

AI Fusion: Enabling Distributed Artificial Intelligence to Enhance Multi-Domain Operations & Real-time Situational Awareness

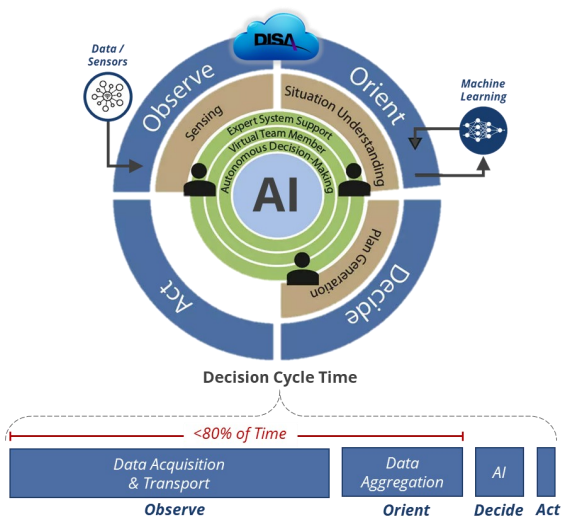
The pace of innovation in Artificial Intelligence (AI) is completely unprecedented, and it continues to accelerate year after year. With research and advances in key technologies such as machine learning, computational game theory, and autonomy, today's AI is capable of augmenting humans and increasing productivity and efficiency for critical tasks that just two years ago would have been impossible. For the Department of Defense and the Intelligence Community, this innovation will significantly enhance situational awareness and decision-making by fusing information from systems and sensors across multiple domains—from the enterprise to the edge of the battlefield—to maximize mission effectiveness, reduce risk, and save lives.

Current Limitations of AI for Multi-Domain Operations:

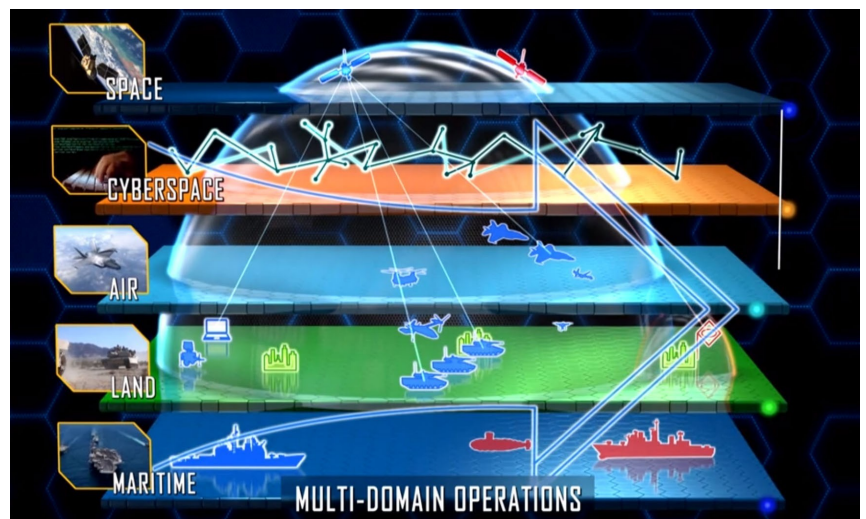
Even with the incredible advances over the past few years, the current approach to leveraging AI for multi-domain operations is still limited in its ability to provide decision makers with real-time situational awareness and to respond quickly and accurately to imminent activities or threats.

- AI requires massive data aggregation and storage from deterministic systems and sensors across the battlefield in a highly dynamic and contested environment.
- The transport and aggregation of massive amounts of data from platforms and sensors operating at the edge to tactical HPC nodes and to the enterprise cloud requires extensive and persistent, high-bandwidth connectivity.
- The warfighter must perform extensive data engineering to consolidate the data into a cohesive ontology to enable processing by AI algorithms. This requires considerable time and supervision by the warfighter to enable AI.

One of the key advantages that AI offers is human augmentation and enhanced decision support as part of the traditional 'OODA Loop' shown below. Today, excessive time and human supervision must be spent acquiring, transporting, aggregating, and engineering the massive data sets required for AI to be accurate and effective and to compensate for uncertainty. When this critical dependency on data aggregation is combined with highly dynamic and opportunistic communications at the edge, it creates a crucial vulnerability that adversaries could exploit to undermine the impact or availability of AI in multi-domain operations—simply by denying or degrading AI's access to critical data being collected or stored on military platforms and sensors operating across the various domains of the battlefield.



Original Source: NATO JAPCC Journal

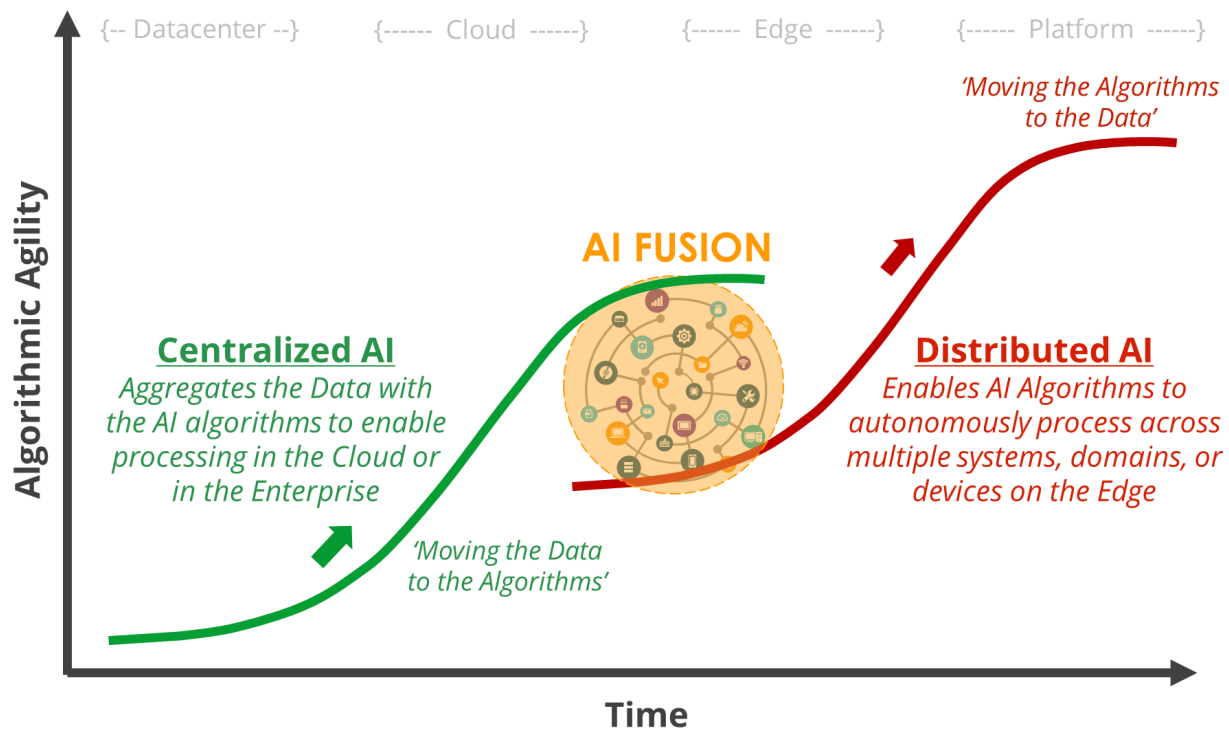


Source: Army Futures Command

The highly deterministic nature of data aggregation and engineering required for AI today greatly limits the ability to adapt to and integrate data sources that may become available on the battlefield over time as a result of US, allied, or coalition partner activities or coordinating multi-domain operations. The ability to access or share data between the US and our allies and between different US military services is also heavily constrained by the varying levels of networks and systems that protect the security, confidentiality, and integrity of the data as well as the sensors collecting and/or storing it. This greatly complicates the aggregation of massive data sets across different classifications and networks, and the ability of AI to provide timely, accurate insights for multi-domain operations with an acceptable level of uncertainty and reliability. In order to build trust through human-machine teaming, AI must work at the speed of relevance—while also compensating for the dynamic and opportunistic communications across the battlefield.

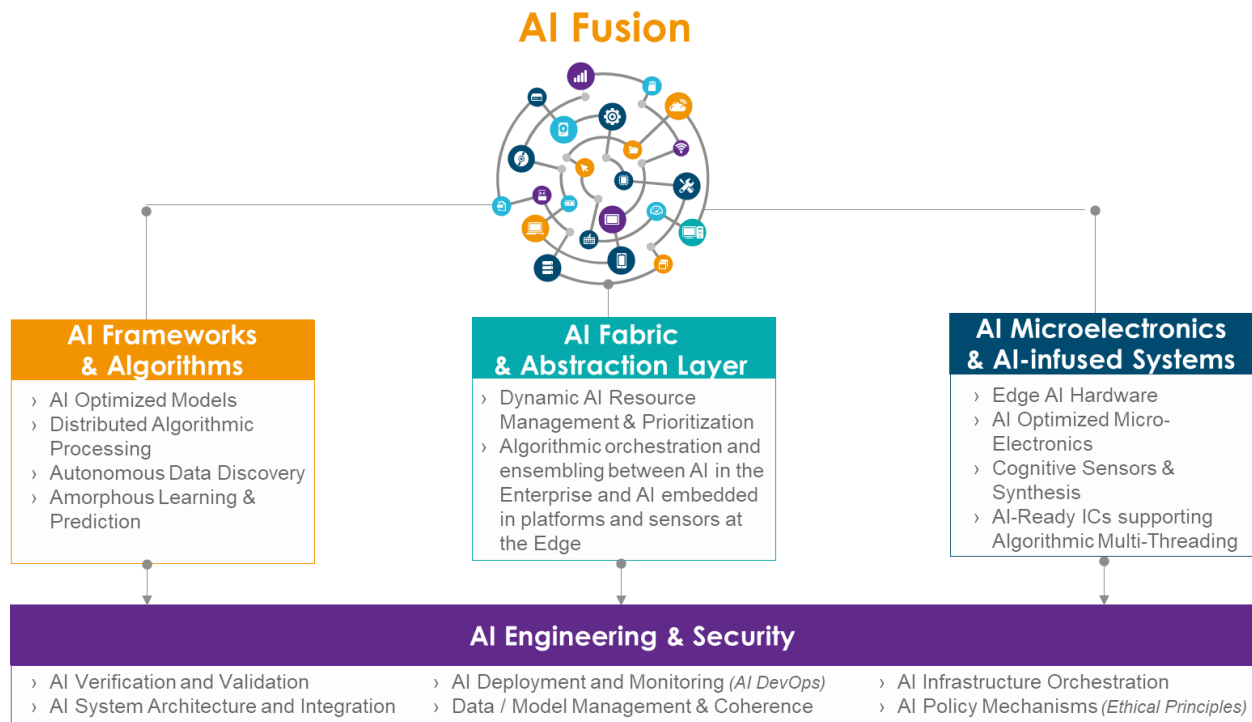
Accelerating AI Fusion for Multi-Domain Operations:

The pace of innovation in AI is not slowing down, but rather continues to accelerate each and every year. Carnegie Mellon envisions a bold, new transformation in AI over the next 6-8 years that will allow it to evolve from today's highly structured and controlled, centralized architecture to a much more adaptive and pervasive, distributed architecture that autonomously fuses AI capability between the enterprise, the edge, and AI-infused systems embedded on-platform. We call this transformation AI Fusion. AI Fusion will minimize AI's dependency on aggregating and engineering massive data sets and the need to 'move data to the algorithms' which are then processed in the cloud or in an enterprise datacenter. Instead, AI Fusion will leverage algorithmic agility to enable autonomous data discovery and have 'algorithms move to the data'. The ability to process data at the edge or on-platform will drastically reduce the need for persistent, high-bandwidth connectivity to transport data, and monolithic networks and deterministic systems to connect decision makers with critical platforms operating across multiple domains or as part of allied or coalition partner activities.



AI Fusion Integrated Research Thrusts:

Precursors to AI Fusion are already being seen with recent advances in federated learning and microelectronics optimized for neural networks—but to truly unlock the potential of Distributed AI for Multi-Domain Operations will require integrated research across four critical thrusts and the co-design and development of AI hardware/software to enhance algorithmic agility and enable distributed algorithmic processing and ensembling:



- 1) **AI Frameworks & Algorithms**—Enabling algorithmic agility and distributed processing will require the development of new theoretical frameworks and algorithms that extend autonomous discovery and processing of disparate data beyond the current limits of federated learning, information theory, and meta learning. With these advances, the cloud will serve as an enabler for algorithmic mapping and orchestration between the enterprise and varied military platforms and systems operating at the edge.
- 2) **AI Fabric & Abstraction Layer**—The co-design and development of an AI Fabric is critical to facilitate distributed algorithmic processing and ensembling between the enterprise and the edge. Extensive research into novel mathematical theorems and frameworks, based on stochastic analysis and models of distributed systems, is necessary to ensure the performance, prioritization, scheduling, resource allocation, and security of the new AI algorithms—especially with the very dynamic and opportunistic communications associated with military operations in contested environments.
- 3) **AI Microelectronics & AI-Infused Systems**—Supporting dynamic, autonomous AI processing at the edge and on-platform will require extensive research into novel architectures, processing, and connectivity for AI microelectronics. More importantly, extensive research in the co-design and development of the AI microelectronics with the AI Algorithms & Frameworks and the AI Fabric is needed to support algorithmic multi-threading on a single embedded chip or AI-infused system or sensor on-platform, and to enable scalable training, inferencing, and prediction for military platforms and sensors operating at the edge across multiple operational domains.
- 4) **AI Engineering & Security**—With the exponential increase in AI applications and deployments, extensive research is needed to establish a new ‘AI Engineering’ discipline for developing resilient, reliable, and secure AI systems. Simply put, AI Engineering and Security brings ‘*confidence in capability*’—knowing when AI systems are going to work and when to fix them—across the AI Fusion research thrusts, a task made more difficult as we embrace algorithmic agility and distributed processing and fuse AI capabilities between the enterprise, the edge, and on-platform AI-infused systems operating across multiple domains.

Transforming the Foundation of the AI Stack

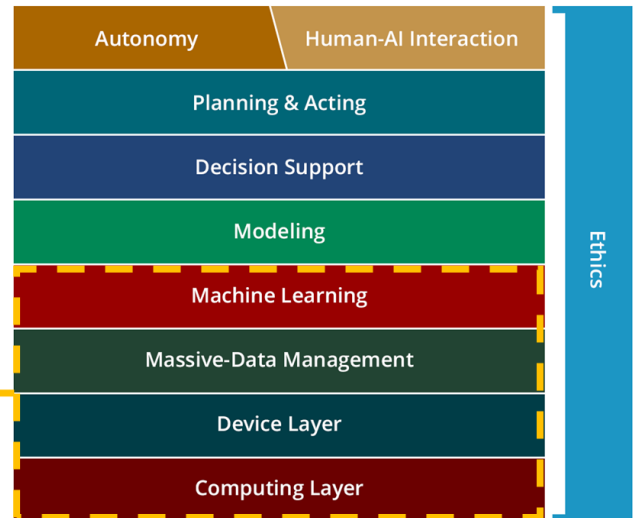
In 2016, Carnegie Mellon created the AI Stack as a technical blueprint to develop and deploy Artificial Intelligence. The premise of the AI Stack is simple—AI isn't just one thing. It's built from technology blocks that work together to enable AI. The AI Stack can also be thought of as a toolbox—each block houses a set of technologies that scientists and researchers can reach for as they work on new projects and initiatives. Each technology block depends on the support of the blocks beneath it, and enhances capability in the blocks above it.

In a traditional, Centralized AI architecture, all of the technology blocks would be collocated or combined in the cloud or a single enclave to enable AI. For Distributed AI, AI Fusion will transform the very foundation of the AI Stack by enabling transformational advances in AI Theory, Frameworks, and Algorithms; AI Micro-Electronics and AI-Infused Systems; and an AI Fabric & Abstraction Layer that will fuse the distributed capabilities together to enable dynamic, autonomous AI processing on the edge or on-platform.

Algorithmic agility and distributed processing will enable AI to perceive and learn in real-time by parallelizing these critical AI functions across disparate systems, platforms, sensors, and devices operating at the edge.

AI Stack

AI FUSION



AI FUSION



Fusing AI capability from the Enterprise to the Edge, and enabling AI to be embedded On-Platform and support parallel and distributed algorithmic processing

Creating AI Frameworks & Algorithms that can autonomously discover and 'Move to the Data' - eliminating the need for massive data aggregation and engineering to enable AI

Driving the co-design and development of AI Microelectronics with AI Algorithms and an immersive AI Fabric that fuses them all together and re-defines the art-of-the-possible in AI

Enabling the Future of Multi-Domain Operations

By allowing for distributed algorithmic processing and the autonomous discovery and integration of new data sources, AI Fusion will drastically reduce the human intervention required to enable AI. Being able to autonomously fuse AI capabilities between the enterprise, the edge, and AI-infused systems and sensors embedded on-platform will enable real-time situational awareness and enhanced decision support. AI Fusion is all about accelerating transformational AI capabilities to connect any system or sensor across any domain to any decision maker that would benefit or gain insights from the information.

AI FUSION CORE RESEARCHERS: AI FRAMEWORKS & ALGORITHMS



Dr. Gauri Joshi

ASSISTANT PROFESSOR, ELECTRICAL & COMPUTER ENGINEERING

gaurij@cmu.edu

Dr. Gauri Joshi's research seeks to democratize machine learning by designing distributed ML algorithms that are system-aware (robust to computation and communication limitations, and seamlessly scale to thousands of devices) and data-aware (can handle heterogeneous amounts of statistically skewed data).

KEY AREAS: Distrib. ML; information theory; ML/cloud infrastructure (PDL)



Dr. Soumya Kar

PROFESSOR, ELECTRICAL & COMPUTER ENGINEERING

soumyak@cmu.edu

Dr. Soumya Kar's work focuses on theory and methods for large-scale decentralized inference, optimization and machine learning. A key thrust is in the design of secure algorithms and architectures in highly distributed IoT-type settings that are resilient to adversarial attacks.

KEY AREAS: Distributed inference, optimization, & machine learning; resilient & secure distributed algorithms



Dr. Virginia Smith

ASSISTANT PROFESSOR, MACHINE LEARNING

COURTESY APPOINTMENT, ELECTRICAL & COMPUTER ENGINEERING

smithv@cmu.edu

Dr. Virginia Smith's work addresses challenges related to optimization, privacy, fairness, and robustness in distributed settings. Her work aims to make federated and on-device learning safe, efficient, and reliable.

KEY AREAS: Federated learning, distributed optimization, privacy, fairness



Dr. Ameet Talwalkar

ASSISTANT PROFESSOR, MACHINE LEARNING

talwalkar@cmu.edu

Dr. Ameet Talwalkar's research focuses on underlying modeling challenges associated with machine learning in federated settings. He aims to devise principled methods to tackle the novel statistical challenges in these settings stemming from data heterogeneity, while simultaneously respecting the underlying systems constraints related to energy, computation, communication and privacy.

KEY AREAS: Federated learning, meta learning, distributed deep learning, MLSys (intersection of systems & machine learning)



Dr. Rashmi Vinayak

ASSISTANT PROFESSOR, COMPUTER SCIENCE
COURTESY APPOINTMENT, ELECTRICAL & COMPUTER ENGINEERING
rvinayak@cs.cmu.edu

Dr. Rashmi Vinayak's research interests broadly lie in two areas: information/coding theory and computer/networked systems. Her current focus is on addressing reliability, availability, scalability, and performance challenges in large-scale distributed systems, with a key thrust on systems for machine learning. Her research involves designing solutions rooted in fundamental theory as well as building systems that employ the resulting insights to advance the state-of-the-art.

KEY AREAS: Resource-efficient resilience for distributed systems including systems for machine learning; Information theory & algorithms

AI FUSION CORE RESEARCHERS: AI FABRIC & ABSTRACTION LAYER



Dr. Mor Harchol-Balter

PROFESSOR, COMPUTER SCIENCE
COURTESY APPOINTMENT, TEPPER SCHOOL OF BUSINESS
harchol@andrew.cmu.edu

Dr. Mor Harchol-Balter works on optimal scheduling and resource allocation algorithms for distributed computing systems. The goals of these algorithms include: minimizing latency, minimizing power/capacity, maximizing fairness, and other performance-related metrics.

KEY AREAS: Theory & frameworks for efficient scheduling & resource allocation of parallel jobs



Dr. Mahadev Satyanarayanan (Satya)

CARNEGIE GROUP UNIVERSITY PROFESSOR, COMPUTER SCIENCE
satya@cmu.edu

Dr. Mahadev Satyanarayanan's current research focuses on how edge computing can support low-latency, compute-intensive, bandwidth-scalable, and privacy-sensitive AI algorithms in areas such as computer vision. He is particularly interested in the intersection of edge-based Augmented Reality (AR) and AI.

KEY AREAS: Edge computing; system performance, scalability, availability & trust in edge computing/cloudlets



Dr. Vyas Sekar

PROFESSOR, ELECTRICAL & COMPUTER ENGINEERING
vsekar@andrew.cmu.edu

Dr. Vyas Sekar's research in Cylab is looking at new AI- and ML-driven ways to improve the security and performance of systems with broad applications to Internet-of-Things, content distribution, software-defined networking, and network functions virtualization.

KEY AREAS: Data-driven networking & security; IoT systems & security



Dr. Srinivasan Seshan

PROFESSOR & DEPT. HEAD, COMPUTER SCIENCE
COURTESY APPOINTMENT, ELECTRICAL & COMPUTER ENGINEERING
srini@cs.cmu.edu

Dr. Srinivasan Seshan's research focuses on distributed systems and networking protocols. He is a pioneer in the design of applications that adapt to network conditions and incorporate Internet-scale sensor data collection. His recent work explores edge computing and in-network computing system designs that address key security and availability issues that occur in many common networked systems.

KEY AREAS: Distributed systems; next generation network architectures



Dr. Peter Steenkiste

PROFESSOR, COMPUTER SCIENCE
PROFESSOR, ELECTRICAL & COMPUTER ENGINEERING
prs@cs.cmu.edu

Dr. Peter Steenkiste's research is broadly speaking in networking. His more recent work is future internet architecture, information-centric networking, wireless, and edge computing.

KEY AREAS: Networking & distributed systems; pervasive computing

AI FUSION CORE RESEARCHERS: AI-INFUSED SYSTEMS & HARDWARE

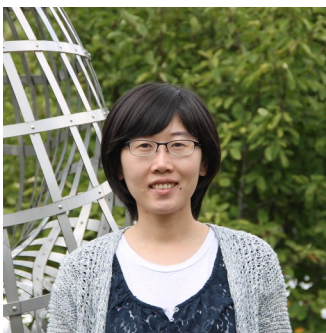


Dr. Shawn Blanton

TRUSTEE PROFESSOR & ASSOC. DEPT. HEAD FOR RESEARCH, ELECTRICAL & COMPUTER ENGINEERING
rblanton@andrew.cmu.edu

Dr. Shawn Blanton's research is in the general area of hardware design and fabrication. A major theme of his current projects focuses on designing formally-secure hardware systems, especially those centered on accelerating AI algorithms.

KEY AREAS: Data-mining algorithms, data analysis, chip design/fabrication; IC design-enabled ML acceleration; trustworthy distributed systems based on integrated circuits



Dr. Yuejie Chi

ASSOCIATE PROFESSOR, ELECTRICAL & COMPUTER ENGINEERING
yuejiec@andrew.cmu.edu

Dr. Yuejie Chi's research focuses on the theoretical and algorithmic foundations of data science, signal processing, machine learning and inverse problems, with applications to high-resolution parameter estimation in sensing and imaging, distributed and streaming information processing, and bioinformatics.

KEY AREAS: Distributed stochastic optimization w/heterogeneous data & privacy guarantees enabling scalable, real-time processing of high-dimensional heterogeneous data at the edge



Dr. Franz Franchetti

PROFESSOR & FACULTY DIR. OF ITS, ELECTRICAL & COMPUTER ENGINEERING

franzf@ece.cmu.edu

Dr. Franz Franchetti's research focuses on automatic performance tuning and performance portability for emerging parallel platforms and algorithm/hardware co-synthesis. He leads the open source SPIRAL project and aims to establish it as the AI system for automatic optimization of computational programs, targeting both classical scientific computing and signal processing as well as emerging AI/ML algorithms. He led or leads four DARPA projects in the PAPP, BRASS, HACMS, and PERFECT programs and is PI/Co-PI on a number of federal and industry grants.

KEY AREAS: Automatic performance tuning & performance portability for emerging parallel platforms & algorithm/hardware co-synthesis



Dr. Saugata Ghose

SYSTEMS SCIENTIST, ELECTRICAL & COMPUTER ENGINEERING

ghose@cmu.edu

Dr. Saugata Ghose's work is on designing new computer architectures and systems from the bottom up to efficiently handle large amounts of data. These new systems eliminate much of the chip-to-chip data movement that takes place during model inference and training, which can allow for ultra-low-power, high-performance computation and enable new uses for mobile computing and smart sensors.

KEY AREAS: Co-designing low-precision ML algorithms & processing-using-memory architectures enabling heavy-weight local computation for ML/AIDr.



Dr. Swarun Kumar

ASSISTANT PROFESSOR, ELECTRICAL & COMPUTER ENGINEERING

swarun@cmu.edu

As head of the Emerging Wireless Technologies (WiTech) lab, Dr. Swarun Kumar's research applies AI/ML techniques to next-generation wireless systems and mobile services. He designs and builds efficient and secure systems for the Internet of Things and cellular networks beyond 5-G.

KEY AREAS: AI/ML techniques for faster, better networking; optimizing wireless communication for AI/ML workloads



Dr. Tze Meng Low

ASSISTANT RESEARCH PROFESSOR, ELECTRICAL & COMPUTER ENGINEERING

lowt@andrew.cmu.edu

Dr. Tze Meng Low's research focuses on the design of portable and efficient implementations of AI models for both training and inference. He targets traditional high performance and resource-constrained computing platforms, and the increasing number of AI accelerators introduced.

KEY AREAS: Algorithmic agility, SWaP/Computing on the edge



Dr. Brandon Lucia

ASSOCIATE PROFESSOR, ELECTRICAL & COMPUTER ENGINEERING

blucia@andrew.cmu.edu

Dr. Brandon Lucia's lab focuses on the intersection of computer architecture and hardware, and software systems with the goal of bringing ML/AI-ready computational capabilities to, and beyond, the edge. He works on intermittent and energy-harvesting computer systems, smart and secure edge sensing and computing systems, and ML and AI hardware and software for emerging nanosatellite applications.

KEY AREAS: Bringing ML/AI to the edge; intermittent computing; embedded, on-platform AI through AI hardware



Dr. Gianluca (Gian) Piazza

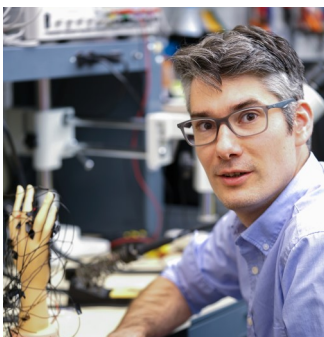
PROFESSOR, ELECTRICAL & COMPUTER ENGINEERING

DIRECTOR, NANOFAB

piazza@ece.cmu.edu

Dr. Gian Piazza's research focuses on the development of low power and miniaturized RF, microwave and ultrasonic microsystems for large sensor network data acquisition and transmission. He also works on developing innovative nanomechanical systems to facilitate low power hardware reconfiguration for edge computing.

KEY AREAS: NEMS relay for low power intermittent/edge computing



Dr. Anthony Rowe

PROFESSOR, ELECTRICAL & COMPUTER ENGINEERING

agr@ece.cmu.edu

Dr. Anthony Rowe's research is in the area of networked embedded sensing systems. Most recently his work has focused on GPS-denied positioning, timing and navigation along with edge computing infrastructure support for mixed reality systems.

KEY AREAS: Networked, real-time embedded systems; large-scale sensor networks; adaptive and distributed communication



Dr. Aswin Sankaranarayanan

ASSOCIATE PROFESSOR, ELECTRICAL & COMPUTER ENGINEERING

saswin@ece.cmu.edu

Dr. Aswin Sankaranarayanan's research takes a holistic view of the design of imaging systems wherein the design of sensors and learning algorithms complement each other. He views sensors as optical computational elements - an unexploited degree of freedom that not just reduces the computational burden of processing but can also make hard problems more tractable.

KEY AREAS: Computational sensing-or the joint design of sensors & processing techniques to optimize Inferencing



Dr. Osman Yağın

ASSOC. RESEARCH PROFESSOR, ELECTRICAL & COMPUTER ENGINEERING

oyagan@ece.cmu.edu

Dr. Osman Yağın's research around ML/AI focuses on developing online learning algorithms to optimally infer the stochastic properties of hidden random phenomena (e.g., distribution estimation, hypothesis testing, etc.); a class of reinforcement learning problems known as multi-arm bandits, with a focus on where arms are correlated; and developing algorithms to perform privacy-preserving data analytics.

KEY AREAS: Statistical inference & decision-making with sequential samples (in tactical networks); secure, reliable, & resilient design of large-scale distributed networks



Dr. Pei Zhang

ASSOC. RESEARCH PROFESSOR, ELECTRICAL & COMPUTER ENGINEERING

peizhang@cmu.edu

Dr. Pei Zhang's research focuses on learning in real-world data-limited cyber-physical systems by integrating data models, with physical knowledge, and dynamically actuating the hardware system.

KEY AREAS: ML in cyber-physical systems; structures as sensors; mobile carriers as sensors

AI FUSION CORE RESEARCHERS: AI ENGINEERING & SECURITY



Dr. Matt Gaston

DIRECTOR, EMERGING TECH. CENTER, SOFTWARE ENGINEERING INSTITUTE

megaston@sei.cmu.edu

Dr. Matt Gaston's areas of expertise include autonomy and autonomous systems, computer science and software engineering, and large-scale computational modeling and simulation. He is currently leading a national initiative to establish and advance the discipline of AI Engineering for Defense and National Security for scalable, robust and secure, and human-centered AI systems.

KEY AREAS: AI verification & validation, AI system architecture & integration, AI infrastructure orchestration



Dr. Eric Heim

SR. RESEARCH SCIENTIST, MACHINE LEARNING

etheim@andrew.cmu.edu

Dr. Eric Heim's research interests revolve around efficiently and effectively learning models of similarity and representations from different forms of human supervision. He currently leads work in the areas of inverse reinforcement learning and engineering AI systems to deal with uncertainty.

KEY AREAS: AI verification & validation, system architecture & integration, data/model management & coherence, AI infrastructure orchestration



Dr. Shing-hon Lau

SR. CYBERSECURITY ENGINEER, SOFTWARE ENGINEERING INSTITUTE

slau@sei.cmu.edu

Dr. Shing-hon Lau currently co-leads the AIDE (AI Defense Evaluation) project which is developing a methodology for evaluating AI & ML-powered network defenses. His research interests include understanding how to build and how to rigorously test trustworthy AI systems.

KEY AREAS: Data/model management & coherence, AI verification & validation



Dr. Grace Lewis

PRINCIPAL RESEARCHER, TACTICAL & AI-ENABLED SYSTEMS LEAD, SOFTWARE ENGINEERING INSTITUTE

glewis@sei.cmu.edu

Dr. Grace Lewis's areas of expertise include software engineering for AI & ML systems, IoT security, edge computing, software architecture (in particular the development of software architecture practices for systems that integrate emerging technologies), and software engineering in society. She currently leads work in characterizing and detecting mismatch in ML-enabled systems and predicting inference degradation in operational ML systems.

KEY AREAS: AI system architecture & integration



Dr. Ipek Ozkaya

TECHNICAL DIRECTOR, ENGINEERING INTELLIGENT SOFTWARE SYSTEMS, SOFTWARE ENGINEERING INSTITUTE

ozkaya@sei.cmu.edu

Dr. Ipek Ozkaya's research interests lie in the intersection of applying and developing software architecture analysis and design techniques for AI and software systems and system sustainability. Her work includes developing techniques for improving software development efficiency and system evolution, with an emphasis on software architecture practices, software economics, agile development, and

managing technical debt in complex, large-scale software-intensive systems.

KEY AREAS: AI system architecture & integration



Carol Smith

SR. RESEARCH SCIENTIST—HUMAN MACHINE INTERACTION, SOFTWARE ENGINEERING INSTITUTE

cjsmith@sei.cmu.edu

Carol Smith's expertise is in improving AI systems and the human experience across industries. Her current work is in guiding the creation of AI systems that are accountable, de-risked, respectful, secure, honest and usable.

KEY AREAS: AI policy mechanisms (ethical principles)



Dr. John Wohlbiel

SR. RESEARCH SCIENTIST, SOFTWARE ENGINEERING INSTITUTE

jgwohlbiel@sei.cmu.edu

Dr. John Wohlbiel's work is focused on the performance of AI workloads on modern and emerging hardware. He is currently involved with the DARPA SDH and DSSoC programs, and the Spiral AI ML project at CMU SEI.

KEY AREAS: AI system architecture & integration, AI infrastructure orchestration



Hasan Yasar

TECHNICAL DIRECTOR, CONTINUOUS DEPLOYMENT OF CAPABILITY, SOFTWARE ENGINEERING INSTITUTE

hyasar@cmu.edu

Hasan Yasar leads an engineering group that enables, accelerates and assures transformation at the speed of relevance by leveraging DevSecOps, Agile, Lean AI and ML and other emerging technologies to create a Smart Software Platform/Pipeline. He has experience as a senior security engineer, software engineer, software architect and manager in all phases of secure software development and information modeling processes.

KEY AREAS: AI deployment & monitoring (AI DevOps)
