Theme Article: IEEE Micro Special Issue on JUMP2.0

CoCoSys: The Co-Design of Cognitive AI from Algorithms to Systems

Zishen Wan, *Georgia Institute of Technology and Columbia University, USA*
Yu Cao, *University of Minnesota, Minneapolis, MN, 55455, USA*
Sumeet K. Gupta, *Purdue University, West Lafayette, IN, 47907, USA*
Larry Heck, *Georgia Institute of Technology, Atlanta, GA, 30332, USA*
Tushar Krishna, *Georgia Institute of Technology, Atlanta, GA, 30332, USA*
Yingyan (Celine) Lin, *Georgia Institute of Technology, Atlanta, GA, 30332, USA*
Azad Naeemi, *Georgia Institute of Technology, Atlanta, GA, 30332, USA*
Bruno Olshausen, *University of California, Berkeley, CA, 94720, USA*
Priyadarshini Panda, *University of Southern California, Los Angeles, CA, 90089, USA*
Jan M. Rabaey, *University of California, Berkeley, CA, 94720, USA*
Vijay Raghunathan, *Purdue University, West Lafayette, IN, 47907, USA*
Priyanka Raina, *Stanford University, Stanford, CA, 94305, USA*
Tajana S. Rosing, *University of California, San Diego, La Jolla, CA, 92093, USA*
Kaushik Roy, *Purdue University, West Lafayette, IN, 47907, USA*
Jae-Sun Seo, *Cornell Tech, New York, NY, 10044, USA*
Naresh Shanbhag, *University of Illinois Urbana-Champaign, Urbana, IL, 61801, USA*
Josh Tenenbaum, *Massachusetts Institute of Technology, Cambridge, MA, 02139, USA*
Anand Raghunathan, *Purdue University, West Lafayette, IN, 47907, USA*
Arijit Raychowdhury, *Georgia Institute of Technology, Atlanta, GA, 30332, USA*

Abstract—The deep-learning scaling that drove the past decade of AI is colliding with hard limits in energy, data, robustness, and explainability, just as emerging applications demand the capabilities neural networks lack: reasoning, abstraction, and collaboration. The CoCoSys JUMP 2.0 center confronts this by co-designing cognitive systems vertically, from algorithms through architectures and silicon. This article highlights its research along four thrusts: (1) unified neural, symbolic, and probabilistic algorithms; (2) algorithm-hardware co-design steered by systematic workload characterization; (3) technology-driven hardware motifs built on memory-centric, non-volatile-memory computing; and (4) collective and collaborative intelligence for embodied, federated, and human-AI systems. Using neuro-symbolic AI as an end-to-end case study, from algorithm restructuring through reconfigurable dataflows to a heterogeneous silicon prototype, we show that co-designing the full software-to-silicon stack yields orders-of-magnitude efficiency gains while preserving the accuracy and interpretability that make cognitive AI compelling. We close with cross-layer lessons and open challenges.

0272-1732 © 2026 IEEE
Digital Object Identifier 10.1109/MM.2026.3734155

The Co-Design Imperative for Cognitive AI

For more than a decade, progress in artificial intelligence has been propelled by a virtuous cycle: larger neural networks, larger datasets, and faster matrix-multiplication hardware. This formula has delivered superhuman perception and fluent language generation, but it is now straining against fundamental limits. Training compute has been doubling every few months, frontier models have crossed into the trillions of parameters, and the resulting energy demand is rapidly driving up global datacenter energy consumption. Equally concerning, the brittleness and opacity of large neural models appear endemic rather than incidental: they interpolate patterns but struggle to generalize compositionally, enforce logical constraints, reason under uncertainty, or explain their conclusions. As AI is deployed in increasingly consequential settings, from autonomous agents to scientific discovery and human-robot collaboration, the link between these limitations and the prevailing scale-everything paradigm looks increasingly less incidental and more causal.

Cognitive AI based on a principled combination of neural perception with symbolic reasoning and probabilistic inference offers an alternative to the conventional trajectory for AI. By unifying neural, symbolic and probabilistic approaches, cognitive systems can represent structured knowledge, manipulate it under explicit rules, and quantify uncertainty, yielding intelligence that is interpretable, robust, and data-efficient. Neuro-symbolic models have already shown promise, reaching state-of-the-art results in abstract visual reasoning and olympiad-level mathematics using a fraction of the data and parameters of neural networks [1]. Compositional systems that pair language models with symbolic solvers match or exceed far larger LLMs on complex reasoning tasks, and embodied agents augmented with symbolic planners cooperate more reliably than purely neural counterparts. Together, these results point toward a path that scales intelligence rather than merely scaling size.

Realizing cognitive AI in practice, however, exposes a deep mismatch between its computational patterns and the hardware available to run it (Figure 1). Modern accelerators, from GPUs to TPUs and NPUs, and even in-memory processors, have converged on one primitive: dense, low-precision matrix multiplication (GEMM). Symbolic reasoning and probabilistic inference are different – they traverse sparse graphs, chase pointers, search combinatorial spaces, and bind high-dimensional vectors, producing irregular memory access, divergent control flow, and high memory intensity that map poorly onto current hardware. As algorithms evolve from CNNs and Transformers toward neuro-symbolic-probabilistic models, this functional gap widens, leaving compute underutilized, memory traffic high, and real-time deployment out of reach.

Closing this gap cannot be delegated to any single layer of the stack. Algorithm restructuring can expose hardware-friendly parallelism and shrink memory footprints; reconfigurable architectures can serve heterogeneous cognitive primitives within one substrate; and heterogeneous silicon that fuses computation with memory can deliver an efficiency general-purpose processors cannot. What this demands is vertical co-design: the synergistic evolution of algorithms, software, architectures, and devices toward a shared goal. This is the founding premise of CoCoSys, a JUMP 2.0 center jointly supported by SRC and DARPA, whose mission is to enable next-generation AI systems through co-design across the software-to-silicon stack. The remainder of this article highlights the center's research across its four thrusts, spanning deliverables such as low-power NPU design, model-to-NPU mapping, and compute-in-memory hardware alongside longer-horizon directions, and then traces neuro-symbolic AI as an end-to-end case study from algorithm to silicon, before closing with broader impact, lessons learned, and open challenges.

A Vertically Integrated Vision for Cognitive Systems

The vision of the CoCoSys center is to enable the next generation of collaborative human-AI systems through synergistic advances in algorithms, hardware motifs, algorithm-hardware co-design, and collective intelligence. AI systems consist of a set of interacting layers. At the top, *applications*, such as abstract reasoning, embodied planning, scientific discovery, and human-AI collaboration, define the cognitive tasks that drive requirements downward. The *models* layer composes these systems from neural networks for perception, symbolic engines for structured reasoning, and probabilistic models for uncertainty-aware inference. The *workloads* layer characterizes the runtime behavior of these models, their operator mix, memory footprints, and data dependencies. The *system software* layer bridges algorithms to hardware through compilers, schedulers, and dataflow mappings tailored to cognitive primitives. The *architecture* layer defines the heterogeneous processing system that executes the workload, and the *technology and silicon* layer realizes

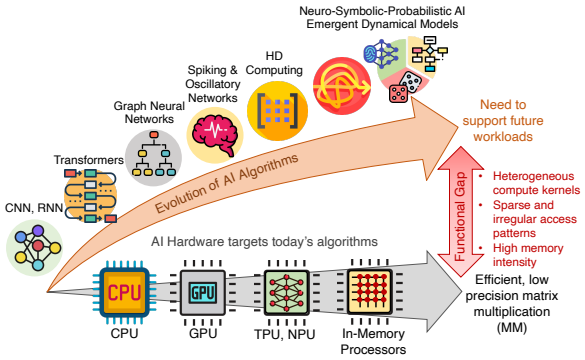


FIGURE 1. The functional gap between AI algorithm evolution and AI hardware evolution. While algorithms have progressed from CNNs and RNNs through Transformers and Graph Neural Networks to neuro-symbolic-probabilistic and hyperdimensional models, hardware has converged on efficient low-precision matrix multiplication (CPU → GPU → TPU/NPU → In-Memory Processors). The resulting gap, characterized by heterogeneous compute kernels, sparse and irregular access patterns, and high memory intensity, is the co-design challenge that CoCoSys addresses.

them using scaled CMOS technology, emerging devices such as resistive memory (RRAM) and advanced integration such as 3D stacking and heterogeneous packaging.

CoCoSys research is organized around four co-designed thrusts that span this stack, each with its own set of research tasks (Figure 3). *Thrust I (Neural, Symbolic, and Probabilistic Algorithms)* creates the next generation of explainable and data-efficient algorithms, unifying neural, symbolic, and probabilistic models, advancing hyperdimensional information representations and neuro-inspired optimization, and uncovering the fundamental security, efficiency, robustness, and accuracy tradeoffs that bound cognitive systems. *Thrust II (Algorithm-Hardware Co-Design)* distills the computational characteristics of these workloads into programmable cognitive architectures through joint algorithm-hardware search, full-stack optimization, and rigorous benchmarking. *Thrust III (Technology-Driven Hardware Motifs)* builds the hardware primitives of cognitive computing by matching CMOS and beyond-CMOS devices to workload needs, spanning fused logic and memory, mixed-mode circuits, technology evaluation, and heterogeneous integration. *Thrust IV (Collective and Collaborative Intelligence)* addresses systems of many agents and their interaction with humans through human-AI collaboration, federated and decentralized learning, robustness to noisy or untrusted agents, and privacy-preserving intelligence.

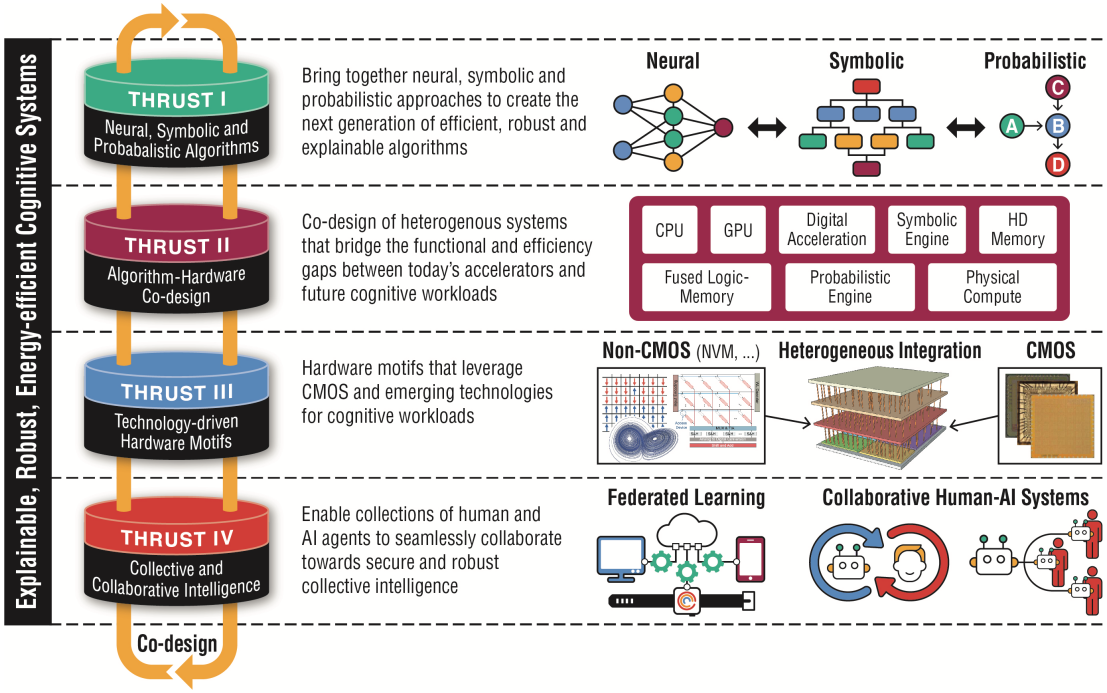
The portfolio deliberately spans two horizons. Near-term deliverables – low-power NPU designs and model-mapping frameworks, CIM and ACIM macros with their design tooling, efficient LLM inference, and communication-efficient federated learning – are transferable to practitioners today, many developed jointly with the center's industry partners (Figure 2, bottom). Long-horizon research, most prominently unified neuro-symbolic-probabilistic systems, targets capabilities beyond incremental scaling.

These thrusts share a common set of grand-challenge goals (Figure 4), pursued across this entire portfolio rather than through any single workload class: to improve the accuracy-robustness-efficiency profile of cognitive systems, to deliver more than an order-of-magnitude reduction in data and compute at equal accuracy, and to drive end-to-end latency to the sub-millisecond range with more than 100× better energy efficiency than the state of the art, all demonstrated in collaborative human-AI applications such as co-operative robotics and untethered mixed reality. The next few sections examine each thrust in turn, after which we use neuro-symbolic AI as an end-to-end case study that threads a single workload from algorithm to silicon.

Thrust I: Neural, Symbolic, and Probabilistic Algorithms

Thrust I builds the algorithmic foundation of CoCoSys. Its goal is a new generation of cognitive algorithms that are explainable, robust, and data-efficient, while also being amenable to efficient hardware implementation. The thrust pursues this along four directions: unifying the three (neural, symbolic and probabilistic) paradigms of cognition, advancing new information representations, developing neuro-inspired solvers for optimization, and characterizing the fundamental limits that bound robustness-accuracy-efficiency tradeoffs in cognitive systems.

Unifying neural, symbolic, and probabilistic models. These three paradigms in cognitive systems are complementary. Neural networks excel at perception and flexible function approximation; symbolic methods bring interpretability and inject structured prior knowledge, reducing data requirements; and probabilistic models reason under uncertainty, providing robustness in unstructured environments. CoCoSys develops unified frameworks that opportunistically combine these methods rather than viewing them as competing alternatives in isolation. A representative effort couples agentic language models with probabilis-



Representative outcomes across the CoCoSys portfolio

Thrust	Representative outcomes (contributing institutions)
I: Algorithms	Residue HD computing for factoring high-dimensional representations (UC Berkeley); LifeHD on-device lifelong HD learning, 34.3× energy efficiency (UCSD); statistical error compensation for in-memory computing, +2.7–6 dB compute SNR (UIUC); agentic LLM + probabilistic physical reasoning (MIT, GT); foveated object-centric robustness, gradient-free unlearning (Purdue)
II: Co-Design	NSAI and compositional-AI workload characterization (GT); heterogeneous NPU design-space exploration, 46.9% iso-area energy saving (Purdue + Intel); GNN / SSM-Mamba / LLM mapping onto commercial NPUs, up to 4.8× (Purdue + Intel); LaCache long-context LLM inference (GT); SCALE-Sim v3 cycle-accurate simulation (GT)
III: Hardware Motifs	Output-stationary SRAM CIM, 20.9–137.2 TOPS/W (Cornell Tech); ADC-less hybrid analog-digital CIM, up to 28% energy reduction (Purdue); ClipFormer write-noise mitigation on memristive crossbars, +10–40% non-ideal accuracy (USC); TAXI SOT-MRAM Ising machine, 8× faster at 86k cities (Purdue); Voyager accelerator generation and DSE (Stanford); digital CIM for N:M-sparse LLMs (GT + Intel); 40-nm RRAM-SRAM NSAI SoC, 10.8 TOPS/W at 0.73 V (GT + TSMC)
IV: Collective Intelligence	ReCA cooperative embodied agents, 10.2× end-to-end speedup (GT, Harvard); MulBERRY bit-error-robust UAV swarms, 4.16× processing energy reduction (GT, IBM); MultimodalHD federated HD learning across heterogeneous sensor modalities, 2–8× faster training (UCSD); subspace federated LLM transfer, conversational grounding (GT)

FIGURE 2. *Top*: the four CoCoSys thrusts and their bidirectional co-design coupling, through which algorithm, architecture, and technology decisions co-evolve toward explainable, robust, and energy-efficient cognitive systems. *Bottom*: representative outcomes across the center's portfolio, with key quantitative results and contributing institutions, spanning near-term deliverables (NPU design and model mapping, CIM/ACIM macros and tooling, efficient LLM inference, federated learning) and long-horizon research (unified neuro-symbolic-probabilistic systems).

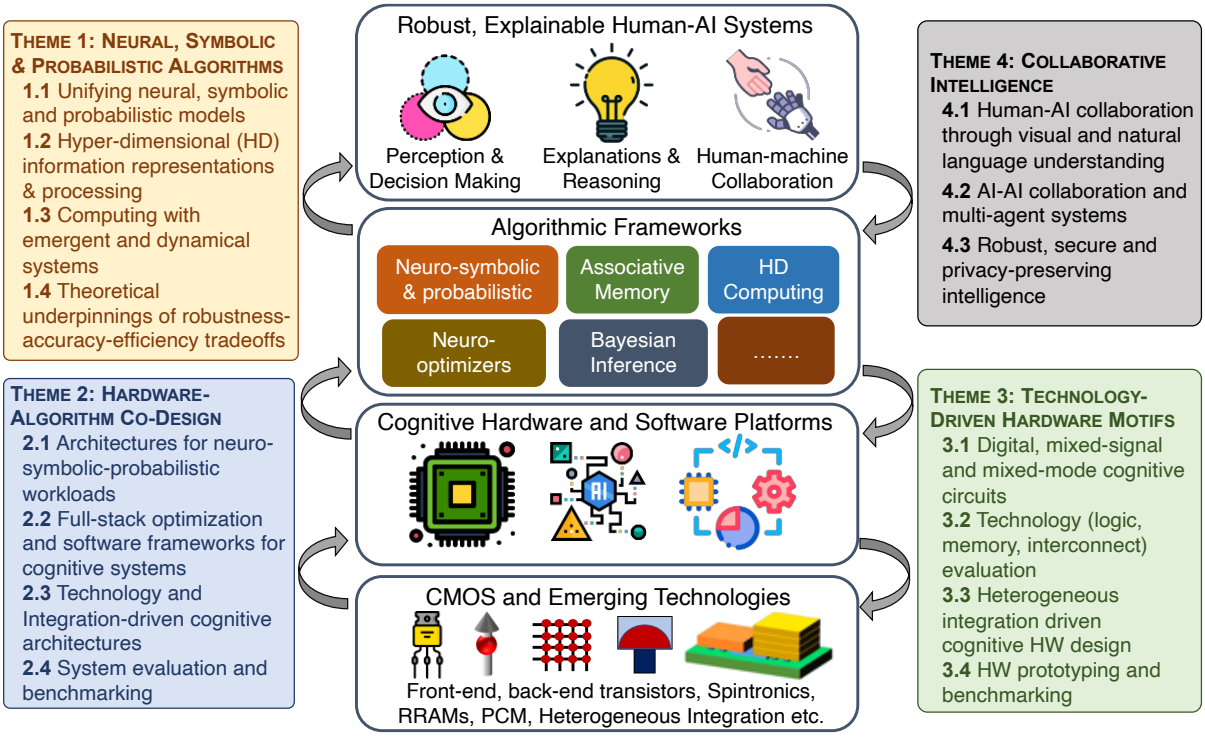


FIGURE 3. The four CoCoSys research thrusts and their constituent tasks, atop a shared cognitive hardware-software platform spanning CMOS and emerging technologies. *Thrust I* develops neural, symbolic, and probabilistic algorithms, hyperdimensional representations, neuro-inspired optimization, and the theory of accuracy-robustness-efficiency tradeoffs. *Thrust II* co-designs algorithms and hardware through joint architecture search, full-stack optimization, and benchmarking. *Thrust III* builds technology-driven hardware motifs spanning fused logic-memory, mixed-mode circuits, and heterogeneous integration. *Thrust IV* enables collective and collaborative intelligence through human-AI collaboration, federated learning, and privacy-preserving methods.

tic programs to answer open-ended physical-reasoning questions, synthesizing question-conditioned models whose predictions align with human judgments and outperform vision-language models that reason directly over each task. This compositional structure is explored further in the case study section.

Hyperdimensional information representation. Hyperdimensional (HD) computing, also called vector-symbolic architectures (VSA), represents concepts, sensory data, and high-level state as high-dimensional vectors in a common space, and reasons over them through algebraic binding, bundling, and permutation. A core challenge is the inverse problem: recovering the constituent factors of a composite hypervector, such as the objects and positions in a visual scene. Residue hyperdimensional computing marries residue number systems with resonator networks, factoring composite hypervectors and solving visual-perception and combinatorial-optimization tasks with far fewer computational resources than generic methods [2]. Complementary work makes HD learning practical at the extreme edge: LifeHD performs unsupervised lifelong learning fully on device, improving clustering accuracy by up to 74.8% at 34.3× better energy efficiency than neural lifelong-learning baselines [3]. Because HD representations are inherently error-resilient and memory-centric, they map naturally onto the low-power, in-memory hardware motifs of Thrust III.

Neuro-inspired optimization. Many cognitive and scientific tasks reduce to hard combinatorial optimization, where neural models can guide classical solvers toward better solutions at scale. Neuro-Ising illustrates this synergy: a graph neural network globally guides a collection of localized Ising solvers, partitioning large traveling-salesman instances into locally solved clusters and sustaining near-optimal quality where monolithic Ising annealers degrade. This insight, that physics-based solvers should be localized and neurally orchestrated, directly motivates the in-memory Ising hardware of Thrust III.

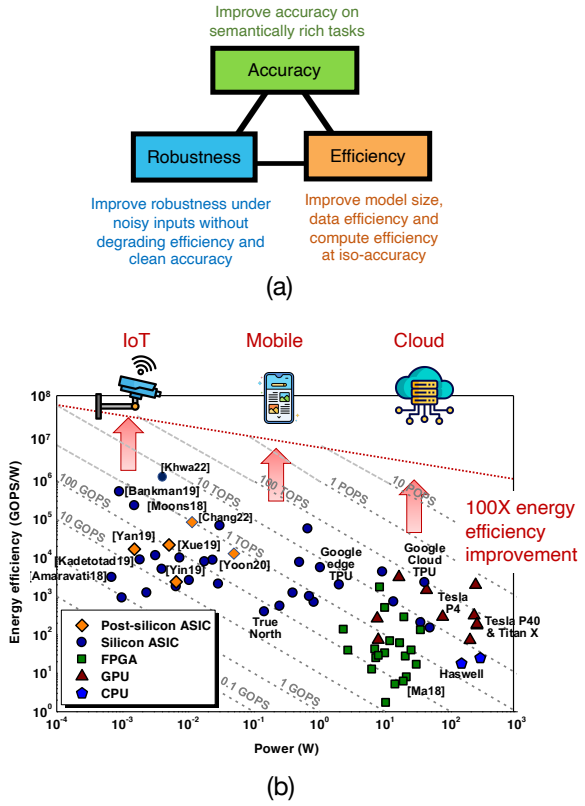


FIGURE 4. CoCoSys grand-challenge goals. **(a)** Improve the joint accuracy-robustness-efficiency profile of cognitive systems, raising accuracy on semantically rich tasks and robustness to noise without sacrificing efficiency or clean accuracy. **(b)** Improve the energy efficiency of cognitive hardware by more than 100× over the state of the art, from cloud to mobile to IoT, through full-stack co-design.

Robustness, security, and fundamental limits. A distinctive aim of Thrust I is to characterize, rigorously, the fundamental tradeoffs among security, efficiency, robustness, and accuracy that bound cognitive systems, using tools from learning theory, signal processing, and information theory. Statistical error compensation carries this thesis into silicon: a 22-nm MRAM-based in-memory computing macro treats analog nonidealities statistically rather than masking them, boosting compute SNR by 2.7-6 dB and recovering ResNet-20 accuracy from 74.8% to 82.0% at only 0.5% overhead, without on-chip training [4]. On the robustness and safety front, center work spans foveated object-centric learning, which biases recognition toward label-consistent object regions for robustness to clutter and distribution shift, and gradient-free deep unlearning, which removes specific concepts or noisy samples from a trained model without full retraining. Together these results turn robustness and security from afterthoughts into first-class design objectives.

Thrust II: Algorithm-Hardware Co-Design

Thrust II is the bridge between the algorithms of Thrust I and the device technologies of Thrust III (Figure 3). Its premise is that the design of highly efficient cognitive systems requires tight coupling between algorithm and hardware: workload characteristics must steer architecture, and hardware constraints must feed back into algorithm design. The thrust advances this co-design loop through systematic characterization and benchmarking, full-stack optimization, and modeling frameworks that let architects explore the cognitive design space before committing to silicon.

Characterization precedes optimization. A central methodological commitment of CoCoSys is that hardware teams must measure before they build. The center profiled two generations of cognitive workloads to localize their true bottlenecks: seven classical neuro-symbolic models across server GPUs, edge SoCs, and a neural accelerator [5], and six compositional LLM-integrated workloads on H100/H200 systems [6]. The campaigns surfaced four recurring bottlenecks, profiled in detail for neuro-symbolic AI in Figure 5: symbolic and probabilistic operations dominate end-to-end latency; neural and symbolic kernels sit at opposite ends of the roofline, the former compute-bound and the latter memory-bound; general-purpose accelerators are severely underutilized on non-GEMM primitives; and in compositional pipelines, parameter-light symbolic components create outsized Amdahl bottlenecks. This benchmarking ethos extends beyond efficiency to reasoning quality: dedicated benchmarks now measure the token efficiency of LLM reasoning jointly with accuracy, exposing that models of equal accuracy can differ several-fold in their compute cost per unit of intelligence.

Low-power NPU design and model-to-NPU mapping. A major near-term deliverable of the thrust, pursued with the center's industry partners, is making commercial low-power neural processing units (NPUs) serve the full breadth of modern model classes. MOSAIC jointly explores heterogeneous NPU architectures that mix "big," "little," and specialized function tiles for state-space models (SSMs), FFT-style long convolutions, Kolmogorov-Arnold networks, and spiking networks, delivering 46.9% mean iso-area energy savings over homogeneous designs across a 20-workload suite and 1.5-2.4× speedup over NVDLA [7]. Complemen-

tary mapping frameworks carry today's mainstream models onto off-the-shelf edge NPUs: GraNNite enables graph neural networks despite their irregular gather-scatter patterns [8]; XAMBA restructures the recurrent scan of SSM/Mamba into NPU-friendly matrix primitives for up to 4.8× speedup on a commercial AI PC [9]; and LLM-NPU targets the memory-bound decode phase of foundation models on low-power NPUs [10]. This industry co-design extends to structured sparsity: VEGETA adds flexible N:M sparse-tile support to CPU matrix engines, co-designed with Intel [11], and the same N:M discipline carries into the digital CIM accelerator for LLM inference in Thrust III. These results close the loop opened in Figure 1: the emerging model classes on the algorithm-evolution trajectory now have concrete, transferable mapping strategies on deployed hardware.

Full-stack optimization. Once bottlenecks are localized, co-design attacks them across the stack. For long-context LLMs, ladder-shaped KV-cache compression restructures the cache to retain both short- and long-range information at a fixed memory budget, extending context length without retraining [12]. For compositional neuro-symbolic pipelines, system-level optimizations such as speculative symbolic execution, cross-component operator fusion, and coordinated cache management cut end-to-end latency by 50-70% with no change to model accuracy [6]. These algorithmic levers are architecture-agnostic: they precede hardware specialization and multiply its eventual gains.

Modeling, simulation, and design frameworks. Co-design at scale depends on fast, accurate models that predict system behavior before fabrication. CoCoSys develops full-stack modeling infrastructure for cognitive systems: cycle-accurate simulators capture the end-to-end behavior of systolic accelerators across diverse dataflows, mappings, and memory hierarchies, enabling rapid design-space exploration before a chip is committed [13]. Complementary frameworks automate chiplet synthesis and heterogeneous integration and model multi-stage distributed inference, giving architects a principled way to navigate tradeoffs among latency, energy, and area for cognitive workloads.

From characterization to custom architectures. The distilled workload signatures, heterogeneous kernels, sparse and irregular access, and high memory intensity, point to a clear conclusion: cognitive AI needs programmable architectures that serve neural, symbolic, and probabilistic primitives within a single substrate, rather than bolting accelerators for each onto a GEMM-centric datapath. CoCoSys has produced the first such architectures, and the case study section develops this story in depth for neuro-symbolic AI, carrying a single workload class from reconfigurable dataflow through to fabricated silicon.

Thrust III: Technology-Driven Hardware Motifs

Thrust III builds the physical building blocks of cognitive hardware (Figure 3) by matching the capabilities of CMOS and beyond-CMOS devices to the needs of cognitive workloads. Because Thrust II identifies memory capacity, bandwidth, and access energy as the binding constraints, the thrust concentrates on memory-centric and physical computing motifs: compute-in-memory, mixed-mode circuits that compute directly in device physics, and emerging non-volatile memories combined through heterogeneous integration.

Compute-in-memory, from macro to system. Compute-in-memory (CIM) embeds multiply-accumulate operations inside the memory array, attacking the data-movement bottleneck at its source. CoCoSys advances output-stationary CIM, in which an outer-product dataflow eliminates the costly partial-sum memory traffic of conventional weight-stationary macros and natively supports structured and unstructured sparsity, reaching 20.9-137.2 TOPS/W in a 28 nm prototype [14]. A key lesson, however, is that macro-level efficiency does not translate directly to systems: once analog-to-digital conversion, data movement, and array underutilization are accounted for, a macro reporting hundreds of TOPS/W can deliver only single-digit TOPS/W in a full workload. CoCoSys addresses this gap with end-to-end accelerator-generation and evaluation frameworks that compile complete workloads onto generated datapaths and report system-level latency, energy, and utilization, enabling fair macro-to-system comparison and rigorous design-space exploration of CIM-based architectures [15]. The same discipline extends to LLM-scale workloads: a fully digital CIM accelerator co-designed with flexible N:M sparsity, developed with the center's industry partners, carries CIM efficiency into transformer inference [16].

ADC-less analog CIM and reliability. Two circuit-level advances make *analog* CIM (ACIM) practical rather than merely dense. First, because analog-to-digital converters dominate ACIM area and energy, HCiM eliminates them: an algorithm-circuit co-designed hybrid analog-digital accelerator replaces ADCs with lightweight analog shift-and-add and ternary-quantization-aware processing, cutting macro energy by up to 28% versus a 7-bit-ADC baseline

at iso-accuracy [17]. Second, deployability hinges on robustness to device variation, conductance drift, and array non-idealities; ClipFormer mitigates memristive write noise for transformers at inference time, boosting non-ideal accuracy by 10-40% in high-noise regimes with no added hardware or retraining [18]. Together with the statistical error compensation of Thrust I [4], these results treat analog non-determinism as a design parameter, exposed to and absorbed by the algorithm, rather than a defect to be masked.

Mixed-mode and physical computing for optimization. Beyond digital datapaths, Thrust III encodes computation directly in the physics of nanoscale devices, where stochasticity becomes a resource rather than a nuisance. TAXI realizes an in-memory Ising machine: crossbar macros in which spin-orbit-torque (SOT) MRAM devices serve as compact random number generators driving natural annealing, each independently solving a hierarchically clustered subproblem with no off-macro data movement [19]. TAXI scales to traveling-salesman instances of nearly 86,000 cities, on average 8× faster than state-of-the-art clustering-based Ising solvers – the hardware embodiment of the Neuro-Ising insight from Thrust I.

Emerging non-volatile memory and heterogeneous integration. For models whose working sets exceed on-chip SRAM, dense non-volatile memory eliminates the DRAM traffic that would otherwise dominate power: CoCoSys's heterogeneous RRAM-SRAM system-on-chip exploits RRAM's density to hold full neuro-symbolic working sets on chip while reserving low-latency SRAM for symbolic datapaths [20], as the case study details. Heterogeneous integration through 3D stacking and chiplets extends the same principle beyond the reticle limit, pursued in collaboration with sister JUMP 2.0 centers.

Thrust IV: Collective and Collaborative Intelligence

Thrust IV raises the unit of analysis from a single model to collections of agents that cooperate with one another and with humans (Figure 3). This is where cognitive AI ultimately earns its keep: the center's target applications, cooperative robots, digital assistants, and mixed-reality systems, are inherently multi-agent and human-in-the-loop, inheriting every efficiency challenge of single-system cognition while adding new ones in communication, coordination, and real-time embodied execution. The thrust pursues algorithms for human-AI collaboration, federated and decentralized learning, robustness to noisy or untrusted agents, and natural-language interfaces, carrying vertical co-design from the chip to the collective.

Cooperative embodied agents. Recent embodied systems compose multiple LLM-powered agents to solve long-horizon, multi-objective tasks such as cooperative transport and human-agent assistance. They are algorithmically capable but far from deployable: a single task can take tens of minutes on desktop GPUs, limited by three bottlenecks, repeated sequential LLM calls for planning and communication, poor scaling as agents are added, and brittle low-level execution. ReCA applies CoCoSys's vertical methodology to this setting, co-optimizing across algorithm, system, and hardware [21]. At the algorithm level, efficient local-model processing replaces cloud calls with on-agent models; at the system level, a dual-memory organization, a hierarchical planner that blends centralized and decentralized coordination, and planning-guided multi-step execution amortize expensive planning calls; and at the hardware level, a heterogeneous platform pairs a GPU subsystem for high-level planning with a dedicated accelerator for low-level control. ReCA delivers a 10.2× end-to-end speedup while improving mission success by 4.3% over state-of-the-art cooperative systems. Complementary work on compositional AI planning, in which a language model translates natural-language goals into a formal task definition that a symbolic planner solves, brings interpretable, verifiable long-horizon planning to multi-agent systems.

Robustness in distributed, resource-constrained settings. Multi-agent autonomy must run within strict size, weight, and power budgets, where aggressive energy savings collide with reliability. MuIBERRY is a multi-agent robust learning framework for autonomous UAV swarms that makes low-voltage operation practical despite the on-chip bit failures it induces, combining robust offline and on-device learning, collaborative agent-server optimization, and dynamic thermal-payload management [22]. Across fault patterns, environments, and swarm sizes it reduces flight energy by up to 19%, increases successful missions by up to 22%, and cuts processing energy by 4.16×, linking collaborative intelligence to the low-voltage device efficiency of Thrust III.

Federated and decentralized learning. On-device learning across many agents is bottlenecked by communication. MultimodalHD federates hyperdimensional learning across agents with heterogeneous sensor modalities, matching state-of-the-art multimodal federated-learning accuracy while training 2-8× faster on edge-class hardware [23], a natural fit for the error-resilient HD representations of Thrust I. At LLM scale, subspace-based federated transfer learning lets agents specialized on different tasks or domains ex-

change knowledge through compact low-rank subspaces, enabling collaboration across heterogeneous models without sharing raw data.

Human-AI collaboration and natural language. Building AI that works *with* people requires interfaces that are intuitive, context-aware, and trustworthy. CoCoSys advances natural-language and conversational agents grounded in personal and multimodal context, including real-time question answering over continuous sensor streams, and pairs them with the interpretability of neuro-symbolic models from Thrust I so that human collaborators can audit an agent's reasoning rather than trust it blindly. Together, these efforts make collective intelligence not only efficient but accountable.

Case Study: Neuro-Symbolic AI from Algorithm to Silicon

No single workload exposes the cognitive-AI hardware gap more sharply than neuro-symbolic AI (NSAI), and none threads the four thrusts more tightly. We present NSAI as the center's end-to-end stress test of vertical co-design: unlike the near-term NPU, CIM, and LLM-efficiency deliverables above, NSAI is a deliberately long-horizon bet whose returns at deployment scale are not yet realized. NSAI systems integrate neural perception, symbolic reasoning over vector-symbolic, first-order-logic, and satisfiability representations, and probabilistic inference, composing them as pipelines and nested loops. Figure 5 lists representative NSAI workloads and the profiling that exposes the bottlenecks summarized in Thrust II. We now follow this workload class end to end to show how vertical co-design compounds across layers, turning gains that are modest in isolation into orders-of-magnitude improvements in concert.

Algorithm restructuring. Restructuring algorithms before designing hardware lightens the load on every downstream layer. CogSys factorizes the large symbolic codebooks of vector-symbolic architectures into a compact basis reconstructed on the fly, shrinking codebook memory 71.4× (13,560 KB to 190 KB) and cutting encoding latency 4.1×, so the working set fits entirely on chip [24]. REASON unifies first-order logic, satisfiability, probabilistic circuits, and Hidden Markov Models under a common directed-acyclic-graph representation, then applies adaptive pruning and two-input regularization to produce hardware-friendly binary trees, compressing HMMs 52.6× with under 1% accuracy loss [25]. System-level optimizations such as speculative symbolic execution and operator fusion cut compositional-pipeline latency by 50-70% at unchanged accuracy [6]. These factored, pruned representations directly shape the dataflows and memory hierarchies described next.

Reconfigurable architectures. Restructuring reduces the burden but cannot erase the heterogeneity of NSAI execution, so the architecture responds with reconfigurability (Figure 6, top). CogSys's reconfigurable neuro-symbolic processing element (nsPE) switches among data loading, dense matrix multiply, and circular convolution, and its Bubble Streaming dataflow pipelines the convolution with $O(d)$ rather than $O(d^2)$ on-chip memory, yielding 75.96× speedup over a TPU and up to 730× better energy efficiency than an RTX GPU in 4 mm^2 at 1.48 W [24]. REASON extends reconfigurability to probabilistic logical reasoning with a tree-structured PE operating in probabilistic-inference, logic-evaluation, and sparse-matrix modes, reaching 12-50× speedup over a CPU and 310-681× energy efficiency over an RTX GPU across ten benchmarks [25].

Silicon realization. Closing the loop, a heterogeneous system-on-chip in 40 nm CMOS with back-end-of-line RRAM integrates ten RRAM neural tiles and two SRAM symbolic tiles on a shared network-on-chip (Figure 6, bottom) [20]. The memory split is a co-design decision grounded in workload characterization: capacity-bound neural kernels exploit RRAM's 2.9× higher cell density, while symbolic kernels use low-latency SRAM datapaths, holding the full working set of all target NSAI workloads on chip. A custom edge-triggered prolonged sensing circuit reaches 4.80 Mb/mm^2 at 0.247 pJ/bit (1.67× denser than prior art at the node), and a software-defined power manager powers down idle tiles for a 6.47× average power reduction. The chip delivers 10.8 TOPS/W, operates down to 0.73 V, and serves the full NSAI paradigm space through a unified 128-bit VLIW interface without re-fabrication.

Holistic gains. The stack shows that the accuracy-efficiency tradeoff long assumed for cognitive AI is an artifact of inadequate hardware: the co-designed accelerators improve energy efficiency by two to three orders of magnitude while sustaining high cognitive-task accuracy, gains that compound multiplicatively across layers. The chip occupies a Pareto frontier, above 90% accuracy at sub-second latency, that neither neural-only nor symbolic-only hardware can reach.

Broader Impact: Physical Intelligence, Responsible AI, and Semiconductor Frontiers

The advances surveyed here reach well beyond the systems that instantiate them. They converge on three

Representative Neuro-Symbolic Workloads		AlphaGeometry	R^2-Guard	GeLaTo	Ctrl-G	NeuroPC	LINC
Deployment Scenario	Application	Math theorem proving & reasoning	Unsafety detection	Constrained text generation	Interactive text editing, text infilling	Classification	Logical reasoning, Deductive reasoning
	Advantage vs. LLM	Higher deductive reasoning, higher generalization	Higher LLM resilience, higher data efficiency, effective adaptability	Guaranteed constraint satisfaction, higher generalization	Guaranteed constraints satisfaction, higher generalization	Enhanced interpretability, theoretical guarantee	Higher precision, reduced overconfidence, higher scalability
Computation Pattern	Neural	LLM	LLM	LLM	LLM	DNN	LLM
	Symbolic	First-order logic, SAT solver, acyclic graph	First-order logic, probabilistic circuit, Hidden Markov model	First-order logic, SAT solver, Hidden Markov model	Hidden Markov model, probabilistic circuits	First-order logic, probabilistic circuit	First-order logic, solver

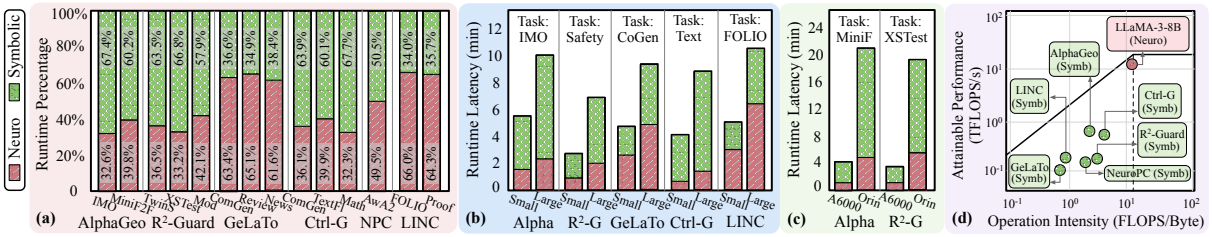


FIGURE 5. Representative neuro-symbolic AI workloads and their characterization. *Top*: six representative NSAI workloads spanning theorem proving, safety verification, constrained generation, and logical inference, with their deployment scenarios and neural-symbolic computation patterns. *Bottom*: end-to-end profiling of these workloads. **(a)** Symbolic and probabilistic stages dominate runtime on a CPU+GPU system. **(b)** Real-time targets are missed as task scale grows. **(c)** The trend holds across server (A6000) and edge (Orin) platforms. **(d)** Roofline analysis shows symbolic and probabilistic kernels are memory-bound while neural kernels are compute-bound, the architectural mismatch that motivates co-designed cognitive hardware.

areas where cognitive AI is poised to reshape computing.

Physical intelligence and edge autonomy. Cognitive AI is a prerequisite for robots and autonomous systems that must perceive, reason, and act in open-ended physical environments under tight power budgets. CoCoSys spans this need from device to swarm: foveated perception makes recognition efficient and robust at the sensor; neuro-inspired and in-memory Ising solvers bring planning-grade combinatorial optimization on chip [19]; the NSAI SoC makes on-device cognitive reasoning feasible at 0.73 V and 10.8 TOPS/W [20]; and ReCA and MuIBERRY carry the same co-design discipline up to cooperative, energy-constrained multi-agent autonomy [21], [22].

Responsible and trustworthy AI. Interpretability, robustness, and data efficiency, intrinsic to neuro-symbolic and probabilistic models, are increasingly prerequisites for AI in safety-critical domains such as healthcare, law, and autonomous transportation. CoCoSys treats trustworthiness as a design axis rather than an afterthought: Thrust I characterizes the fundamental security-efficiency-robustness-accuracy limits that bound what any cognitive system can guarantee [4], and supplies concrete mechanisms for safety, from gradient-free unlearning of harmful or private concepts to bit-error and adversarial robustness for deployed agents [22]. Communication-efficient federated learning keeps data on-device by design [23], and the symbolic structure of NSAI exposes reasoning chains that humans can audit rather than merely trust. Modular composition lets neural and symbolic components be validated separately and updated without full retraining [6], charting a concrete path toward certifiable cognitive systems.

Semiconductor and memory frontiers. The center's hardware work shows that emerging devices are not merely denser drop-in replacements but enablers of qualitatively new architectural organization: resistive memory lets full cognitive working sets sit on chip [20], output-stationary compute-in-memory rethinks the dataflow of the array itself [14], spin-orbit-torque devices turn intrinsic stochasticity into a computational resource [19], and error-resilient hyperdimensional representations invite low-power stochastic circuits [2], [3]. Co-designing algorithms around device properties such as density, write energy, and variability opens an axis of efficiency orthogonal to transistor scaling, pursued with sister JUMP 2.0 centers on memory, integration, and devices.

Lessons Learned and Open Challenges

Five years of CoCoSys research have produced not only systems and silicon but a set of transferable design principles, along with a clear map of where the hardest problems remain.

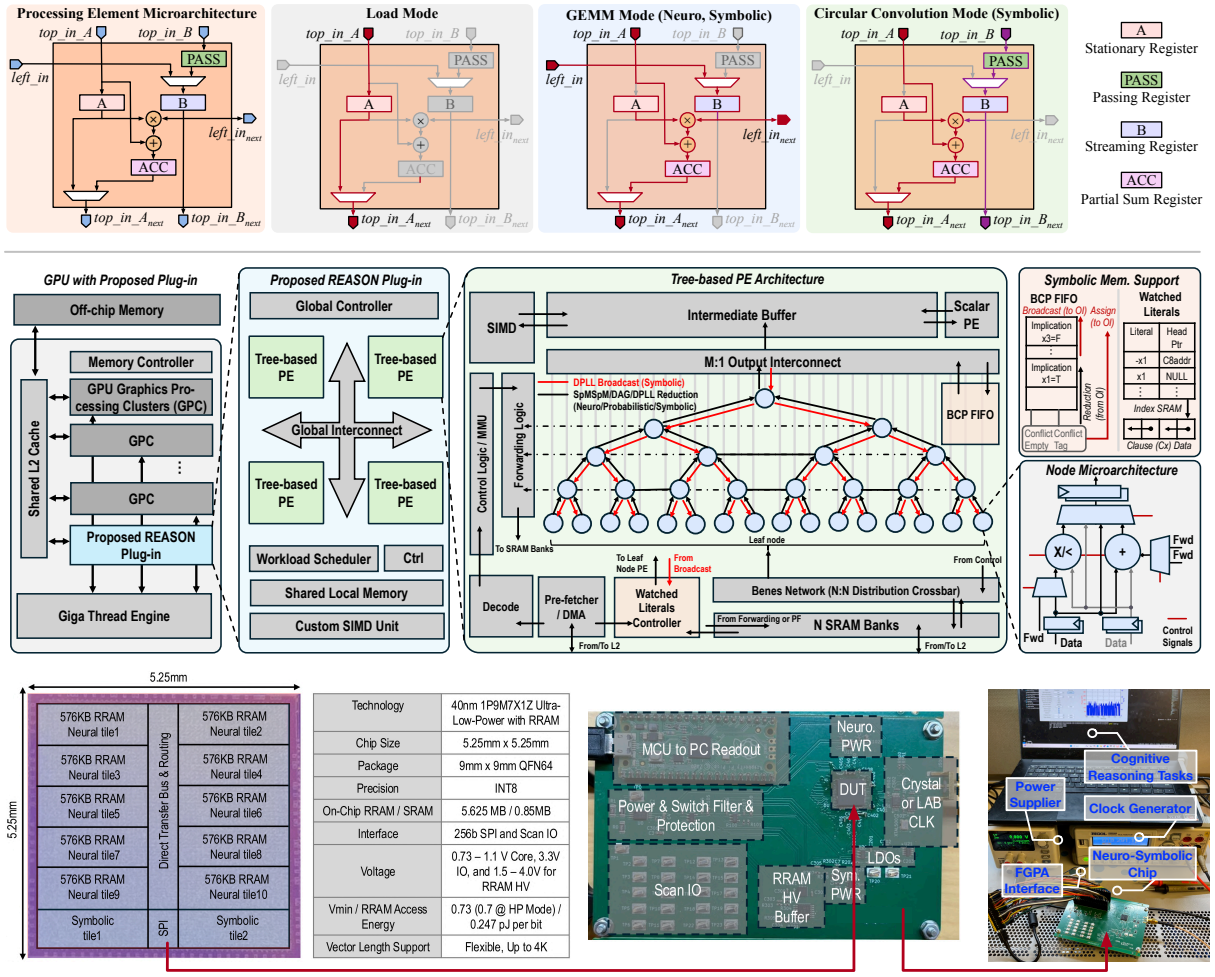


FIGURE 6. From reconfigurable architecture to silicon in the NSAI co-design stack. *Top*: CogSys's neuro-symbolic PE (nsPE) reconfigures among load, GEMM (neural), and circular-convolution (symbolic) modes with a Bubble Streaming dataflow [24]. *Middle*: REASON's tree-structured PEs unify probabilistic inference, symbolic logic, and sparse matrix multiply under Benes-network routing [25]. *Bottom*: the NSAI system-on-chip realizing the stack in silicon [20] – die photo of the $5.25 \times 5.25\,\text{mm}^2$ chip in 40 nm ULP CMOS with BEOL RRAM (ten 576-KB RRAM neural tiles, two SRAM symbolic tiles), key specifications (INT8, 5.625/0.85 MB RRAM/SRAM, 0.73–1.1 V core, 0.247 pJ/bit RRAM access), and the evaluation board running cognitive-reasoning benchmarks.

Lessons learned. *Co-design must be vertical, not horizontal.* Optimizing algorithms, architectures, and silicon independently and then composing them yields additive gains at best. The compounding efficiency observed across the center, where algorithmic restructuring enables microarchitectural specialization, which in turn shapes memory-hierarchy and device choices, arises only when each layer informs the others from the outset. This vertical integration is the central methodological contribution of CoCoSys, and it transfers across scales: the same algorithm-system-hardware co-design that accelerates a single cognitive model also accelerates cooperative embodied agents by an order of magnitude.

Characterization precedes optimization. The bottlenecks that motivated the center's hardware were not obvious a priori. Systematic profiling across platforms and workloads was the prerequisite for principled design. Without measurement, hardware teams optimize the wrong kernels.

Reconfigurability pays for itself. A persistent assumption is that reconfigurability costs area and power relative to fixed-function designs. The evidence from CogSys, REASON, and the heterogeneous NSAI SoC

consistently contradicts this: across long inference sequences with heterogeneous primitive mixes, reconfigurable datapaths amortize their switching overhead and rival workload-specific designs.

Memory and device physics are first-class. Across cognitive workloads, memory capacity, bandwidth, and access energy dominate the efficiency budget, which is why compute-in-memory, dense non-volatile memory, and heterogeneous integration recur throughout the center's hardware. A related lesson runs the other way: rather than masking device nondeterminism, the most efficient designs expose it to the algorithm, turning stochasticity and error resilience into computational resources, as in hyperdimensional computing and Ising annealing.

Open challenges. Several fundamental challenges remain open. *Hardware-efficient training*: current accelerators target inference, while cognitive-AI training, especially gradient passing through discrete symbolic operators, remains largely unsolved. *LLM-scale cognition*: integrating symbolic and probabilistic reasoning into transformer-scale models at hardware efficiency is nascent; the compositional-pipeline work localizes the system bottlenecks but does not yet close the hardware gap. *Collaborative intelligence at scale*: real-time, energy-constrained coordination among many agents and humans demands co-design methodologies that span devices, runtimes, and learning algorithms simultaneously. *Standardized benchmarks*: the field lacks agreed-upon cognitive-AI benchmarks that jointly evaluate accuracy, reasoning quality, robustness, and efficiency. *Formal and certifiable guarantees*: connecting hardware-level numerical approximations and device variability to formal correctness and trust guarantees remains open. These challenges define the frontier where the vertical co-design methodology demonstrated by CoCoSys will be essential.

Conclusion

Cognitive AI, the integration of neural learning, symbolic reasoning, and probabilistic inference, represents a qualitative shift in what AI systems can do, and it demands a qualitative shift in how we build the hardware to run them. This article has presented the CoCoSys program as a systematic response, spanning four co-designed thrusts: neural, symbolic, and probabilistic algorithms; algorithm-hardware co-design; technology-driven hardware motifs; and collective and collaborative intelligence. Threaded through all four, neuro-symbolic AI served as an end-to-end case study, carrying a single workload class from algorithm restructuring through reconfigurable dataflows to a heterogeneous RRAM-SRAM chip.

The central lesson is one of composability: gains that are modest in isolation become transformative when stacked vertically across the full algorithm-to-silicon stack. The order-of-magnitude efficiency improvements demonstrated by CogSys, REASON, the NSAI system-on-chip, and ReCA are not independent achievements; they are facets of a single co-designed system view that reaches from silicon to the collective.

Looking ahead, these results define the center's next steps: extending co-design from inference to training, making gradient flow through discrete symbolic operators hardware-tractable; scaling neuro-symbolic-probabilistic composition to LLM-scale systems by carrying compositional-pipeline optimizations into specialized silicon; integrating the RRAM-SRAM cognitive fabric with 3D stacking and chiplets alongside sister JUMP 2.0 centers; consolidating the center's profiling campaigns into an open benchmark suite that jointly scores accuracy, reasoning quality, robustness, and efficiency; and driving end-to-end demonstrations in cooperative robotics and untethered mixed reality toward the sub-millisecond-latency and 100× energy-efficiency grand-challenge targets.

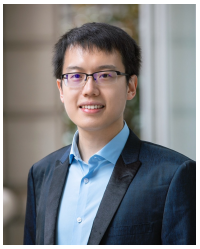
As AI moves from perception toward cognition, the architecture of intelligent systems is still being written. The co-design principles, silicon demonstrations, and open challenges catalogued here offer a foundation, and an invitation, for the next decade of cognitive computing research.

REFERENCES

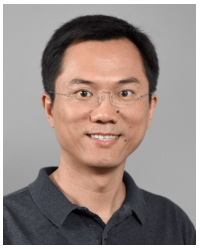
1. T. H. Trinh, Y. Wu, Q. V. Le, H. He, and T. Luong, "Solving olympiad geometry without human demonstrations," *Nature*, vol. 625, pp. 476–482, 2024.
2. C. J. Kymn, D. Kleyko, E. P. Frady, C. Bybee, P. Kanerva, F. T. Sommer, and B. A. Olshausen, "Computing with residue numbers in high-dimensional representation," *Neural Computation*, vol. 37, no. 1, pp. 1–37, 2024.
3. X. Yu, A. Thomas, I. Gomez Moreno, L. Gutierrez, and T. Rosing, "Lifelong intelligence beyond the edge using hyperdimensional computing," in *23rd ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, 2024.
4. S. K. Roy, H.-M. Ou, M. G. Ahmed, P. Deaville, B. Zhang, N. Verma, P. K. Hanumolu, and N. R. Shanbhag, "Compute sndr-boosted 22-nm mram-based in-memory computing macro using statistical error compensation," *IEEE Journal of Solid-State Circuits*, vol. 60, no. 3, pp. 1092–1102, 2024.

5. Z. Wan, C.-K. Liu, H. Yang, R. Raj, C. Li, H. You, Y. Fu, C. Wan, A. Samajdar, Y. C. Lin *et al.*, "Towards cognitive ai systems: Workload and characterization of neuro-symbolic ai," in *2024 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. IEEE, 2024, pp. 268–279.
6. Z. Wan, H. Yang, J. Qian, R. Raj, J. Park, C. Wang, A. Raychowdhury, and T. Krishna, "Compositional ai beyond llms: System implications of neuro-symbolic-probabilistic architectures," in *Proceedings of the 31st ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), Volume 1*, 2026, pp. 67–84.
7. A. Das, H. Kim, S. Lee, A. Raha, D. A. Mathaikutty, and V. Raghunathan, "MOSAIC: A workload-driven simulation and design-space exploration framework for heterogeneous NPUs," *arXiv preprint arXiv:2606.05362*, 2026.
8. A. Das, A. Raha, S. Kundu, S. K. Ghosh, D. Mathaikutty, and V. Raghunathan, "GraNNite: Enabling high-performance execution of graph neural networks on resource-constrained neural processing units," in *International Joint Conference on Neural Networks (IJCNN)*, 2025.
9. ——, "XAMBA: Enabling efficient state space models on resource-constrained neural processing units," *arXiv preprint arXiv:2502.06924*, 2025.
10. A. Raha, S. Kundu, S. N. Sridhar, S. Kundu, S. K. Ghosh, A. Palla, A. Das, D. Crews, and D. A. Mathaikutty, "Llm-npu: Towards efficient foundation model inference on low-power neural processing units," in *2025 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*. IEEE, 2025, pp. 1–8.
11. G. Jeong, S. Damani, A. R. Bambhaniya, E. Qin, C. J. Hughes, S. Subramoney, H. Kim, and T. Krishna, "VEGETA: Vertically-integrated extensions for sparse/dense GEMM tile acceleration on CPUs," in *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2023.
12. D. Shi, Y. Fu, X. Yuan, Z. Yu, H. You, S. Li, X. Dong, J. Kautz, P. Molchanov, and Y. C. Lin, "LaCache: Ladder-shaped KV caching for efficient long-context modeling of large language models," in *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.
13. R. Raj, S. Banerjee, N. Chandra, Z. Wan, J. Tong, A. Samajdhar, and T. Krishna, "Scale-sim v3: A modular cycle-accurate systolic accelerator simulator for end-to-end system analysis," in *2025 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. IEEE, 2025, pp. 186–200.
14. X. Hu, H. Mun, J. Meng, Y. Liao, A. Sridharan, and J.-s. Seo, "A 28nm 20.9-137.2 tops/w output-stationary sram compute-in-memory macro featuring dynamic look-ahead zero weight skipping and runtime partial sum quantization," in *2025 IEEE Custom Integrated Circuits Conference (CICC)*. IEEE, 2025, pp. 1–3.
15. K. Prabhu, J. Yu, X. A. Pan, Z. Xie, A. Aleshire, Z. Chen, A. A. Ratnani, and P. Raina, "Voyager: An end-to-end framework for design-space exploration and generation of dnn accelerators," *arXiv preprint arXiv:2509.15205*, 2025.
16. A. Ramachandran, S. Kundu, A. Raha, S. Kundu, D. K. Mathaikutty, and T. Krishna, "Accelerating LLM inference with flexible N:M sparsity via a fully digital compute-in-memory accelerator," in *IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, 2025.
17. S. Negi, U. Saxena, D. Sharma, and K. Roy, "HCiM: ADC-less hybrid analog-digital compute in memory accelerator for deep learning workloads," in *Proceedings of the 30th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2025.
18. A. Bhattacharjee, A. Moitra, and P. Panda, "ClipFormer: Key-value clipping of transformers on memristive crossbars for write noise mitigation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2024.
19. S. Yoo, A. Holla, S. Sanyal, D. E. Kim, F. Iacopi, D. Biswas, J. Myers, and K. Roy, "Taxi: Traveling salesman problem accelerator with x-bar-based ising macros powered by sot-mrams and hierarchical clustering," in *2025 62nd ACM/IEEE Design Automation Conference (DAC)*. IEEE, 2025, pp. 1–7.
20. C.-K. Liu, Z. Wan, Y.-S. Noh, M. Ibrahim, S. D. Spetalnick, T. Krishna, W.-S. Khwa, A. S. Lele, Y.-D. Chih, M.-F. Chang *et al.*, "A 40-nm programmable heterogeneous soc with 5.625/0.85 mb rram/sram for accelerating neuro-symbolic ai models," *IEEE Journal of Solid-State Circuits (JSSC)*, 2026.
21. Z. Wan, Y. Du, M. Ibrahim, J. Qian, J. Jabbour, Y. Zhao, T. Krishna, A. Raychowdhury, and V. J. Reddi, "ReCA: Integrated acceleration for real-time and efficient cooperative embodied autonomous agents," in *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, 2025, pp. 982–997.
22. Z. Wan, N. Chandramoorthy, K. Swaminathan, P.-Y. Chen, K. Bhardwaj, V. J. Reddi, and A. Raychowdhury, "MuIBERRY: Enabling bit-error robustness for energy-efficient multi-agent autonomous systems," in *Proceedings of the 29th ACM International Conference on Architectural Support for Programming*

Languages and Operating Systems, Volume 2, 2024, pp. 746–762.
23. Q. Zhao, X. Yu, S. Hu, and T. Rosing, "MultimodalHD: Federated learning over heterogeneous sensor modalities using hyperdimensional computing," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2024.
24. Z. Wan, H. Yang, R. Raj, C.-K. Liu, A. Samajdar, A. Raychowdhury, and T. Krishna, "CogSys: Efficient and scalable neurosymbolic cognition system via algorithm-hardware co-design," in *2025 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2025, pp. 775–789.
25. Z. Wan, C.-K. Liu, J. Qian, H. Yang, A. Raychowdhury, and T. Krishna, "REASON: Accelerating probabilistic logical reasoning for scalable neuro-symbolic intelligence," in *2026 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2026, pp. 1–16.



Zishen Wan is an Assistant Professor of Computer Science at Columbia University, New York, NY, USA. His research interests include AI systems, computer architecture, and VLSI design. Wan received the Ph.D. degree in electrical and computer engineering from Georgia Tech. Contact him at zishen.wan@columbia.edu.



Yu Cao is the Louis John Schnell Professor of Electrical and Computer Engineering at the University of Minnesota, Minneapolis, MN, USA. His research interests include neural-inspired computing, hardware-algorithm co-design, and heterogeneous integration. Cao received the Ph.D. degree from the University of California, Berkeley. Contact him at yucao@umn.edu.

Sumeet K. Gupta is the Elmore Associate Professor of Electrical and Computer Engineering at Purdue University, West Lafayette, IN, USA. His research interests include device-circuit co-design, neuromorphic computing, and spintronics. Gupta received the Ph.D. degree from Purdue University. Contact him at guptask@purdue.edu.

Larry P. Heck is the Rhesa S. Farmer Distinguished Chair and a Professor at Georgia Tech, Atlanta, GA, USA. His research interests include conversational AI, multimodal learning, and machine intelligence. Heck received the Ph.D. degree from Georgia Tech and is an IEEE Fellow. Contact him at larryheck@gatech.edu.

Tushar Krishna is an Associate Professor at Georgia Tech, Atlanta, GA, USA. His research interests include computer architecture, on-chip networks, and AI accelerator design. Krishna received the Ph.D. degree in electrical engineering and computer science from MIT. Contact him at tushar@ece.gatech.edu.

Yingyan (Celine) Lin is an Associate Professor of Computer Science at Georgia Tech, Atlanta, GA, USA. Her research interests include efficient AI, trustworthy machine learning, and algorithm–hardware co-design. Lin received the Ph.D. degree from the University of Illinois Urbana-Champaign. Contact her at celine.lin@gatech.edu.

Azad Naeemi is a Professor of Electrical and Computer Engineering at Georgia Tech, Atlanta, GA, USA. His research interests include nanoelectronics, spintronics, and interconnects. Naeemi received the Ph.D. degree from Georgia Tech. Contact him at azad@gatech.edu.

Bruno Olshausen is a Professor at the University of California, Berkeley, Berkeley, CA, USA, and directs the Redwood Center for Theoretical Neuroscience. His research interests include sensory coding, visual perception, and brain-inspired intelligence. Olshausen received the Ph.D. degree from Caltech. Contact him at baolshausen@berkeley.edu.

Priyadarshini Panda is the Lloyd F. Hunt Early Career Chair and an Associate Professor at the University of Southern California, Los Angeles, CA, USA. Her research interests include neuromorphic computing, spiking neural networks, and compute-in-memory systems. Panda received the Ph.D. degree from Purdue University. Contact her at priya.panda@usc.edu.

Jan M. Rabaey is Professor Emeritus and the Donald O. Pederson Distinguished Professor at the University of California, Berkeley, Berkeley, CA, USA. His research interests include low-power circuits, wireless systems, and ubiquitous computing. Rabaey received the Ph.D. degree from KU Leuven and is an IEEE Fellow. Contact him at jan_rabaey@berkeley.edu.

Vijay Raghunathan is a Professor of Electrical and Computer Engineering at Purdue University, West Lafayette, IN, USA. His research interests include embedded systems, wearable electronics, and low-power design. Raghunathan received the Ph.D. degree from UCLA. Contact him at vr@purdue.edu.

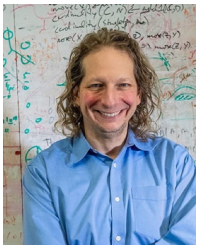
Priyanka Raina is an Associate Professor of Electrical Engineering and, by courtesy, of Computer Science at Stanford University, Stanford, CA, USA. Her research interests include domain-specific architectures, hardware–software co-design, and energy-efficient machine learning. Raina received the Ph.D. degree from MIT. Contact her at praina@stanford.edu.

Tajana S. Rosing is the Fratamico Endowed Chair and a Professor at the University of California San Diego, La Jolla, CA, USA. Her research interests include energy-efficient computing, cyber-physical systems, and distributed systems. Rosing received the Ph.D. degree from Stanford University and is an IEEE Fellow. Contact her at trosing@ucsd.edu.

Kaushik Roy is the Edward G. Tiedemann, Jr. Distinguished Professor at Purdue University, West Lafayette, IN, USA. His research interests include AI hardware, neuromorphic computing, and low-power electronics. Roy received the Ph.D. degree from the University of Illinois Urbana-Champaign. Contact him at kaushik@purdue.edu.

Jae-Sun Seo is an Associate Professor of Electrical and Computer Engineering at Cornell Tech, New York, NY, USA. His research interests include machine-learning accelerators, neuromorphic hardware, and hardware-efficient AI. Seo received the Ph.D. degree from the University of Michigan. Contact him at js3528@cornell.edu.

Naresh R. Shanbhag is the Jack Kilby Professor at the University of Illinois Urbana-Champaign, Urbana, IL, USA. His research interests include machine-learning systems, communication systems, and integrated circuits. Shanbhag received the Ph.D. degree from the University of Minnesota and is an IEEE Fellow. Contact him at shanbhag@illinois.edu.



Josh Tenenbaum is the Whitehead Professor of Computational Cognitive Science at MIT, Cambridge, MA, USA, and a principal investigator in CSAIL. His research interests include perception, learning, and common-sense reasoning. Tenenbaum received the Ph.D. degree from MIT. Contact him at jbt@mit.edu.

Anand Raghunathan is the Silicon Valley Professor of Electrical and Computer Engineering at Purdue University, West Lafayette, IN, USA. His research interests include system-on-chip design, domain-specific architectures, and embedded security. Raghunathan received the Ph.D. degree from Princeton University. Contact him at raghunathan@purdue.edu.

Arijit Raychowdhury is the Steve W. Chaddick Chair and a Professor at Georgia Tech, Atlanta, GA, USA. His research interests include energy-efficient circuits, power management, and emerging computing hardware. Raychowdhury received the Ph.D. degree from Purdue University. Contact him at arijit.raychowdhury@ece.gatech.edu.