PUBLIC SUBMISSION

Received: May 28, 2025
Tracking No. mb8-ea7i-zzvr
Comments Due: May 28, 2025 **Submission Type:** Web

Docket: NSF-2025-OGC-0001
NITRD_FRDOC_0001

Comment On: NSF-2025-OGC-0001-0001
Request for Information: Development of a 2025 National Artificial Intelligence Research and Development Strategic Plan

Document: NSF-2025-OGC-0001-DRAFT-0153
Comment on FR Doc # 2025-07332

Submitter Information

Organization: Purdue University

General Comment

On behalf of Purdue University, we appreciate the opportunity to provide comment on the Development of a 2025 National Artificial Intelligence Research and Development Strategy. Please see attached file.
Thank you,

Jennifer Wonder
Assistant Vice President, Strategic Initiatives
Office of Research
Purdue University

Attachments

Purdue University_NAIR_RFI_Response_May 2025

Physical AI: The Next Frontier in AI Research

Joerg Appenzeller, Aniket Bera, Dennis Buckmaster, Ignacio Ciampitti, Tamal Dey, David F. Gleich, Ananth Grama, Anand Raghunathan, Kaushik Roy, Wojciech Szpankowski, and Dongyan Xu

Purdue University.

Physical AI is the science, engineering, and responsible implementation of AI systems that are embedded within and strongly interact with elements of the physical world. In contrast to current AI systems that interact primarily with humans, physical AI systems must account for the constraints and characteristics of the physical world, while simultaneously accounting for human interfaces. Realizing the true potential of Physical AI hinges on fundamental developments in AI hardware, models, methods, software, infrastructure, and applications. Recognizing this significant need, Purdue University has initiated a signature program, the *Institute for Physical AI (IPAI)*, focusing in part on next-generation technologies that are beyond the commercial horizon of industry. Here, we highlight important areas in Physical AI in need of focused effort and investment to sustain and build on US' intellectual and economic leadership.

State of the art AI systems, commonly deployed as large language models (LLMs), have a number of drawbacks that prevent their use in critical application contexts for Physical AI. These include their lack of verifiability, generalizability, robustness, formal reasoning capability, high energy cost, and sample complexity. Furthermore, diverse deployment contexts (e.g., agentic/ distributed intelligence, potentially faulty compute platforms, real-time constraints, autonomous operation, and need for strong security guarantees) pose challenges that current systems do not adequately support. We propose three critical areas for focused investment in Physical AI: AI hardware, models and methods, and software and applications. These are summarized in Figure 1, highlighting a five-year plan in emerging technologies for Physical AI. We highlight a strong interplay between platforms, models and methods, software, and applications, nothing that isolated advances in any of these areas without corresponding development in others is unlikely to result in significant application impact.

Emerging Physical AI Platforms

Hardware support for critical AI applications must focus on the following five considerations: (i) significantly reducing energy consumption beyond current industry approaches; (ii) need for novel reasoning platforms in support of next generation of AI models; (iii) distributed and federated plat-

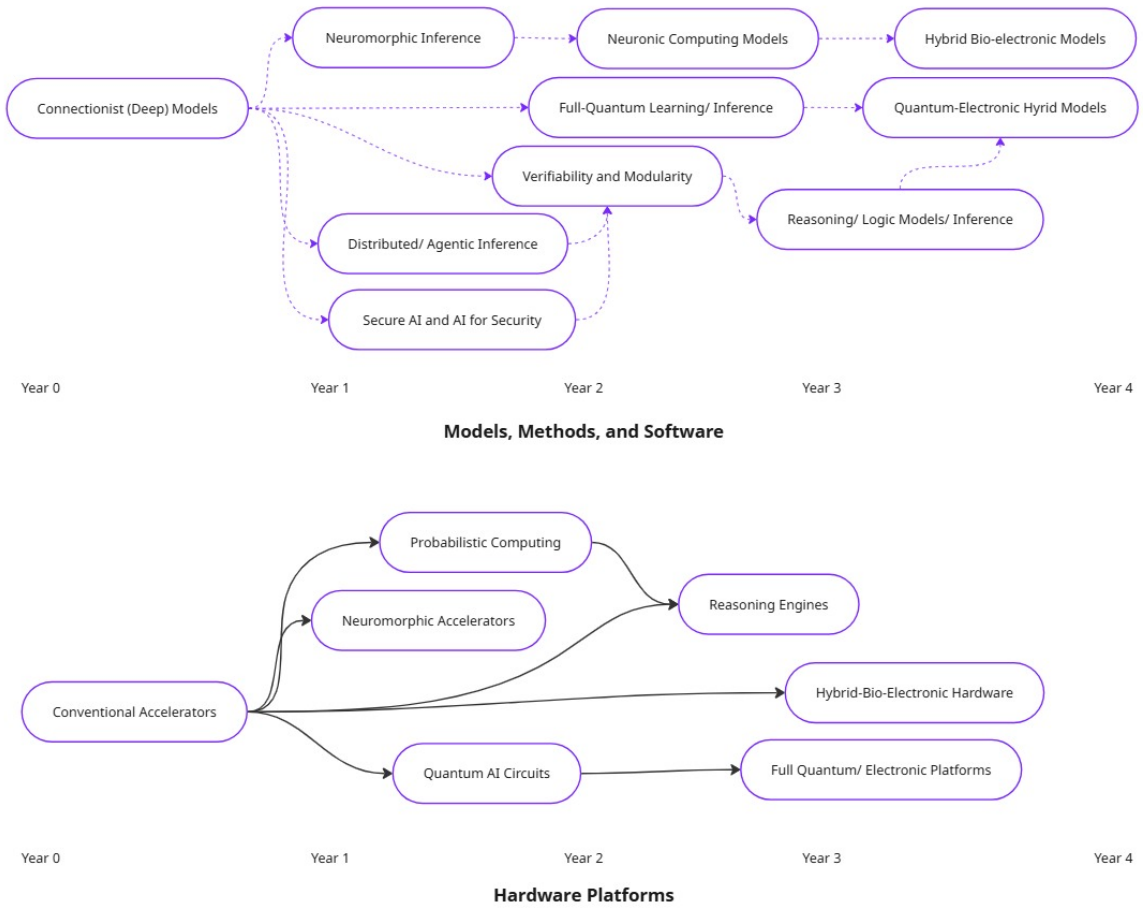

Figure 1: Proposed efforts in hardware platforms, models and methods, and software for AI over a five-year timeframe for sustained leadership.

forms for embedded and embodied intelligence; (iv) full quantum platforms with quantum sensing, communication, storage, and learning; and (v) integrating biological and electronic elements into hybrid platforms.

Neuromorphic Hardware. The human brain performs impressive feats (for example, simultaneous recognition, reasoning, control, and movement), with a power budget of nearly 20W. In contrast, a standard computer performing only recognition among 1,000 different kinds of objects expends about 250W. Although the brain remains vastly unexplored, its remarkable capability may be attributed to three foundational observations from neuroscience: vast connectivity, structural and functional organizational hierarchy, and time-dependent neuronal and synaptic modeling. Spike-based temporal processing allows sparse and efficient information transfer in the brain.

Despite this superficial resemblance, there exists a sharp contrast between the computing principles of the brain and silicon-based computers. A few key differences include: (i) the segregation of computations (the processing unit) and storage (the memory unit) in computers contrasts with the co-located computing (neurons) and storage (synapses) mechanisms found in the brain; (ii) the massive three-dimensional connectivity in the brain is currently beyond the reach of silicon

technology, which is limited by two-dimensional connections and finite number of interconnecting metal layers and routing protocols; and (iii) transistors are largely used as switches to construct deterministic Boolean (digital) circuits, in contrast to the spike-based event-driven computations in the brain that are inherently stochastic. Nevertheless, silicon computing platforms have been one of the enabling factors in the current deep-learning revolution. However, a major bottleneck impeding the realization of 'ubiquitous intelligence' (spanning cloud-based servers to edge devices) is the large energy and throughput requirement.

Guided by the brain, hardware systems that implement neuronal and synaptic computations through spike-driven communication may enable energy-efficient machine intelligence. Neuromorphic computing efforts originated in the 1980s to mimic biological neuron and synapse functionality with transistors, quickly evolving to encompass the event-driven nature of computations (an artifact of discrete 'spikes'). Eventually, in the early 2000s, such research efforts facilitated the emergence of large-scale neuromorphic chips. Today, the advantages and limitations of spike-driven computations (specifically, learning with 'spikes') are being actively explored by algorithm designers to drive scalable, energy-efficient 'spiking neural networks' (SNNs). In this context, we can describe the field of neuromorphic computing as a synergistic effort that is equally weighted across both hardware and modeling domains to enable spike-based artificial intelligence. The first challenge focuses on 'intelligence' (or algorithmic) aspects, including different learning mechanisms (unsupervised and supervised spike-based or gradient-descent schemes), while highlighting the need to exploit spatio-temporal event representations. The second challenge focuses on 'computation' (or hardware) aspects including analog computing, digital neuromorphic systems, beyond both von Neumann and silicon (representing the basic field-effect-transistor device that fuels today's computing platforms) technology. Finally, we must advance algorithm–hardware codesign, wherein algorithmic resilience can be used to counter hardware vulnerability, thereby achieving the optimal trade-off between energy efficiency and accuracy. Advances in neuromorphic computing hold the potential for orders of magnitude improvement in energy requirement, reasoning capabilities, and sample complexity – critical to future generation intelligent systems.

Embedded and Embodied Intelligence. Embedded and embodied intelligence integrates sophisticated AI methodologies into physically grounded and resource-constrained environments, targeting diverse deployments from robotic agents to broader intelligent systems. Embedded intelligence focuses on developing computationally efficient techniques, such as lightweight neural networks employing dynamic sparsification and adaptive quantization methods optimized for real-time operation on specialized hardware like neuromorphic chips, FPGA-based accelerators, and custom ASIC designs. This approach enables energy-efficient processing essential for edge computing and autonomous embedded platforms. Embodied intelligence addresses the capacity of agents to perceive, understand, and dynamically interact within their environments. This necessitates advancements in multimodal sensor fusion methods, real-time self-supervised learning frameworks, and context-aware hierarchical control architectures. Research priorities include modular neuro-symbolic models to enhance interpretability and generalizability, real-time predictive analytics leveraging advanced probabilistic and Bayesian inference techniques, and adaptive planning and control mechanisms with embedded formal verification approaches to assure safety-critical operations. Additionally, collective embedded intelligence among distributed, interacting agents requires innovative algorithms for secure decentralized decision-making, robust distributed infer-

ence frameworks, and scalable multi-agent coordination under uncertainty. Research efforts must explore graph neural network-based decentralized representations, distributed federated learning protocols with differential privacy guarantees, and rigorous formal analyses of collective emergent behaviors to achieve resilient and effective collaborative intelligence. Breakthroughs in these areas will significantly enhance intelligent systems' applicability and reliability in critical domains including industrial automation, autonomous transportation, healthcare, disaster response, and environmental monitoring, ultimately fostering resilient, adaptive, and trustworthy embodied AI systems.

Novel Probabilistic Computing Substrates. An important problem for current AI systems is the mismatch between evolving probabilistic computational needs and the entirely deterministic nature of current models that operate sequentially with a focus on precision. Contrary to the capabilities of current computer systems, many real-world problems solved using AI systems, such as drug discovery, data encryption and decryption, supply chain logistics, and large data handling, require hardware to embrace "uncertainty" in complex data to provide a weighted range of possible answers. The critical absence of such probabilistic computer hardware capabilities remains a barrier to accelerating next-generation computational power for AI and Machine Learning (ML) applications. This grand challenge defines the urgent need to build next-generation AI hardware beyond deep learning with new architectures that do not solely depend upon the use of silicon complementary metal oxide semiconductor (CMOS) technology.

To design and fabricate probabilistic computer hardware, naturally stochastic probabilistic materials and devices require use of true random number generation (TRNG) with suitable methods to demonstrate tunability. Fluctuation speeds, device dimensions, and energy efficiency are key areas of fundamental research required to integrate individual stochastic devices efficiently. Industry is unlikely to fund research on these fundamental questions due to these key gaps remaining in the design and fabrication of probabilistic hardware, but several companies have expressed strong interest in hardware accelerators if government funding can first expand upon the basic science of this field. While the global focus is on using existing CMOS technology and design layouts to improve on the performance of AI, the US can maintain its competitive leadership in AI with disruptive and energy efficient probabilistic hardware. Our military will also benefit from these advanced technologies used to protect confidential data from cyberattacks through enhanced encryption beyond that possible with current CMOS-based computing. A key aspect of p-computing decryption and encryption of data is the source of stochastic electrical behavior at room temperature, which may involve non-semiconducting materials. These stochastic advancements also include magnetic tunnel junctions (MTJs) that are naturally radiation hardened, making them ideal for defense applications. To meet the computational demands for ever increasing AI usage from industry and the public, p-computing can help alleviate energy concerns due to its reduced power consumption from use of stochastic magnetic tunnel junctions.

Fundamental advances in key scientific fields are critical to make probabilistic computing a reality: (i) Variability: Understanding how various parameters that characterize stochastic devices ultimately enter into the performance of probabilistic circuits (p-circuits) is critical to delivering the desired system-level and technology-level testbeds. While stochastic devices, by definition, are different from each other, it remains unclear what "similarities" the devices must show, i.e. what variability between devices is acceptable. (ii) Magnetic dynamics: Building on existing accom-

plishments using stochastic magnetic tunnel junctions (sMTJs), gaining a better understanding of the dynamics of coupled probabilistic bits (p-bits) are imperative. These insights will enable the engineering of fast, functional technology, improve circuit architectures, and optimize individual p-bit hyper-parameters. (iii) Geometry impact: The p-bit geometry is known to be a key player for the actual energy barrier between the two states that characterizes a magnetic tunnel junction (MTJ). A high resistive state is characterized by an antiparallel orientation between the magnetization of the two magnets constituting a conventional MTJ, while a low resistive state results from their parallel arrangement. Fluctuating randomly between these two states at room temperature constitutes the p-bit core. While the industry has a good idea of the geometric impact on stable MTJs, much less is known about the design rules for p-bits. Advancing the fundamental knowledge in this context will be central to the research effort. and (iv) Materials impact: The MTJ material greatly influences the quality of the tunable true random number generator (tunable TRNG), including the choice of materials for both magnetic layers, the insulating spin filter between them, as well as their respective thicknesses and geometric layout. Moreover, choosing magnets with in-plane anisotropy (IMA) versus perpendicular anisotropy magnets (PMA) influences the switching dynamics and, therefore, the switching speed of p-bits. The limited fundamental knowledge in this critical area restricts substantial advances in the development of probabilistic AI hardware.

Symbiotic Bio-Electronic Systems. Computing adjuvants – technologies that enhance the capabilities and performance of computing systems, hold tremendous promise for current and emerging applications. In contrast to conventional accelerators such as GPUs and FPGAs, adjuvants rely on fundamentally different physical, chemical, or biological processes to complement computational capabilities in fundamental ways. Examples of such adjuvants include photonic and quantum computers, DNA processors, and (brain) organoid learning platforms. Each of these technologies *complements* traditional semiconductor-logic-based processing to provide significant benefits – DNA computing excels at high-density durable storage, photonic and quantum computing have the potential to solve classically hard problems, and organoid intelligence provides significant energy efficiency and contextual reasoning. This representational and processing complementarity of adjuvants presents tremendous opportunities, while also posing profound challenges. Traditionally, adjuvants have been viewed from the perspective of classical computing models and data representations. Yet, we know, for instance, that (biological) cellular signals are markedly different from traditional data representations and associated stimuli, and neuronal processing is much more sophisticated than its current computational abstractions. The current myopic perspective on data and compute abstractions for adjuvants significantly limits the power of these technologies. To address this, we need an ambitious research and education program aimed at investigating bioadjuvants based on neural organoids as powerful new paradigms in problem-solving.

Novel Models and Methods for Physical AI

Novel models and methods must complement AI hardware, data characteristics, and applications' considerations. For example, while neuromorphic hardware has been shown to be up to three orders of magnitude more energy-efficient, these gains have not translated to application-level improvements due to representational and computational inefficiencies. Novel data representation and processing techniques are necessary to realize the potential of neuromorphic hardware.

Similarly, while there is widespread recognition of the need for reasoning in AI models, associated logic formalisms, rules of inference, scalable inference engines, and hardware interfaces are ill-understood. The potential of quantum AI models is widely recognized, however, techniques for full quantum-AI loops, specification of suitable Hamiltonians, appropriate quantum loss functions, optimization procedures, and associated sample complexity are unresolved. Embedded and embodied intelligence presents challenges for small-data learning, verifiability, modularity, and composition, and system-level assurances. Finally, emerging bio-electronic platforms pose significant challenges in application factoring to optimally leverage strengths of respective platforms. These and other challenges are organized into the following themes, which require focused effort and investment:

Accuracy, Robustness, and Generalizability. AI systems must generalize beyond training data in ways that are consistent with underlying specifications. Despite ongoing efforts, guarantees from such frameworks often rely on rudimentary complexity measures of the underlying hypothesis class. Hallucinations in LLMs are a glaring example, which fundamentally limit application of AI frameworks to critical applications. These shortcomings must be addressed through significantly refined characterizations that target a broad range of applications, and validated on important application classes.

Modern AI systems typically exhibit much more complex behaviors than classical theories can predict, including the fact that: (i) training data may have distributions that significantly diverge from the target distribution; (ii) data processed by the system may have heterogeneous structures; (iii) a model trained on one task may be applied to other tasks; and (iv) the hypothesis space may encompass more complicated structures (such as transformers), where it is not appropriate to view AI models as simple functions. We must derive new formalizations of generalizability for explaining, predicting, and guiding the design of AI models.

AI systems can be dramatically improved using models that have well-specified *semantics*, which can be represented using (learnable) logic, which notably covers data, rules, and (formal) linguistic statements as special cases. This motivates research to leverage the power of *reasoning*, and chaining of logical statements, rules, and facts, to reach conclusions that may lie far beyond the scope of observed training data, thus fundamentally addressing challenges inherent in current approaches.

Modularity, Composability, and Verification of Distributed Models. Complex connected AI systems, commonly encountered in critical systems, may violate important constraints relating to safety, performance, and robustness. Reasoning about connected AI systems must appropriately account for their interactions and emergent behavior. Verification and validation (V&V) of AI systems under complex temporal objectives and real-time constraints in highly dynamic and unpredictable environments can be, in general, intractable. Focused research and investment must be made on formal methods to verify robustness properties of AI systems. Current approaches tend to offer only component-level point solutions to specific design objectives. In the absence of effective compositional abstractions, it is not obvious how to relate component-level robustness of AI elements to system-level guarantees for connected AI systems. The challenges posed by the specification, verification, and validation of AI systems should be addressed within a unifying, formal, compositional framework, encompassing neural, or data-driven, representations as

well as symbolic, or model-based, representations. There is a critical need for investments in: (i) novel formalisms that capture heterogeneous AI system requirements, alignment objectives, and different forms of uncertainty while retaining computational efficiency; (ii) techniques for modular development, and (iii) AI system validation algorithms that scale to the complexity of forthcoming AI architectures, and quantify the risks.

Full Quantum AI Models. Quantum systems have immense potential for fundamentally advancing Artificial Intelligence (AI) and conversely, AI provides critical technologies for advancing quantum systems. While these fields have made tremendous progress in the recent past, their advances have largely been independent of each other. A focused *co-design effort* that synergizes developments in quantum sciences and quantum AI has the potential to revolutionize both areas and enable the next generation of systems bringing new insights, reason in new ways, and solve important socio-economic challenges. Realizing this potential now requires fundamental advances in quantum materials, devices, computing, and novel AI methodologies.

Current AI approaches for quantum systems rely on variational models in which classical optimizers are used to parametrize quantum circuits in noisy intermediate-scale quantum (NISQ) platforms. Variational models are motivated by challenges in stable storage of qubits for training, reliable communication, and transduction of qubits across quantum platforms (e.g., photonic communications and trapped-ion circuits). Current variational circuits pose significant challenges relating to noise mitigation and tolerance, non-convex optimization landscapes, the Barren Plateaus problem, sample complexity, and scalability to large circuits. To address these challenges, we must invest in the design of novel quantum materials and circuits, and full quantum AI solutions that do not rely on classical elements for training models.

Full quantum AI models have tremendous potential for vastly more efficient and accurate parametrization of quantum circuits – they have smoother objective function landscapes, do not introduce measurement errors, and are capable of scaling to full quantum models supported by emerging platforms (10K qubits and beyond). Realizing these benefits will provide massive impetus for the development of new quantum materials, design of novel quantum circuits, and feed-back into powerful full-quantum AI models. A key innovation that we aim to enable through these full quantum models is to empower the next generation of reasoning engines capable of highly sophisticated and scalable logic descriptions and inference mechanisms. Supporting these advances in quantum AI are powerful new technologies in quantum materials, devices, computation, and communication. These advances rely on AI techniques that will overcome challenges in quantum devices related to efficiency, controllability, and scalability. Building on these innovations, efforts must focus on solving important challenges relating to error mitigation, long-time coherence and quantum storage, and robust quantum sensing and transduction.

Integrating quantum devices with proposed quantum AI models and methods requires a full stack of solutions associated with design of ansatz (quantum circuit), ancillary technologies for error control and mitigation, suitable measurement-guided robustness, quantum coding techniques for reliable communication, novel quantum logic and semantics, and algorithms for key kernels. In the short term, effectively leveraging quantum-classical hybrid platforms requires support for variational quantum algorithms, efficient quantum resource management, and scalable decoding for quantum error correction. In the medium to long term, the efforts must focus on novel formulations of optimization of full-quantum AI loss functions, quantum sample complexity and gener-

alizability, and integrated learning and noise mitigation. The goal of these efforts is to provide a theoretical and systems' foundation for efficient execution of full-quantum AI models on stable, efficient, and robust platforms.

System Software for Physical AI

AI platforms must be supported by flexible and powerful system software to enable efficient and effective application deployment. In addition to providing facile interfaces between applications and hardware, system software must provide security and modeling substrates for robustness, verifiability, composition, adaptation, and elasticity. These core services are organized into the following research themes:

Robustness, Composability, and Verifiability. As AI systems increasingly find use in embedded environments, software substrates must support effective information exchange and orchestration of agents operating (semi)autonomously. Such interfaces must account for real-time requirements, fault tolerance (e.g., if an AI agent stops responding), compositional reasoning (e.g., exchange local guarantees to ensure global system invariants), and resource allocation (e.g., ensuring that real-time model retraining has the data, compute, and communication resources). Current support for AI systems offers minimal support for complex connected components, leading to significant gaps in our ability to reason about their behavior. Efforts and investments must be made to develop software systems that allow development, deployment, and efficient operation of connected AI systems as an integrated unit. This would require standardization of interfaces, information semantics and data formats, and specification of accuracy, robustness, and performance invariants, along with integration of environmental sensing and actuation capabilities. This is a critical need for supporting AI applications in complex real-world settings.

AI System Security. Ubiquitous use of AI in critical systems exposes massive infrastructure vulnerabilities, requiring a strong focus and investment in security of AI infrastructure – from hardware and system support to software and application verification. Attacks such as data poisoning and model trojans are currently being investigated. However, connected and embodied AI systems expose significant new attack surface at the interface of AI systems and between the integrated AI system and the underlying physical plant. A comprehensive research program on the security of connected AI systems that integrates physical plant models, software systems, and AI methodologies is critical to safe and secure operation of future generation smart infrastructure.

Novel AI Applications

While AI applications in domains such as manufacturing, materials processing, supply chains and urban environments are well-recognized, areas such as agriculture and rural-upliftment are underserved. We advocate for sustained effort and investment in AI for the rural sector, which has traditionally been neglected by large-scale private-sector technology investment.

Agriculture is a logistics business. Input suppliers, farmers, commodity firms, processors, transportation companies, and even retailers need to have the right product, person, tool, machine, etc. at the right place at the right time. Labor is a major constraint in many agricultural settings and

AI can serve as an important decision-support tool to complement human decision-making. When used effectively to improve logistics, financial analysis, or strategic decisions, AI can increase the efficiency of labor and other inputs to improve sustainability and profits. While AI can facilitate optimization of subsystems, it can be difficult to incorporate all the physical constraints into these models, such as labor availability, machine locations and speeds, soil moisture, weather forecasts, etc. In typical settings, the physical environment is highly variable, but there are natural ranges for the many biological, chemical, and physiological variables. Incorporation of these constraints into AI has been gradual and generally tightly constrained to specific functions.

For example, current AI/ML applications include the identification of livestock, weeds, crop nutrient and water status, or product quality, and can directly inform or automate specific actions such as sorting, spraying, very precise tillage, etc. However, the data collected as part of this operation is rarely shared for other uses of the same information; even when it is accessible and reusable, interoperability is still a challenge due to the siloed systems. Strong Ag-AI requires massive data for training and a reasonably complete context for interpretation. Even when effective models are generated, implementation may require data streams for inference that are not widely available in commercial production or processing operations.

A key hurdle for AI in agriculture is data access, which starts with the data pipeline. Modern production machines can serve as mobile sensing, computing, and communication platforms, but connectivity in rural areas is often poor and these machines do not often traverse the landscape. Aerial platforms, such as satellites or drones, are a complement, but these datasets have spatiotemporal challenges in their resolution and require specialized skills for interpretation. Progress is being made with in-situ (planted) sensors in soils, on plants, in containers, etc., but most platforms for collecting and analyzing such data are siloed in proprietary platforms, making integration and more holistic analysis difficult or very time consuming. In addition, because many decisions in the agricultural value chain are time-sensitive, we are increasingly dealing with data streams (status) on top of data sets (historical or current). This data must be time indexed, and there is usually an identity index (geolocation or animal ID) that is fundamental to aligning and using such data. Digital agriculture is an immature field with many competing data standards and structures. Collecting complete contextual metadata in production situations is cumbersome and often involves a variety of disciplines and sources. Therefore, robust, complete, and high-quality curation of agricultural training data is a major challenge. One opportunity to be seized is the use of AI on the front end to improve consistency and reduce the tedium of collecting, describing, and curating data that can drive AI for decision making and autonomy on the back end.

Automation and autonomous vehicles in agriculture face a complex challenge beyond navigation in ever changing environments in that the proper adjustment of a plethora of machine settings to successfully complete tasks must involve awareness of the size, texture, color, behavior, etc. of the plants, soil, and animals involved. On top of general machine functions, navigation, and other AI training, these machines will require task and situation specific training comparable to skilled workers and it often involves multiple senses and modalities.

Finally, the Food Safety and Modernization Act (FSMA) has an additional traceability requirement that goes into effect in 2026. There are several characteristics associated with products in the agricultural value chain that require traceability. AI can help with traceability of whole foods, ingredients, and processed products or meat, but there are gaps in robustness, privacy and security, and accuracy that must be addressed.