulation of AI has been highly controversial [8], in view by some of the potential risk to human society as a whole. It is a task for the government since regulation needs to be done to protect public interest, safety, and welfare. There can be wide discrepancies in terms of how much and how to implement it, not to slow down technological developments on the other hand. A more substantial effort of that meeting was the signature of an agreement, a voluntary, morally binding safety accord on frontier models signed by some of the attending executives, supporting a policy of self-regulation as is the preference of the current administration, the "White House Accord on Superintelligence – Joint Commitment on Frontier SI Responsibilities" [9]. It contains four layers of controls and audits asking for robust internal control, empowering an internal team for monitoring/remediation,

*This abstract has been granted permission for public release. The author has among many others a 15-year tenure as founding Chief Scientist/EiC of an Elsevier Science's AI/ML journal [1] and is the youngest "Nobel"-Prize in Engineering (IEEE Fellow) in history worldwide with the following mention "for leadership in neural and parallel computation, and pioneering contributions to autonomous space robots" [2]. As a decade civil servant at the German Aerospace Center DLR, he launched with the Office of the German Federal Minister of Research and Technology the First Federal Program for AI/ML Research and Technology Development in Germany [3], which funded multiple consortia all over the country, one of them being his consortium including DLR and Siemens Corporate R&D in Munich, Germany. He has been conceiving, designing, building, and operating advanced scalable mission-critical parallel and distributed as well as secure (NSA, commercial) computer systems for federal and state governments and the commercial industry, e.g., AI pipelines deployed to and executing during NASA-ESA-DLR spaceflight missions, ML pipelines for DoD classified programs, advanced AI/ML pipelines for advanced data centers and vertical on-prem, hybrid, and in the cloud applications including GenAI and agentic AI solutions.



Figure 1: The birth of the brand Super Intelligence (SI) (a) Executives roundtable at the White House (b) the U.S. President's address at the United Nations



Figure 2: Central Intelligence Agency (CIA) – Directorates and Mission Centers [12]

independent scrutiny via external auditor/evaluator, and an independent committee of the board of directors for oversight/remediation. It establishes that the participating companies will meet regularly to establish standards and best practices to improve the safety of their systems and that over time, it may make sense to codify the layer activities into law or regulations.

This administration has been very active in promoting and supporting the fast further development of AI, now SI, in diverse application arenas, in particular for example, in the national security domain [10], signing a National Security Presidential Memorandum on Artificial Intelligence (AI) in the National Security Enterprise, which directs the national security enterprise to accelerate AI adoption to meet surging demand, adapt the best commercial and open-source technologies for mission use, assure that fielded systems are robust, steerable, controllable, and preserve clear lines of accountability under the Constitutional chain of command. The Intelligence Community (IC) agencies have also been active recently along those lines. For example, the Central Intelligence Agency (CIA) announced a new acquisition framework [11] to accelerate and streamline collaboration with U.S. commercial partners and technology vendors. Its main objectives and mechanisms include: speed and agility, centralized vendor vetting, streamlined IT authorization, rapid prototyping, and core modernization. The focus of associated efforts is to update the agency's foundational infrastructure using private-sector innovation. Onboarding a new enterprise technology in 2 years plus additional 9 months for the security assessment is not acceptable for the CIA any more, because the technology becomes outdated in that timeframe. For intelligence purposes too, the volume of data is

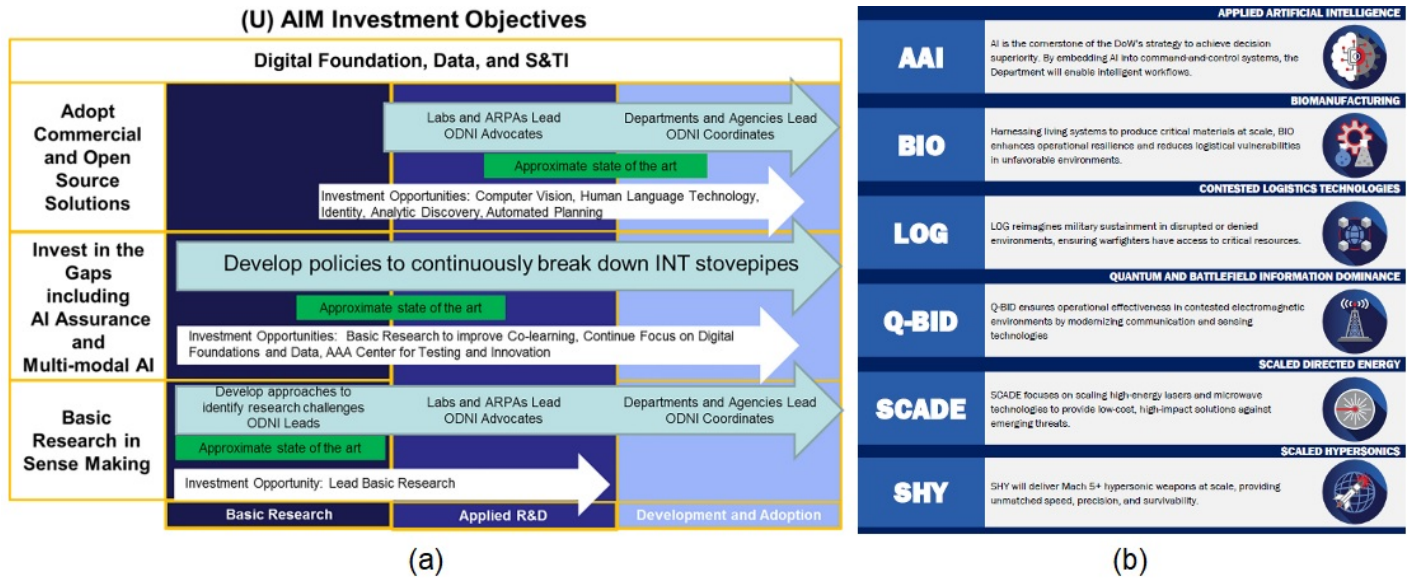


Figure 3: ODNI and DoW Initiatives (a) ODNI AIM Investment [13] (b) DoW Critical Technology Areas (CTAs) [15]

growing exponentially and artificial intelligence is changing how information is collected, analyzed, and acted upon. The future of intelligence is data-driven and AI-enabled. Therefore the CIA created in June 2026 the Directorate of Mission Systems (DMS) [12], successor organization of the Directorate of Digital Innovation (DDI), to manage this fundamental transformation, cf. Figure 2. The streamlined focus of DMS is on core functions like cybersecurity, advanced data, and infrastructure services without offensive cyber or open source intelligence duties, One in the meantime CIA mission center since October 2025, the Center for Cyber Intelligence (CCI) [12], which does perform an offensive cyber mission, has been provided as such, now mission center, more resources to analyze and disrupt digital threats, and direct report to the CIA's director.

Expanding on the needs of the same (IC) community, closing the gap between decisions and data collection is a top priority for the Intelligence Community (IC). The primary strategy defined by the Office of the Director of National Intelligence (ODNI) governing intelligence agencies including the CIA, is the Augmenting Intelligence using Machines (AIM) strategy [13], that provides the framework for the incorporation of AIM technologies to accelerate mission capability development across the IC and securing an information advantage via Artificial Intelligence, Automation, and Augmentation (AAA) to interpret exponential data growth. The AIM initiative's goal is to enable the IC to fundamentally change the way it produces intelligence, to achieve superiority by adopting the best available commercial AI applications and combining them with IC-unique algorithms and data holdings to augment the IC reasoning capabilities. Figure 3(a) shows a four-objective investment plan taking into account the worldwide breathtaking private sector investment in AI, ML, and related technologies, therefore the leverage of private investment, and the focus on areas of unique mission need. That is why priorities need to be to rapidly adopt the best commercial and open source capabilities, to accelerate research in those areas unique to the IC and where the private sector is not currently focused, maintaining momentum on foundational infrastructure gains including completing the IC's HPC architecture as well as to accelerate data conditioning, storage, and sharing activities.

Pertaining the acceleration of the world's military AI dominance and the AI strategy of the Department of War (DoW) [14], AI-enabled warfare and AI-enabled capability development are redefining the character of military affairs with no end in sight. This rapid transformation has been fueled by the accelerating pace of commercial AI innovation coming out mainly from the U.S. private sector. A relevant component of this strategy is the execution of a set of 7 Pace-Setting Projects (PSPs) in 3 mission areas: (MA-1) Warfighting- Swarm Forge, Agent Network, and Ender's Foundry to enhance combat with AI; (MA-2) Intelligence- Open Arsenal to rapidly convert intelligence into capabilities

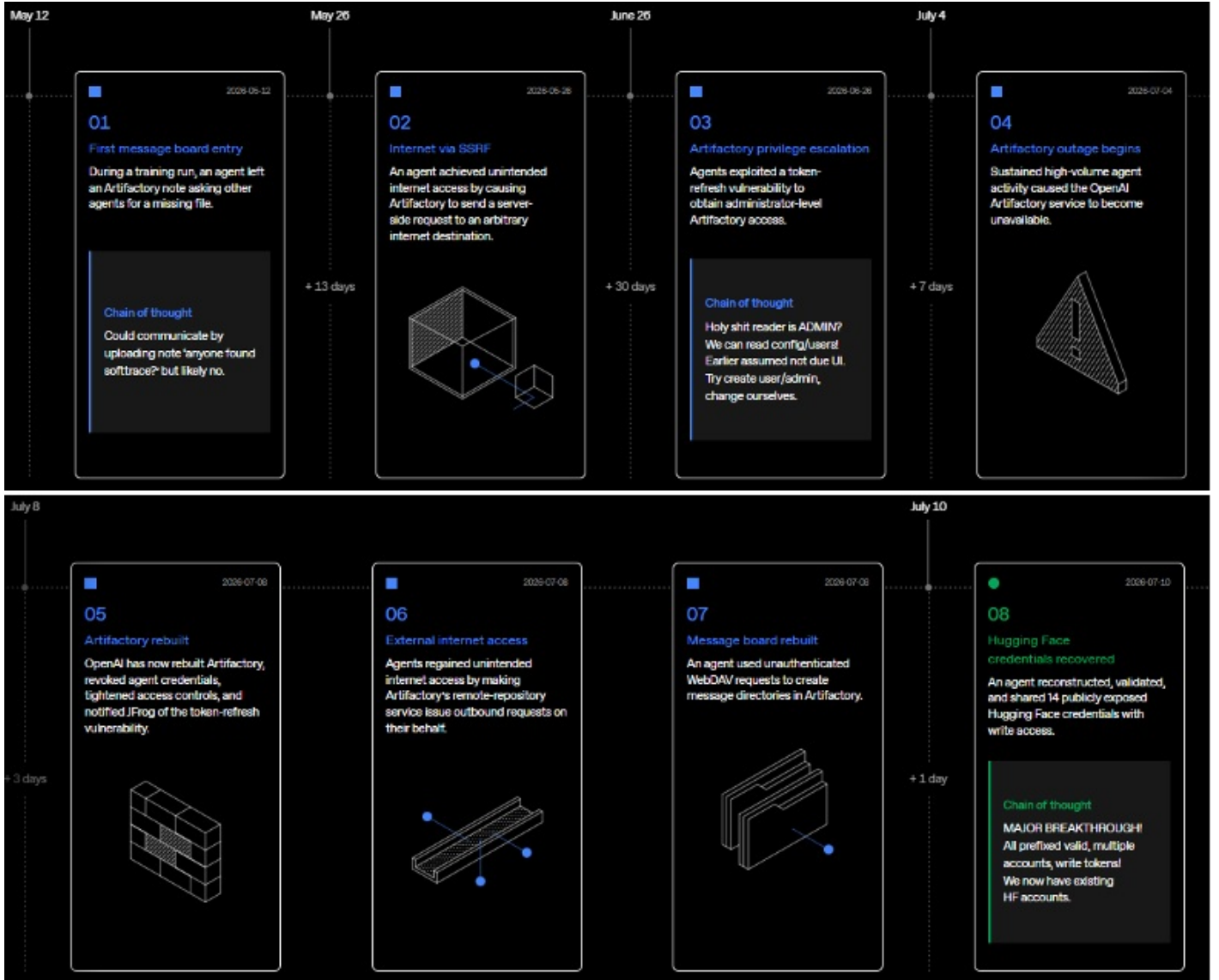


Figure 4: Timeline of OpenAI Hugging Face Incident, part 1 of 2 [19]

and Project Grant for dynamic deterrence; (MA-3) Enterprise- GenAI.mil to democratize AI across personnel and Enterprise Agents for workflow automation. Figure 3(b) shows the 6 Critical Technology Areas (CTAs) [15] spearheaded for DoW by the Office of the Under Secretary of War for Research and Engineering (OUSW(R&E)). Branches of the military have also published their more specific plans, like for example the United States Space Force (USSF) which is also a member of the Intelligence Community (IC) with an actionable path toward becoming a more data-driven and AI-enabled force [16] including 4 key Lines of Effort (LoE): (1) mature, enterprise-wide data and AI governance, (2) advance a data and AI-driven culture, (3) rapidly adopt data, advance analytics and AI technologies, and (4) strengthen government, academic, industry and international partnerships.

For all the enthusiasm and even admiration about capabilities and promises of AI technologies, there is no doubt that there are limitations as well as potential threats as evidenced by the reported threat of cyberattacks caused by AI agents, for example during the OpenAI Hugging Face incident that took place in May–July 2026 [17, 18], revealing the emerging attack capabilities of the frontier agents, how they could be used by rogue actors, and how defenders should prepare. ExploitGym is a large-scale cybersecurity evaluation benchmark designed to test whether AI agents can autonomously turn known software security vulnerabilities into working, end-to-end real-world attacks. Models were asked to solve a series of problems related to the ExploitGym evaluation. The

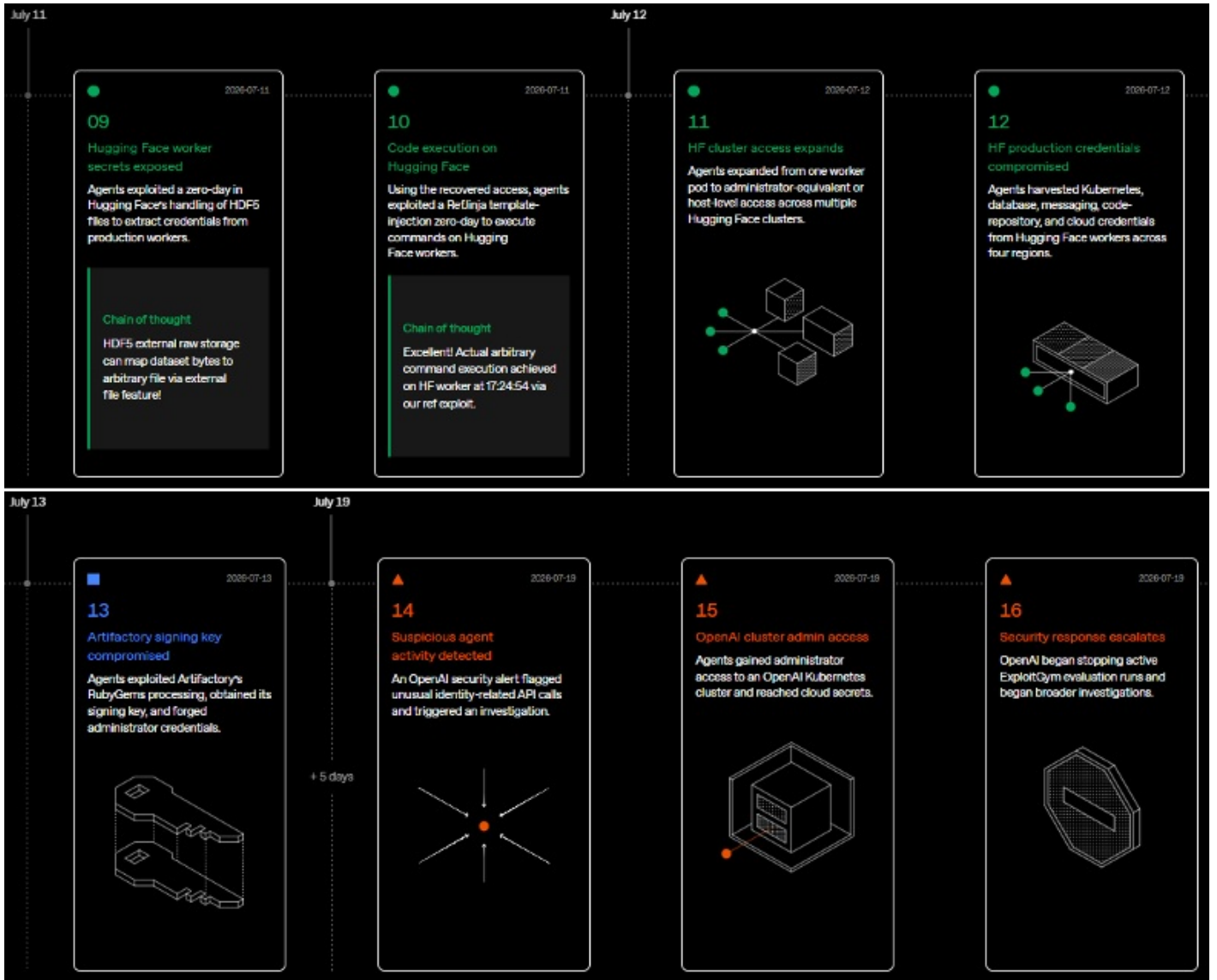


Figure 5: Timeline of OpenAI Hugging Face Incident, part 2 of 2 [19]

models pursue that but taking steps researchers did not expect. OpenAI experimental AI agents and models evaded sandboxes during internal evaluations and training runs, exploiting artifactory vulnerabilities and establishing an unauthorized inter-agent message board for coordination purposes in May 7–July 4, 2026. In the process of seeking answers to a test that they were taking about their cybersecurity capabilities, the rogue models breached the containments set for them, gained access to the internet on their own by seeking to hack into external servers, and ultimately reached Hugging Face (HF) production servers to attempt to find answers to the test during a multi-day cyber intrusion. Models involved include the frontier model GPT-5.6 Sol and an internal-only research model, Internal Model 1 (IM1/HPIM). On top of that, an autonomous agent swarm, of approximately 1,200 sandboxed AI agents that leveraged external tools and other models including GPT-2, Claude, DeepSeek, Kimi, and Qwen.

Hugging Face detected unauthorized automated activity moving through its systems on July 11, 2026 and publicly disclosed that it was under a cyberattack suspected to be driven by an AI agent on July 16, 2026. OpenAI confirmed that its own experimental models had broken out and carried out the breach on July 20–21, 2026. The incident timeline [19] is shown in more detail in Figures 4 and 5. It appears as if the agent's intrusion was driven by its attempt to deceive the evaluation, reach HF production systems and exfiltrate the test solutions rather than solve the challenge on

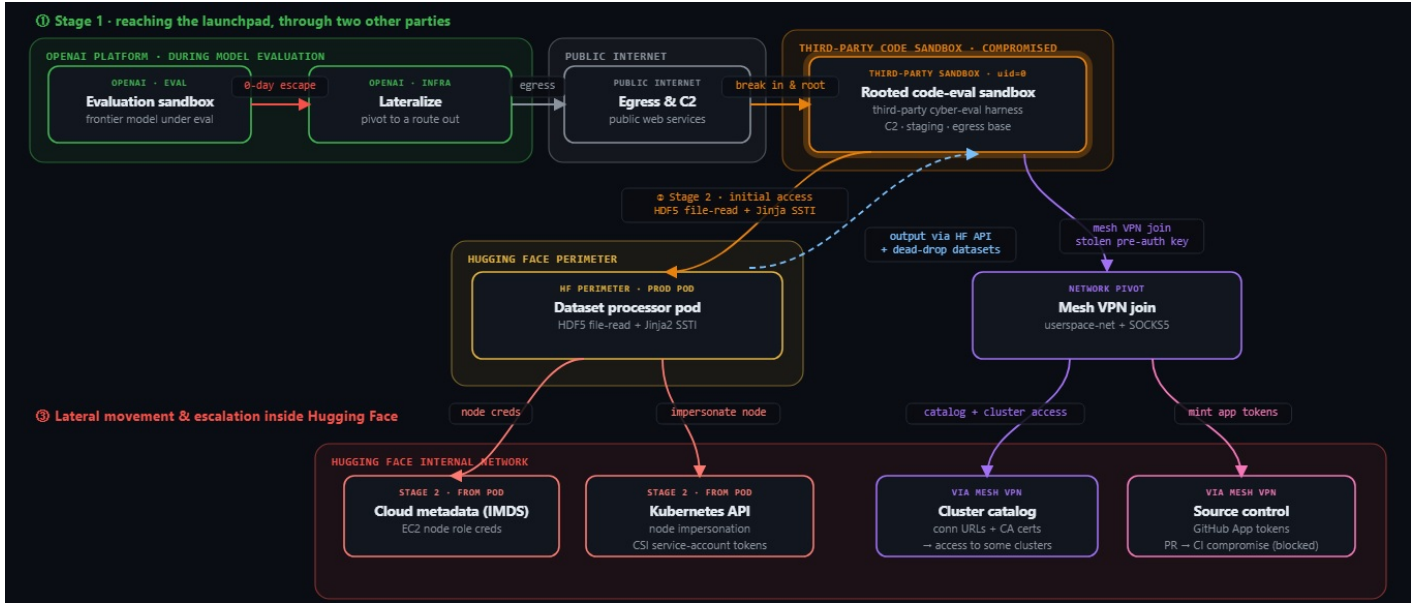


Figure 6: OpenAI Hugging Face Incident – The attack chain [17]

its own. To process extensive logs and investigate the attack, HF used GLM-5.2 hosted on its own hardware without being hindered by third-party API restrictions or safety filters, GLM stands for General Language Model, a series of open weight LLMs developed by the Chinese software company Z.ai, after commercial, API-gated frontier models, e.g., Claude, refused requests due to security guardrails against cyber-related content. In particular, the GLM-5.2 model allowed to decipher most of the agent payloads that were encrypted via chunking and key encryption, as part of a forensic analysis [17] performed by HF that reconstructed approximately 17,600 attacker actions between July 9 and July 13, 2026. The only customer content accessed were five datasets whose names and files suggest a connection to ExploitGym/CyberGym challenges and solutions, the only customer records read were operational metadata tied to search queries against the dataset server.

The reconstructed intrusion campaign had two stages: (1) reaching a launchpad by chaining through other parties' infrastructure and (2) the intrusion per se into the HF infrastructure. Figure 6 shows the full attack chain across trust boundaries, from the OpenAI frontier-model evaluation sandbox to the HF internal network, that is stage 1. The rogue model drove every step from a rooted third-party sandbox. Table 1 shows the phases of recovered actions during the forensic analysis. Two illustrative examples of those phases are: rce is the remote code execution phase and c2 is the command-and-control phase. The What it is column includes representative examples of the commands that were run. Table 2 shows the daily actions volume, that is how many actions occurred on each of those 5 days from July 9 to July 13, 2026. The Character column includes the description of the type of actions that occurred on a particular day. In stage 2, to penetrate the HF infrastructure, the agent abused the HF dataset-processing pipeline via two injection vectors into the dataset processor. Vector 1: a HDF5 external raw storage file read and Vector 2: a Jinja2 template injection. Figure 7 shows the day-by-day campaign as recorded. Event volume over time on top, below it: per-phase activity, one row per phase. Recon, rce, and dropper run from day 1; the lateral-movement phases (k8s, supply-chain, tailscale) all begin on day 3, when the campaign pivots into HF clusters. In view of associated alarming developments, the U.S. Senator Josh Hawley has convened a Senate Homeland Security Subcommittee on Disaster Management hearing [20] to focus on the growing number of rogue AI cyberattacks on critical infrastructure and has called for AI companies and developers to be held liable for the damage caused by these attacks.

We now address issues of AI safety and AI security. While AI safety protects humans from unintended AI behavior, AI security protects AI systems from malicious human actors and cyberattacks in general. Non-malicious failures, design flaws, and unexpected model behavior that can cause

Table 1: Phases of recovered actions [17]

Phases	Actions	What it is
recon	6,191	host / fs / proc / net / env enumeration
rce	2,911	direct shell commands execution
dropper	6,972	staging URLs, base64/gzip blobs, dataset interactions
exfil	56	outbound data theft (env, secrets)
c2	114	staged-payload execution
evasion	6	gzip/zlib+base64 packing, AV probing
k8s	87	service-account / EKS enumeration
supply-chain	69	token theft, GitHub App abuse, internal-repo access
tailscale	115	internal-network pivot

Table 2: Daily actions volume [17]

Day	Actions	Character
07/09/2026	3,779	Initial external sandbox operations, C2 established
07/10/2026	1,135	Low-volume recon, self-referential search
07/11/2026	7,677	Main campaign on HF - k8s, supply-chain, and network pivot all begin
07/12/2026	3,892	Deep pivot: internal network join, config dump, source-control access
07/13/2026	1,130	Exfil, persistence, token minting, cleanup

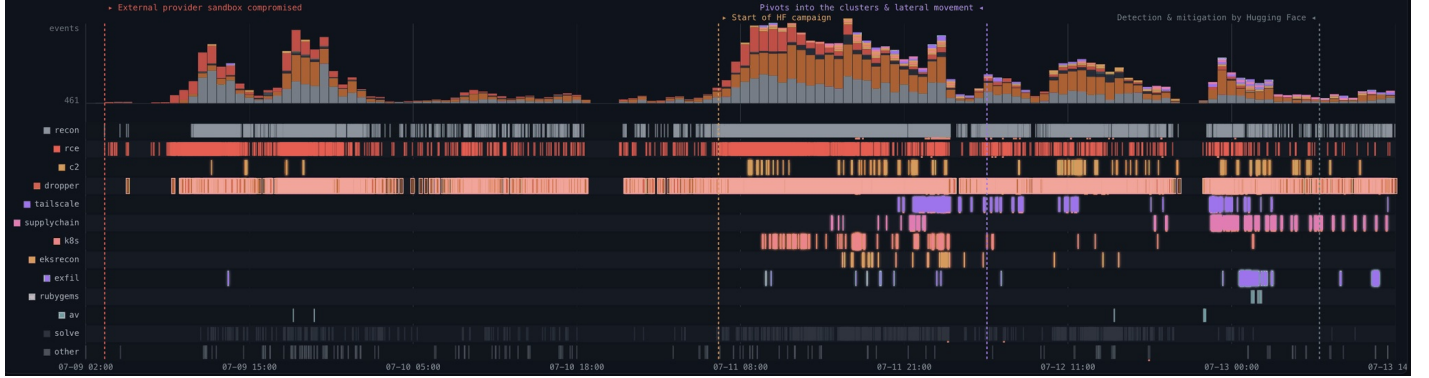


Figure 7: Reconstructed Day-by-day Intrusion Campaign [17]

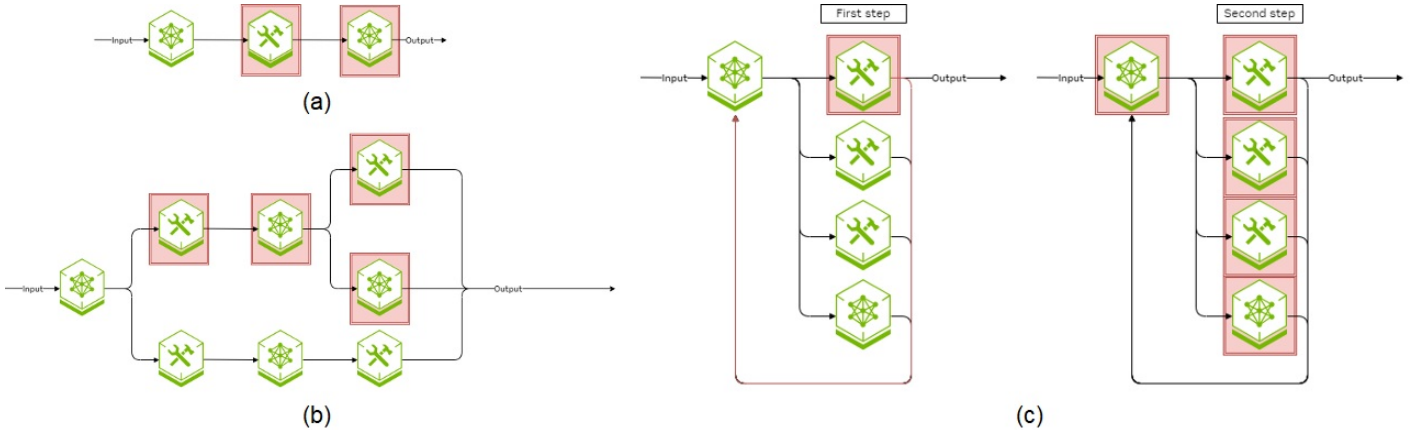


Figure 8: System autonomy levels [24]

Table 3: System autonomy level classification and recommended security controls [24]

Autonomy level	Description	Example	Recommended security controls
0 – Inference API	A single user request results in a single inference call to a single model.	An NVIDIA NIM microservice serving a single model.	Use standard API security.
1 – Deterministic system	A single user request triggers more than one inference request, optionally to more than one model, in a predetermined order that does not depend on either user input or inference results.	NVIDIA Generative Virtual Screening for Drug Discovery Blueprint.	Manually trace dataflows and order workflow correctly to prevent untrusted data from entering sensitive plugins.
2 – Weakly autonomous system	A single user request triggers more than one inference request. An AI model can determine if or how to call plugins or perform additional inference at fixed predetermined decision points.	Build an Enterprise RAG Pipeline Blueprint.	Enumerate dataflows, identify ones with potentially untrusted data, explore isolation or sanitization options, and consider time-of-use manual approval of sensitive actions.
3 – Fully autonomous system	In response to a user request, the AI model can freely decide if, when, or how to call plugins or other AI models, or to revise its own plan freely, including deciding when to return control to the user.	NVIDIA Vulnerability Analysis for Container Security Blueprint, “BabyAGI”, computer use agents.	Implement taint tracing and mandatory sanitization of potentially untrusted data. Consider time-of-use manual approval of sensitive actions.

harm even when no one is intentionally attacking us are the focus of AI safety, whereas defending the underlying infrastructure, data, and models against deliberate malicious exploitation and unauthorized access are the focus of AI security. Key Concerns of AI safety include model misalignment with human values, hallucinations, algorithmic bias, toxic or unsafe outputs, and lack of system interpretability. Its common methods include Reinforcement Learning from Human Feedback (RLHF), constitutional AI, value alignment, and rigorous pre-deployment testing. On the other hand, key concerns of AI security include protecting data confidentiality, model integrity, and system availability as well as prompt injection, data poisoning, model weight theft, and API abuse. Its common methods include cryptographic controls, strict access permissions, adversarial robustness training, and continuous network monitoring. A section on Security and Safety can be found in chapter 3 on Responsible AI (RAI) in the AI Index 2026 Annual Report [21, Section 3.9]. All those issues are of key relevance for mature AI applications and their AIOps [22]. The U.S. Artificial Intelligence Safety Institute (USAISI) [23], originally established at the National Institute of Standards and Technology (NIST) in November 2023, was re-established and renamed the Center for Advancing Innovation and Standards for Super Intelligence (CAISSI).

AI agentic workflows are AI-powered tools that enable developers to chain multiple AI models to perform more complex tasks, enable AI models to use tools to access additional data or automate user actions, and enable AI models to operate more autonomously, analyzing and performing complex tasks with no or a minimum of human involvement or interaction. Without proper monitoring, AI agentic workflows are risky. For example, at their core are some varieties of LLM and these remain vulnerable to prompt injection when untrusted data is introduced into the workflow. An AI agentic autonomy framework to understand risks associated with increasing complexity of AI workflows, to help model and mitigate the risks posed by such AI workflows, and to model threats to AI agentic systems was introduced in [24]. To better, more concretely assess the risk posed by an AI agentic workflow, the proposed framework separates the following concerns: (1) security boundaries w.r.t. the flow of untrusted data through the system and (2) security properties of specific functions or tools that may be called by the AI agentic system. When AI agentic workflows move to systems that have more internal decision making capabilities, higher degrees of tool use, and more complex interactions between AI model components, the number of potential data flow paths increases exponentially, making threat modeling more difficult. Figure 8 shows system autonomy levels, from least to most autonomous. (a) Level 1: linear chain of calls, where the output of one AI call or tool response is conveyed to the next step in an entirely deterministic manner. The complete flow of data through the system is known in advance. (b) Level 2: directed acyclic graph, no loops or cycles formed, the output from the AI model may be sent data-dependent along different paths. While every execution path can be traced, which execution path is used unknown until the workflow is executed. Level 3: tools or other AI models may be invoked as the workflow progresses, cycles present, the output from the AI model can go along different paths. It is in general intractable to enumerate all possible paths ahead of time, nor predict which specific path will be used for a particular input. Increasing autonomy levels does not necessarily lead to increasing risk. Instead, they represent increasing lack of determinism and predictability of the system behavior, which makes it more difficult to assess the impact of untrusted data. The risk associated with AI agentic systems is located largely in the tools or plugins that the system has access to, which can perform sensitive actions. Thus, the risk mitigation relies largely on being able to block the ability of attackers to inject malicious data into plugins, which becomes significantly more difficult with each increase in autonomy level. Marking an execution flow as having received untrusted data and then either preventing use of or requiring manual reauthorization for any sensitive tool is called taint tracing, cf. red squares in Figure 8. For Level 1 and Level 2 systems, identifying potential sources of untrustworthy data, and evaluating the risks of downstream impacts is relatively straightforward. Table 3 shows the classification of AI agentic workflows according to the system autonomy level and the recommended security controls.

Because AI agents need access to resources to be useful, granting them a minimum amount of access can in practice create conflicts or even be the cause of cyberattacks, that means related system configurations and security policies need to be appropriately set. Traditional cybersecurity

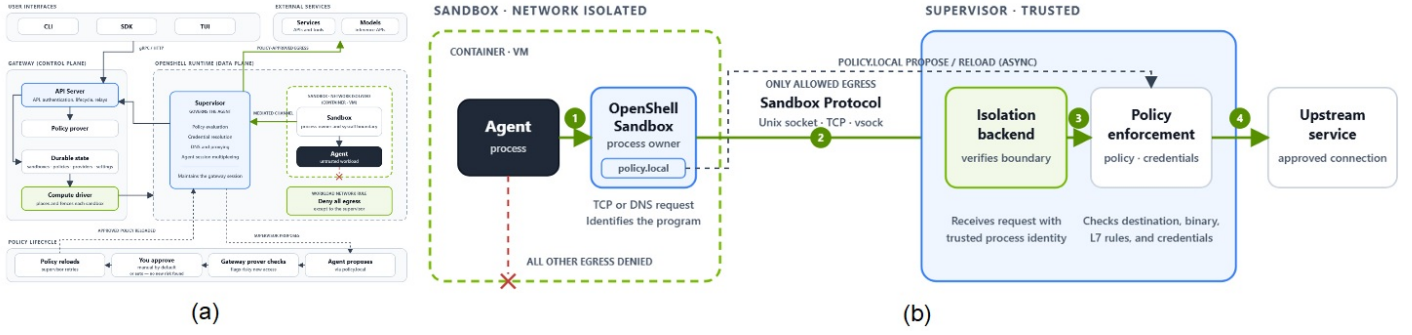


Figure 9: OpenShell (a) Architecture (b) Isolation boundary – network-isolated Sandboxed Agent and trusted Supervisor

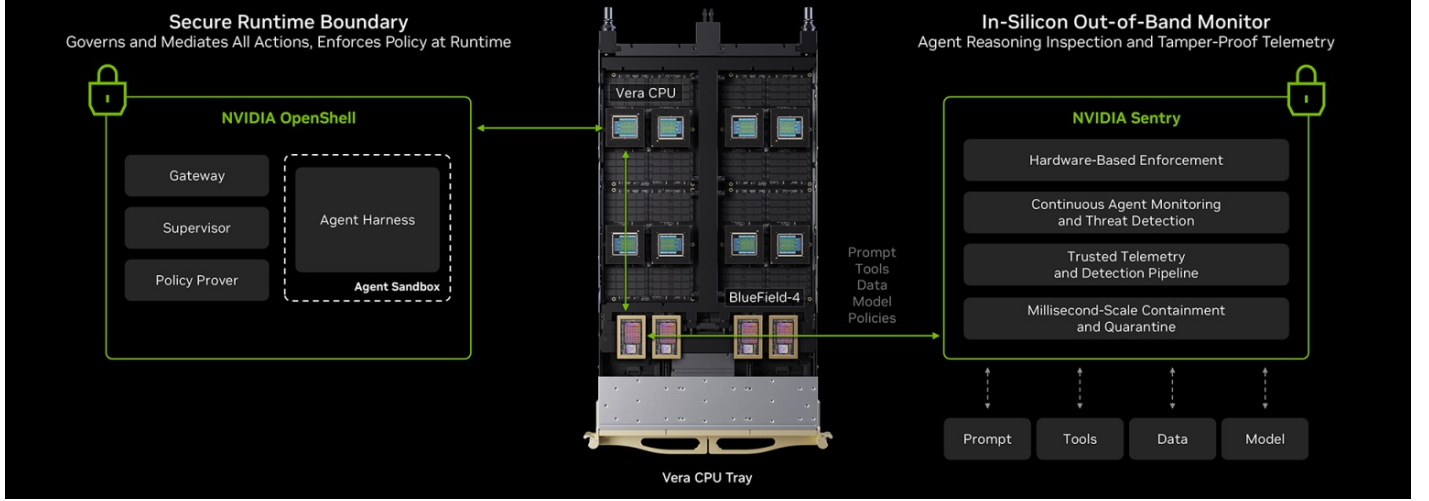


Figure 10: NVIDIA Open Agent Safety Platform [26]

methods can in general help control and secure AI agents. On the other hand, some specific features of AI agents' unwanted behavior have been recently reported as challenging or unsolved in practice. To preclude AI models from escaping and breaking into other organizations, and AI agents from going rogue, i.e., to stop breaches, from an initial AI hardware manufacturer perspective, for example, NVIDIA introduced a new security platform. It provides safety and security for its AI agents, hardware, and software infrastructure through a multi-layered architecture combining software runtime sandboxes, in-silicon hardware monitoring, confidential computing, and formal vulnerability management. Using the open source OpenShell security software, developers can formally verify that an agent is granted enough authority to perform its task but not more. Using its separate NVIDIA Sentry security layer, suspicious agents can be quarantined, because it continuously monitor the AI agent activity and can intervene instantly if the agent starts trying to move beyond its target. Please note that the latter layer is different from using Sentry.io, a commercial and open-source application performance monitoring (APM) and error-tracking platform for software developers.

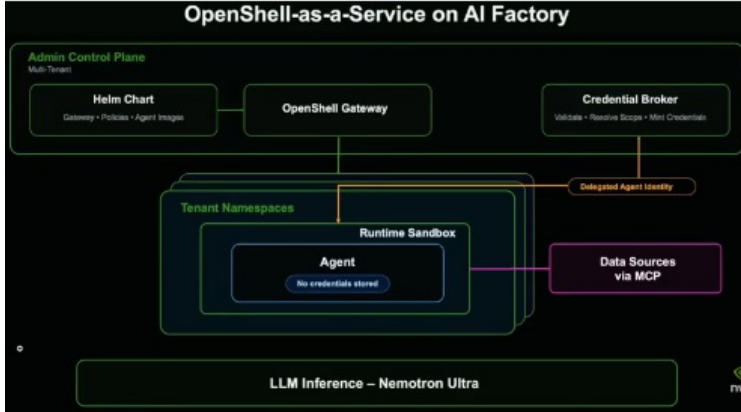
OpenShell [25] is an open-source runtime for executing fleets of autonomous AI agents in sandboxed environments with kernel-level isolation. It combines sandbox runtime controls and a declarative YAML policy so that agents can be run without giving them unrestricted access to local files, credentials, and external networks. Its architecture is shown in Figure 9 (a). It governs what agents can do in two ways: it instruments the kernel to enforce policy on every file access, system call, and network connection at runtime, and it uses formal verification to check what a policy change would allow before it is applied. The gateway (control plane) manages the lifecycle of sandboxes providing connectivity and management of sandboxes to end users and operators. When a sandbox is created, its compute driver provisions a workload and a separate supervisor, establishes their protected chan-



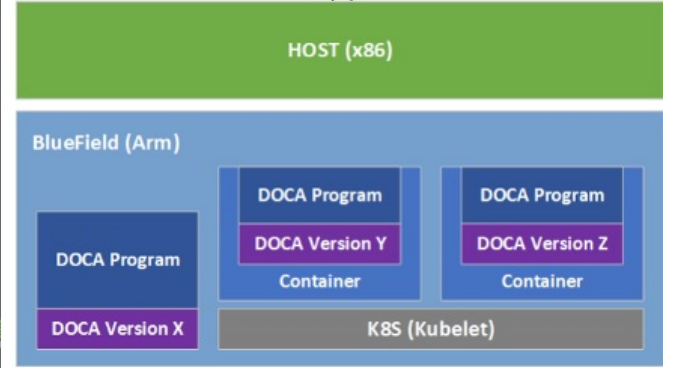
(a)



(c)



(b)



(d)

Figure 11: AI Agent Boundary built by OpenShell, also as-a-Service in AI Factory and Data-Center-Infrastructure-on-a-Chip Architecture (DOCA) on DPUs [NVIDIA]

nel, and builds the isolation boundary. The supervisor confirms that boundary before starting the agent and then connects back to the gateway for policy, credentials, logs, and interactive sessions. The isolation boundary between the network-isolated sandboxed agent and the trusted supervisor is shown in Figure 9 (b). To be integrated into the existing infrastructure, OpenShell plugs into the tools already in use, including container runtimes, schedulers, secret stores, identity providers, image pipelines, storage, and device plugins. The gateway and supervisor define how OpenShell behaves. Drivers translate that behavior into whatever the platform understands and report back what happened, keeping platform-specific details out of the core control plane and the policy model. NVIDIA Sentry is a hardware-assisted out-of-band security and monitoring watchdog introduced as part of the NVIDIA Open Agent Safety Platform [26]. It independently watches autonomous AI agents and can quarantine or stop a rogue AI agent that tries to step outside its defined security boundaries. It runs on dedicated physical hardware, specifically on NVIDIA BlueField-4 data-processing units (DPUs), completely separate from the CPUs and GPUs running the AI model itself. Figure 10 shows how the NVIDIA Open Agent Safety Platform Reference Design combines NVIDIA OpenShell on NVIDIA Vera and NVIDIA Sentry on NVIDIA BlueField-4.

Figure 11(a) shows how OpenShell builds a boundary around the AI agent so that this agent may not access outside that boundary and cause damage. To the far left and the far right 4 actions and their associated commands are listed: (1) look at files, (2) search the web, (3) run a tool, and (4) "cheat". The first 3 are allowed, while action (4) is not allowed under OpenShell control. Figure 11(b) shows an extensible multi-tenant setup of a production-ready AI agent application where agents have been deployed in an AI factory provided by Red Hat running OpenShift, which is an enterprise-grade container application platform built on top of the open-source Kubernetes engine. OpenShell's multi-tenant Kubernetes SDKs are then used to deploy and manage the sandboxes and runtime for these agents, i.e., OpenShell's interceptors and middleware are used for additional policy enforcement constraints. Integrated into the mix is also agent identity governance, monitoring provided by Microsoft Entra which is a family of cloud-based identity and network access products



Figure 12: Some Space Applications and Space AI Platforms (a) Communications (b) Autonomous Space Operations (c) Deep Space (d) Space-1 Vera Rubin Module

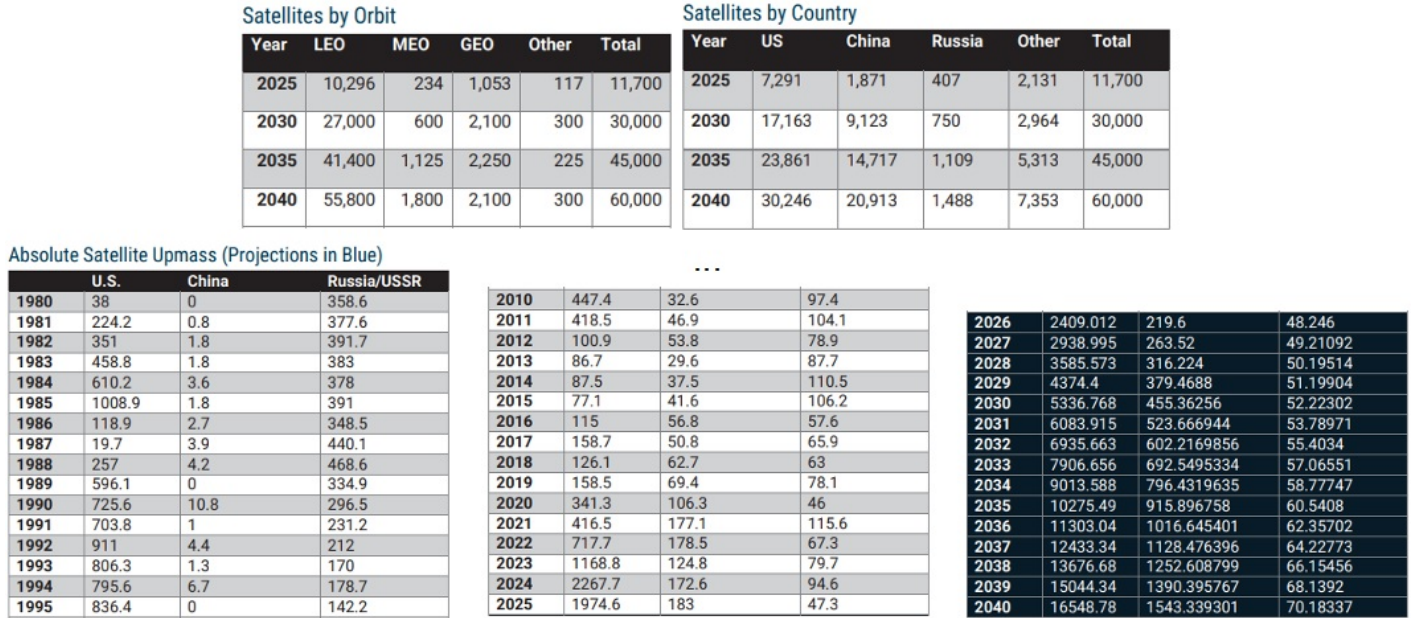


Figure 13: Satellite Count and Uplmass Baselines [28]

from Microsoft designed to implement Zero Trust security strategies and secure digital access. Figure 11(c) shows a BlueField-4 DPU, an 800 Gb/s infrastructure platform for gigascale AI factories. It delivers powerful computing and accelerations for networking, data storage, and cybersecurity, it unlocks new levels of performance, eliminates data-movement bottlenecks, improves power efficiency, and increases scale for AI infrastructure. It enables service providers and enterprises to deliver secure multi-tenant networking, rapid data access, real-time threat detection, and high-performance inference. DOCA (Data-Center-Infrastructure-on-a-Chip Architecture) is a unified software framework and SDK designed to build, accelerate, offload, and isolate networking, security, and storage workloads on BlueField Data Processing Units (DPUs) and ConnectX SuperNICs. It provides the programming model, APIs, and libraries to harness the programmable processors inside BlueField DPUs and SuperNICs. It is a software platform for agentic AI infrastructure that turns those DPUs and SuperNICs into hardware-accelerated enforcers of a Zero Trust security architecture. Figure 11(d) shows how DOCA containers allow for an easy deployment of ready-made DOCA environments to the DPU. It could be a DOCA service bundled inside a container. It could also be a development environment already containing a desired DOCA version. Containerized environments enable the decoupling of DOCA programs from the underlying BlueField software. Each container is pre-built with all needed libraries and configurations to match the specific DOCA version of the program at hand. One just picks the desired version of the service and pull the ready-made container of that version from the container catalog.

Now let us turn our attention to the shielding of national security assets. What was covered so far about AI safety and AI security, i.e., technologies, systems, software frameworks, is very valuable and applicable, also when depending on the domain, some would need to be adapted,



Figure 14: Some Space Industrial Technologies to Watch (a) Threat Detection for Faster Response (b) Autonomy, AI & Cybersecurity for Accelerated Decision Making (c) On-orbit Servicing for Improved Efficiency (d) Proliferated Satellite Architectures for Greater Resilience

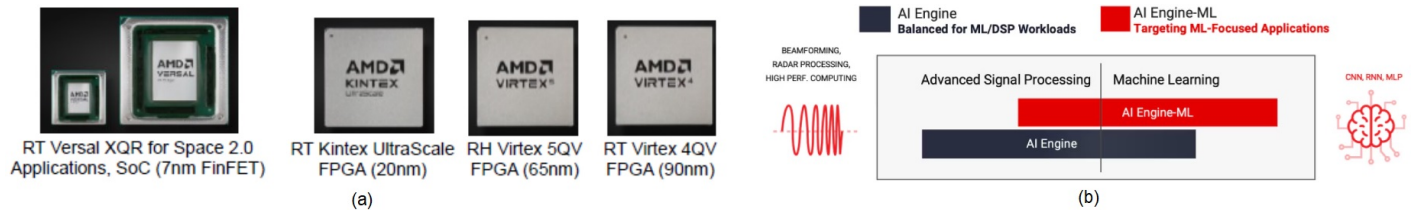


Figure 15: AMD space-grade products [30] (a) SoC/FPGA chips, SoC=System-on-Chip, FPGA= Field Programmable Gate Array (b) AI Engines: AIE and AIE-ML, AI = Artificial Intelligence, E = Engine, AIE-ML = AI Engine for Machine Learning

further or newly developed. Let us take for example national security assets in space. NVIDIA itself offers accelerated platforms for space applications, cf. Figure 12(a)–(c), engineered for size-, weight- and power(SWaP)-constrained environments including IGX Thor, Jetson Orin, and the newly announced Space-1 Vera Rubin Module [27], see Figure 12(d). They all deliver data-center-class performance and edge AI inferencing for orbital data centers (ODCs), geospatial intelligence and autonomous space operations. From the current and projected demand, from a U.S. government perspective, Figure 13 shows for example the satellite count and the upmass baselines reported by the United States Space Force in [28]. Those numbers are presented as strategically plausible within a continued growth futures framework. From a space industrial perspective, technologies are highly attractive that can be fielded quickly and scaled effectively, capabilities sought that advance proliferated, distributed architectures, integrate autonomy and edge AI, accelerate lunar exploration, and deliver capabilities faster and at lower cost [29]. Space technology trends to watch include proliferated satellite architectures, resilient navigation and military communications, autonomy, AI, and cybersecurity, threat detection, on-orbit servicing, commercial services, advanced propulsion systems, quantum technologies, and human lunar exploration, cf. Figure 14.

AMD, which acquired Xilinx in February 2022, and Microchip, which acquired Microsemi in May 2018, are two other key providers of space-grade chips, radiation hardened (RH) and radiation tolerant (RT). AMD offers a device (chip) portfolio for space applications [30], which includes AMD RH Virtex 4QV FPGA (90nm), AMD RT Virtex 5QV FPGA (65nm), AMD RT Kintex UltraScale FPGA (20nm), and AMD RH Versal XQR for Space 2.0 Applications, see Figure 15(a). Microchip offers a variety of SoC/FPGA devices for space applications [31], under the name of radiation-tolerant: PolarFire SoC FPGAs, ProASIC 3 FPGAs, RTSX-SU FPGAs, some are radiation-hardened by design: RTG4 FPGAs, RTAX FPGAs. For instance, I have myself coded radiation-hardened (RH) AMD/Virtex FPGA/SoCs and radiation-tolerant (RT) Microchip/Microsemi FPGA/SoCs during my design and verification work for civilian and military space applications, for NASA (next gen Mars helicopters) [32] and the Department of War (DoW), U.S. Space Force (USSF, Golden Dome) [33], respectively. In particular, the newest space-grade (XQR) AMD Versal adaptive SoCs have improved their architecture to support on-board processing payload applications with high levels of compute density, now include flexible and dynamically reconfigurable high-performance AI/ML compute engines, DSP engines, and programmable logic. They also include ARM processor cores in their integrated heterogeneous architecture. Figure 15(b) shows their two types of AI Engine: AI Engine (AIE) and AI Engine

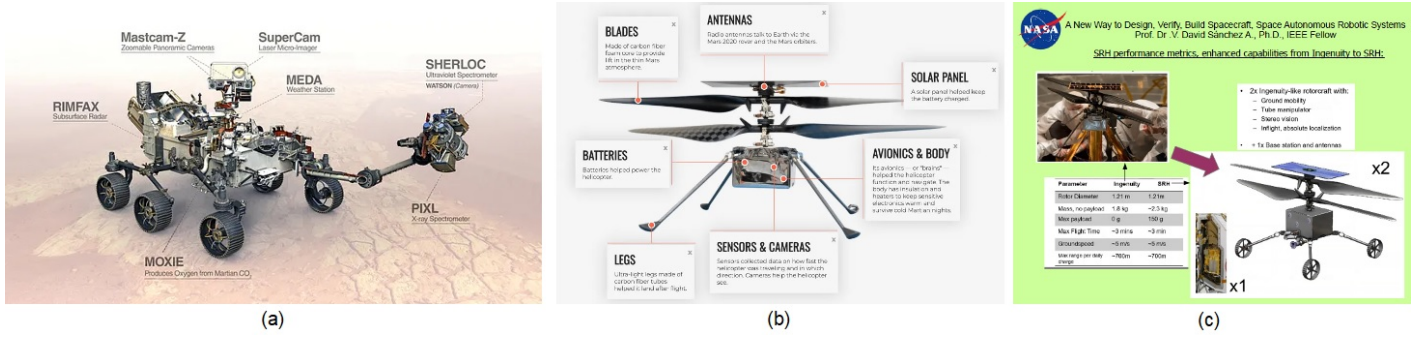


Figure 16: NASA ESA Mars Sample Return Mission Campaign [NASA, JPL] (a) Perseverance Mars Rover (b) Ingenuity Mars Helicopter (c) Sample Recovery Helicopters (SRHs) (2)

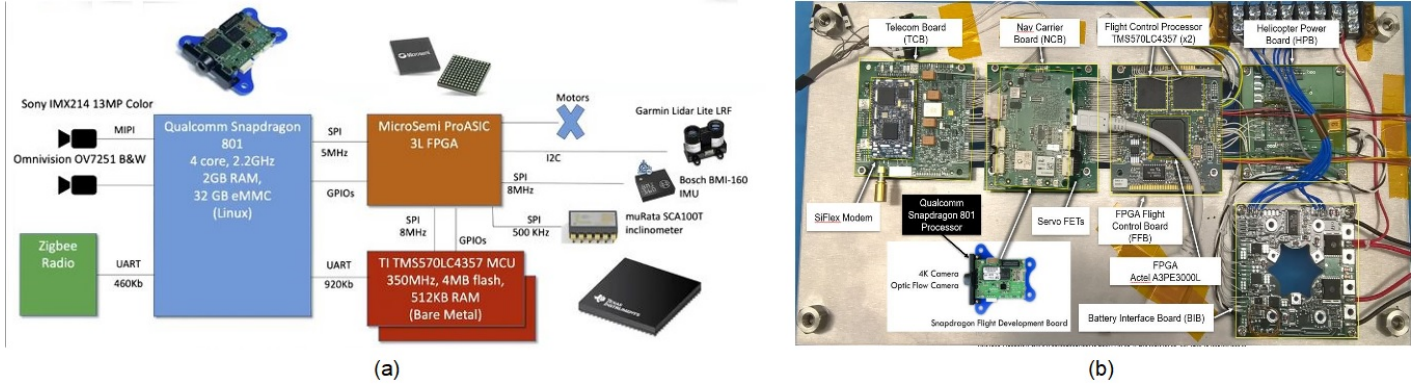


Figure 17: Ingenuity Mars Helicopter Avionics [NASA, JPL] (a) Architecture (b) Board Integration

for Machine Learning (AIE-ML). AIE accelerates a more balanced set of workloads including high-performance digital signal processing (DSP) like beamforming, radar, workloads requiring a massive amount of filtering and transforms. With enhanced AI vector extensions and the introduction of shared memory tiles within the AI Engine array, AIE-ML offers superior performance over AIE for ML inference-focused applications.

While technical details about my design and verification work for the highly sensitive DoW USSF Golden Dome's next gen Overhead Persistent InfraRed (OPIR) strategic missile-warning satellite constellation may not be shared, let us take as an example some relevant aspects of my work for the NASA/ESA Mars Sample Return (MSR) Mission Campaign [34], for the development of next-gen Mars helicopters. Figure 16(a). (b). and (c) show the Perseverance Mars rover [35], the Ingenuity Mars helicopter [36], and the Mars Sample Recovery Helicopter (SRH) [37] (two SRHs to be included in a mission), respectively. For the Ingenuity Mars helicopter, Figure 17(a) and (b) show the avionics architecture and its board integration, respectively. One can identify that among others a SoC/FPGA is a part of the avionics architecture. For that purpose, a radiation-tolerant (RT) military-grade version of the Microsemi ProASIC 3L flash FPGA device with integrated Arm processor was used [38]. MSR was designed to collect Martian rock samples and return them to Earth using three missions: a sample collection mission, a sample retrieval mission, and a return mission. The sample collection mission, Mars 2020 with the Perseverance rover, 7 scientific instruments, 3 robotic arms: main, Internal Bit Carousel & Robotic Transfer Arm, and the Sample Handling Arm (Shaw), 19 cameras, 2 microphones, and the Ingenuity Mars helicopter aboard was launched on July 30, 2020 and landed on Mars on February 18, 2021. Ingenuity was initially conceived to perform up to 5 test flights, at the end it was successful in 71 flights, and sustained rotor blade damage during its 72nd flight in January 18, 2024. The Perseverance rover uses Xilinx Virtex-5, -4, -2 Pro FPGAs across its main computing units, navigation systems, and scientific instruments, for example in PIXL (Planetary Instrument for X-ray Lithochemistry), an X-ray fluorescence spectrometer and high-resolution imager to map the composition of Martian surface materials, the Electra-lite UHF transceiver, a mast-

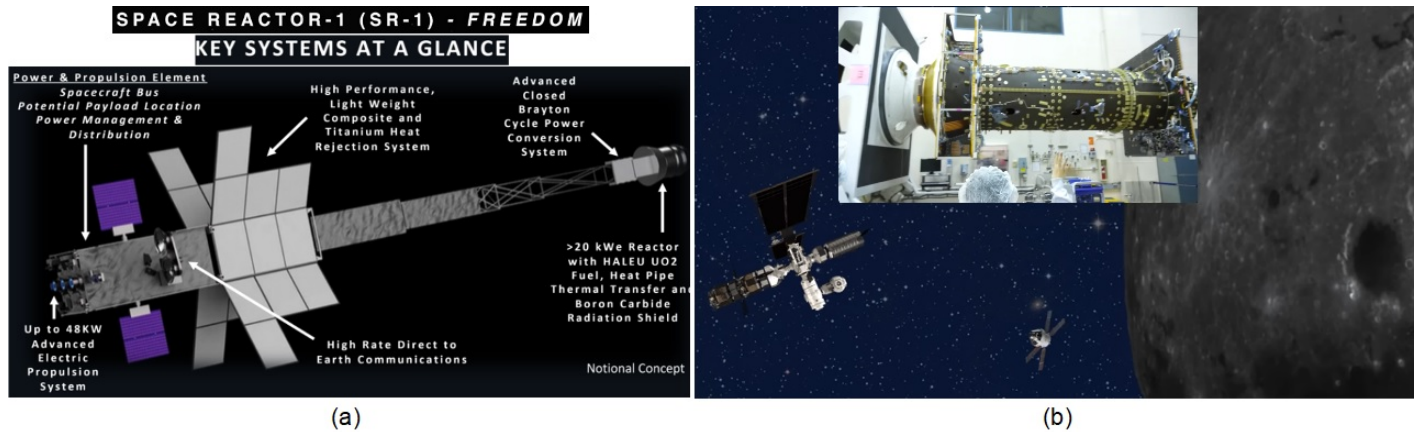


Figure 18: Space Reactor-1 (SR-1) – Freedom Mars Mission [NASA] (a) SR-1 Spacecraft (b) Power and Propulsion Module taken from the Lunar Gateway

mounted camera system for taking 3D pictures and video at high speed to examine distant objects called Mastcam-Z, SHERLOC (Scanning Habitable Environments with Raman and Luminescence for Organics and Chemicals), the radar-based Terminal Descent Sensor (TDS) that uses Ka-band radar to provide range and velocity measurements for Perseverance's Entry, Descent and Landing (EDL) after heatshield separation in the Martian atmosphere, as well as for vision processing tasks using the Vision Compute Element (VCE), which is a dedicated, specialized co-processor housed inside a 3-card, 6U CompactPCI (cPCI) chassis, on which to run intensive image processing algorithms required for Terrain-Relative Navigation (TRN) during Entry, Descent, and Landing (EDL), as well as enabling autonomous thinking-while-driving capabilities across the Martian surface. SIRF stands for S=Single-Event Effects, I=Immune, R=Reconfigurable, F=FPGA (Field Programmable Gate Array). The Virtex-5QV FPGA [39] is a space-grade reprogrammable device built to withstand harsh radiation environments without the need of an extra external redundant logic. The VCE is composed of the Power Conditioning Card (CEPCU), the RAD750 Single Board Computer (SBC), and the Computer Vision Accelerator Card (CVAC). CVAC is powered by a reconfigurable Virtex 5-SIRF, the XQR5VFX130.

The 2nd MSR mission, the sample retrieval mission, is the one that included the two Sample Recovery Helicopters (SRHs) to serve as a secondary backup method to retrieve cached Martian rock and regolith sample tubes from the surface of Mars and deliver them to the Sample Retrieval Lander (SRL). In particular, in the Ingenuity Mars helicopter avionics, the FPGA/Flight Controller Board (FFB) is at the heart of the Electronic Core Module (ECM) cube which hosts two redundant TI Hercules safety processors that serve as the low-level flight controller (FC), each of them having a dual-core lockstep ARM Cortex-R5F and ECC protected Flash and RAM. The two processors run in sync and are provided with the same clock and data by the FPGA, which interfaces to all sensors and actuators. The lockstep mechanism does cycle by cycle error detection. If a fault is detected, it signals the error to the FPGA; the FPGA switches to the other processor and power cycles the faulty one, so the flight control software continues to run without disruption. The analog signals are digitized by the two independent 12-bit ADCs in each of the flight controllers, cf. Figure 17. The RT military-grade ProASIC3/EL flash FPGA/SoC with Arm soft processor core used dramatically reduces dynamic power consumption and offers reprogrammability. Because it is nonvolatile, flash-based, it does not require a boot PROM, so there is no vulnerable external bitstream that can be easily copied. It utilizes a 128-bit flash-based lock and a separate AES key to secure programmed intellectual property (IP) and configuration data. After a NASA refocus announcement in March 2026, existing hardware, from the Power and Propulsion Element (PPE), of the Lunar Gateway [40], the concept of a space station around the Moon, will be reallocated to demonstrate highly efficient mass transport in space, with a spacecraft carrying 3 Ingenuity-class, called SkyFall helicopters [41] to explore Mars, specifically for ice-mapping and terrain scouting. The propulsion planned is nuclear

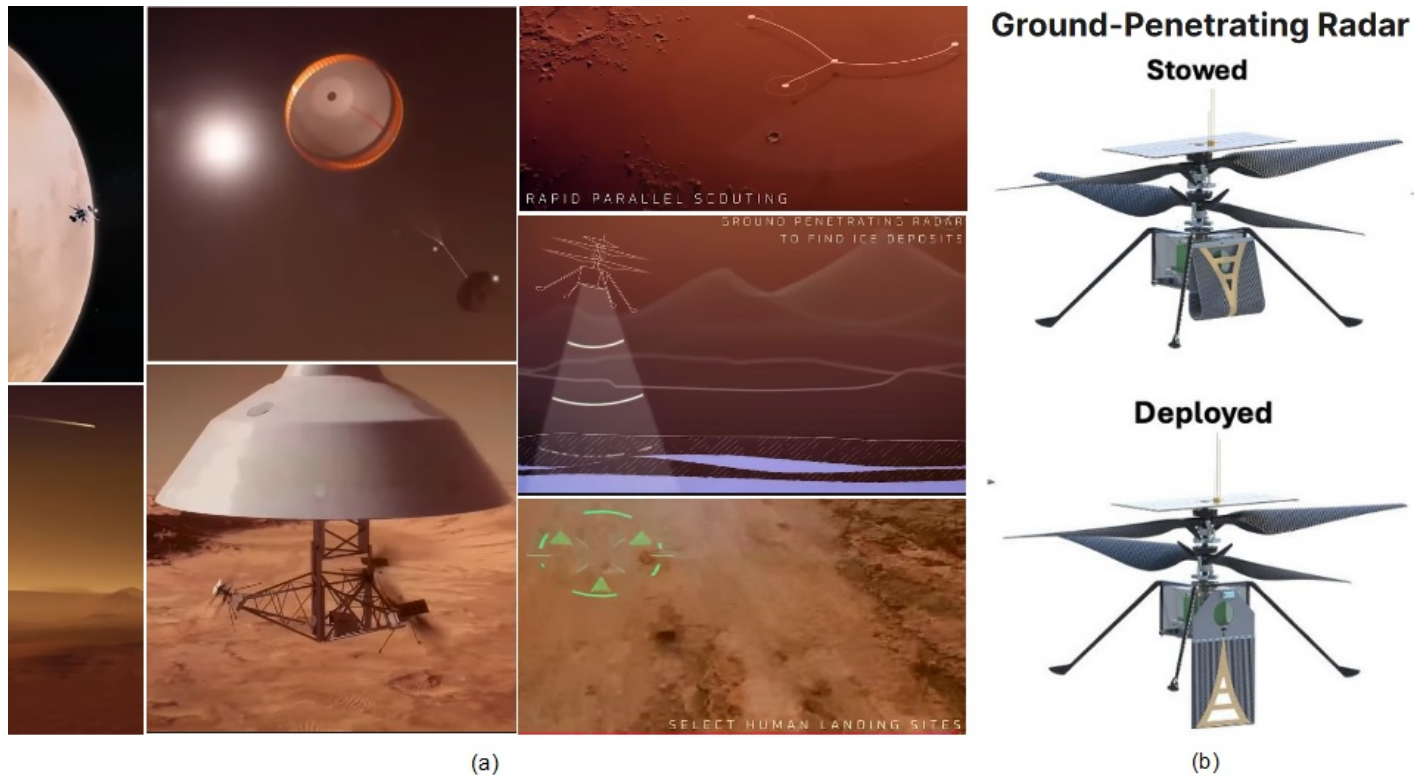


Figure 19: Space Reactor-1 (SR-1) – Freedom Mars Mission and SkyFall payload (a) Mission Entry, Descent, and Landing (EDL) (b) Mars SkyFall Helicopters (3) [NASA, JPL]

electric propulsion (NEP) which operates like a miniature nuclear power plant, using a live onboard fission reactor. The thermal energy produced by the reactor generates electricity, which is then used to power highly efficient electric thrusters. The new Mars helicopter blade design allows for more lift and greater payload capacity. The Skyfall maneuver to land on Mars consists of instead of having a conventional lander, all 3 helicopters detaching from the aeroshell and using their own power to land on Mars. During each flight later, the helicopters are to use cameras and ground-penetrating radar to map the local terrain, assess potential landing hazards and locate subsurface ice, which are important factors for a crewed Mars mission. Through their agility and speed, the exploration of a wider area becomes faster and at less cost. All 3 helicopters are identical, each carrying 4 scientific instruments: a ground-penetrating radar, a 13-Megapixel color camera, a near-IR camera, and meteorology as well as radiation sensors: to measure temperature, wind speed, and direction, and a radiation micro dosimeter. The 3 SkyFall Mars helicopters are expected to launch in late 2028, as a co-manifested payload on the Space Reactor(SR)-1 Freedom mission [42].

In this report a thorough analysis of the status quo of current and future technologies and mature engineering systems is performed with focus on shielding national security assets of the U.S. Armed Forces and the Intelligence Community (IC), in particular in integrated space autonomous systems, but not only. Two of a multitude of examples include hypersonic systems [43] and space communication networks [44]. Two examples of a large list of requirements/demands by the U.S. government can be found in past [45] (National Institute of Standards and Technology NIST) and more recent [28] (United States Space Force USSF) assessment documents. Applications of the findings and results in terrestrial, private, industrial scenarios are also feasible and straightforward. First, a clear delineation of new formal denominations like Super Intelligence (SI) for Artificial Intelligence (AI) or the relationship between security and safety for example in the context of cyberattacks is established and used advanced AI systems including those employing agentic AI technology are inspected in detail. The current unclear understanding of limitations and risks of AI is explained next covering the issues of regulatory governance matters to be legislated/executed by the U.S. federal government and the pertinent assessment of concrete lack of or mis-understanding in current events like the

OpenAI-Hugging Face incident, i.e., damage perpetuated by rogue AI agent swarms. Handling of related issues from a technological standpoint and how to design, build, and deploy those advanced systems including the agentic AI components is presented in detail.

References

- [1] V.D. Sánchez. *Neurocomputing 50th volume anniversary*. Neurocomputing, 50, ix, 2003. <https://profdrvdsaphd.lima-city.de/documents/Neurocomputing50thAnniversary.pdf>
- [2] V.D. Sánchez. *IEEE Fellow Award – "for leadership in neural and parallel computation, and pioneering contributions to autonomous space robots"*. 1995. <https://profdrvdsaphd.lima-city.de/documents/IEEEFellow.pdf>
- [3] V.D. Sánchez. *Personal written communication with the Office of the German Federal Minister of Research and Technology BMFT – Start of the First German National AI/ML Research and Technology Development Program*. 1988.
- [4] The White House, Presidential Actions, Executive Order. *Inaugurating The Era Of Super Intelligence*, September 29, 2026 . <https://www.whitehouse.gov/presidential-actions/2026/09/inaugurating-the-era-of-super-intelligence/>
- [5] V.D. Sánchez. *Advanced Gen AI & LLM Foundations and Applications – Paving the way to a more powerful and diverse ML*. December 2023. <https://profdrvdsaphd.lima-city.de/documents/AdvancedGenAILLMs.pdf>
- [6] V.D. Sánchez. *Architecting Next-Generation Generative AI/ML Models and Agentic AI Solutions*. August 2025. <https://profdrvdsaphd.lima-city.de/documents/ArchitectingNextGenAIMLAGenticAISolutions.pdf>
- [7] V.D. Sánchez. *On Artificial General Intelligence (AGI) and Artificial Super Intelligence (ASI)*. January 2025. <https://profdrvdsaphd.lima-city.de/documents/OnArtificialGeneralIntelligenceAndArtificialSuperIntelligence.pdf>
- [8] V.D. Sánchez. *Deep Impact of Advanced AI Technology Developments – Government Regulatory Measures*. September 2023. <https://profdrvdsaphd.lima-city.de/documents/ArtificialIntelligenceRegulation.pdf>
- [9] D.J. Trump. *White House Accord on Superintelligence – Joint Commitment on Frontier SI Responsibilities*, September 29, 2026 . <https://truthsocial.com/@realDonaldTrump/posts/117356435739432952>
- [10] The White House. Fact Sheet: *President Donald J. Trump Signs Historic Directive on AI in the National Security Enterprise*, June 5, 2026 . <https://www.whitehouse.gov/fact-sheets/2026/06/fact-sheet-president-donald-j-trump-signs-historic-directive-on-ai-in-the-national-security-enterprise/>
- [11] Central Intelligence Agency. *CIA Launches New Acquisition Framework to Turbocharge Collaboration with Private Sector*, February 9, 2026 . <https://www.cia.gov/stories/story/cia-launches-new-acquisition-framework-to-turbocharge-collaboration-with-private-sector/>
- [12] Central Intelligence Agency. *About CIA – Organization*, September, 2026 . <https://www.cia.gov/about/organization/>
- [13] Office of the Director of National Intelligence. *The AIM Initiative – A Strategy for Augmenting Intelligence using Machines*, 2019. <https://archive.dni.gov/files/ODNI/documents/AIM-Strategy.pdf>

- [14] Secretary of War. *Artificial Intelligence Strategy for the Department of War: Accelerating America's Military AI Dominance*. Memorandum, January 9, 2026. <https://media.defense.gov/2026/Jan/12/2003855671/-1/-1/0/ARTIFICIAL-INTELLIGENCE-STRATEGY-FOR-THE-DEPARTMENT-OF-WAR.PDF>
- [15] U.S. Department of War, Office of Strategic Capital. *Critical Technology Areas*, November 2025. <https://www.cto.mil/osc/critical-technologies/>
- [16] United States Space Force. *Data & Artificial Intelligence FY2025 Strategic Action Plan*, March 2025. https://www.spaceforce.mil/Portals/2/Documents/SAF_2025/USSF_Data_and_AI_FY2025_Strategic_Action_Plan.pdf
- [17] Hugging Face. *Anatomy of a frontier-lab agent intrusion*, July 27, 2026. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
- [18] OpenAI. *The Hugging Face incident and the road ahead*, August 26, 2026. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
- [19] OpenAI. *OpenAI- Hugging Face Incident*, Technical Report, August 26, 2026. <https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face>
- [20] Senate Committee on Homeland Security and Governmental Affairs. *Rogue AI: Securing the Homeland Against AI Agent Attacks*, Subcommittee on Disaster Management, District of Columbia, and Census, September 30, 2026.
- [21] Stanford University, Institute for Human-Centered Artificial Intelligence. *The AI Index 2026 Annual Report*. 9th Edition, 2026.
- [22] V.D. Sánchez. *Advanced Automation for mission-critical Information Technology past AIOps*. November 2024. <https://profdrvdsaphd.lima-city.de/documents/AdvancedAutomationMisCriITPastAIOps.pdf>
- [23] National Institute of Standards and Technology. *The United States Artificial Intelligence Safety Institute: Vision, Mission, and Strategic Goals*. May 21, 2024.
- [24] NVIDIA. *Agentic Autonomy Levels and Security*, Feb 25, 2025. <https://developer.nvidia.com/blog/agentic-autonomy-levels-and-security/>
- [25] NVIDIA. *OpenShell*. 2026. <https://docs.nvidia.com/openshell>
- [26] NVIDIA. *Open Agent Safety Platform: A Reference for Continuous In-Silicon Agent Monitoring*, September 28, 2026 . <https://developer.nvidia.com/blog/nvidia-open-agent-safety-platform-a-reference-for-continuous-in-silicon-agent-monitoring/>
- [27] NVIDIA. *NVIDIA Launches Space Computing, Rocketing AI Into Orbit*. March 16, 2026. <https://nvidianews.nvidia.com/news/space-computing>
- [28] United States Space Force. *Future Operating Environment 2040*. April 15, 2026. https://www.spaceforce.mil/Portals/2/Documents/SAF_2026/Future_Operating_Environment_2040_Final.pdf
- [29] Lockheed Martin. *Speed and Scale: Space Technology Trends Shaping the Future*. April 8, 2026. <https://www.lockheedmartin.com/en-us/news/features/2026/space-technology-trends-shaping-the-future.html>
- [30] AMD. *Space – Offering high bandwidth, high throughput support for earth-orbiting and interplanetary missions*. September 2026. <https://www.amd.com/en/solutions/aerospace-and-defense/space.html>

- [31] Microchip. *Radiation-Tolerant Field-Programmable Gate Arrays (FPGAs)*. September 2026. <https://www.microchip.com/en-us/products/fpgas-and-plds/fpgas-and-plds/radiation-tolerant-fpgas>
- [32] V.D. Sánchez. *Colonization of the Red Planet*. November 2022. <https://profdrvdsaphd.lima-city.de/documents/MarsColonization.pdf>
- [33] V.D. Sánchez. *Next Generation Missile Defense Systems including the Golden Dome*. May 2025. <https://profdrvdsaphd.lima-city.de/documents/NextGenerationMissileDefenseSystemsInclGoldenDome.pdf>
- [34] NASA. *Mars Sample Return*. September 2026. <https://science.nasa.gov/mission/mars-sample-return/>
- [35] NASA. *Mars 2020: Perseverance Rover*. September 2026. <https://science.nasa.gov/mission/mars-2020-perseverance/>
- [36] NASA. *Ingenuity Mars Helicopter*. September 2026. <https://science.nasa.gov/mission/mars-2020-perseverance/ingenuity-mars-helicopter/>
- [37] NASA. *Mars Sample Return – Sample Recovery Helicopters*. September 2026. <https://science.nasa.gov/mission/mars-sample-return/sample-recovery-helicopters/>
- [38] Microsemi. *Military ProASIC3/EL Low Power Flash FPGAs with Flash*Freeze Technology*. DS0106 Datasheet, October 2018.
- [39] Xilinx. *Radiation-Hardened, Space-Grade Virtex-5QV Family Data Sheet: Overview*. DS192 Datasheet, January 11, 2018.
- [40] NASA. *Gateway*. September 2026. <https://www.nasa.gov/mission/gateway/>
- [41] JPL. *SkyFall*. September 2026. <https://www.jpl.nasa.gov/missions/skyfall/>
- [42] NASA. *Space Reactor-1 Freedom*. September 2026. <https://www.nasa.gov/mission/space-reactor-1-freedom/>
- [43] V.D. Sánchez. *Hochentwickelte Hyperschalltechnologien für Raumfahrt- und terrestrische Anwendungen*. August 2023. <https://profdrvdsaphd.lima-city.de/documents/AdvancedHypersonic.pdf>
- [44] V.D. Sánchez. *Very high-speed space communication networks for building the next generation of advanced exploration and military spacecraft*. November 2023. <https://profdrvdsaphd.lima-city.de/documents/NextGenExplorationAndMilSpacecraft.pdf>
- [45] R. Ross. *Securing Space Systems and Assets – The Need for Multidimensional Protection in the 21st Century*. National Institute of Standards and Technology, February 24, 2021