ICML 2026 · POSITION PAPER

Artificial Intelligence Needs Meta Intelligence

The Case for Metacognitive AI — self-monitoring plus resource-rational control as a design principle

Sergei Chuprov[1], Richard D. Lange[2], Leon Reznik[2], Paulo Shakarian[3], Raman Zatsarenko[2], Dmitrii Korobeinikov[2]

[1]University of Texas Rio Grande Valley · [2]Rochester Institute of Technology · [3]Syracuse University

✉ sergei.chuprov@utrgv.edu · rdlvcs@rit.edu · leon.reznik@rit.edu · pashakar@syr.edu · rz4983@rit.edu · dk9148@rit.edu

The University of Texas Rio Grande Valley · RIT · Syracuse University

Access this paper here · LDQIS Lab · *InteFL* Framework GitHub Repo · Sergei Chuprov's Faculty Website

POSITION OF THIS PAPER

Metacognition, encompassing self-monitoring and resource-allocation inspired by cognitive science, should be a **general design principle for AI systems**. Further, we call for a **theoretical framework** to unify current efforts, and **actionable strategies** to translate metacognition research into practical AI.

Why this matters now

Recent gains in inference and training come at a **soaring computational cost**. Reasoning LLMs and large-scale training strain time, memory, and energy, yet give few guarantees on correctness or robustness — and the way we decide how much computation a problem deserves stays crude, usually a fixed budget set by hand.

The same pressure appears everywhere compute is finite and stakes vary across inputs: from perception on the edge to multi-step reasoning in the cloud. Treating every input the same wastes effort on the easy ones and starves the hard ones.

Humans tend to do better in this: we *sense* when a task is hard, *deliberate* longer when it matters, and *stop* when it does not. A century of cognitive science calls this **metacognition**. A metacognitive system spends resources **up front**, per instance (**Fig. 1**).

The goal is not less computation, but **computation spent where it counts**.

Objective, InteFL, case study

A resource-rational objective trades task loss against compute:

$$\min_{w,w'} \frac{1}{\mathrm{N}} \sum_{i=1}^{N} \mathbb{E}_{\mathbf{c_i} \sim \pi_{\mathbf{w}'}} \left[\mathrm{loss}\left(\mathbf{x_i}, f_{\mathbf{w}}(\mathbf{x_i}; \mathbf{c_i})\right) + \alpha\, \mathrm{cost}(\mathbf{c_i}) \right] + \beta\, \|\pi_{\mathbf{w}'}\|$$

Federated Learning (FL) is naturally well-positioned for metacognition: it already separates an object level (local client training) from a meta level (server aggregation). A monitoring function $\mathcal{M}$ scores each update by trust and distance from consensus; control accepts, rejects, or re-weights it (**Fig. 2**), self-adapting under drift or attack.

In **InteFL** (**Fig. 3**) [1], the aggregation strategy is the control mode c: the object-level step is one aggregation over client updates, π selects among strategies (FedAvg [6], Multi-Krum, PID-based exclusion), and its parameters w' are the exclusion thresholds tuned against observed performance. In **learning**, investing in a controller is especially worthwhile because its overhead is **amortized over the model's lifetime**.

On OctMNIST (**Fig. 4**), a tuned PID rule (2.5σ) converges lowest and best excludes malicious clients; Other baselines miss attackers in most cases.

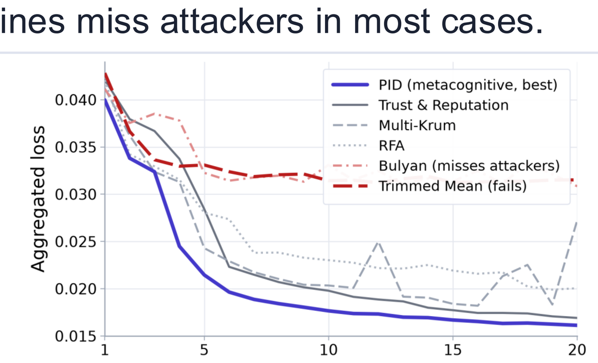

Figure 4. OctMNIST, 20 clients (4 adversarial): the tuned PID rule converges lowest; Trimmed Mean and Bulyan miss attackers and stall.

What metacognition is

Following Nelson and Narens [2], cognition is **bi-level**: object-level processes (perception, learning, reasoning) are watched by a **meta-level**.

It adds **monitoring** — reading one's own state via cheap cues — and **control**, reallocating resources or switching strategy: a System-2 intervention [3].

Research tasks

We call for concerted progress on three research tasks:

1. **Characterize the performance–compute trade-off** — relate accuracy, robustness, and generalization to time, memory, and energy (PAC learnability, anytime algorithms, efficient models).
2. **Develop robust, low-overhead cues** — error-detection rules, early-exit signals, XAI attribution — reliable yet computable in near-constant time.
3. **Build well-calibrated controllers** with uncertainty quantification (temperature scaling, conformal prediction, abduction) [4].

Solving these turns scattered tricks into **cumulative science** and provably good metacognitive control.

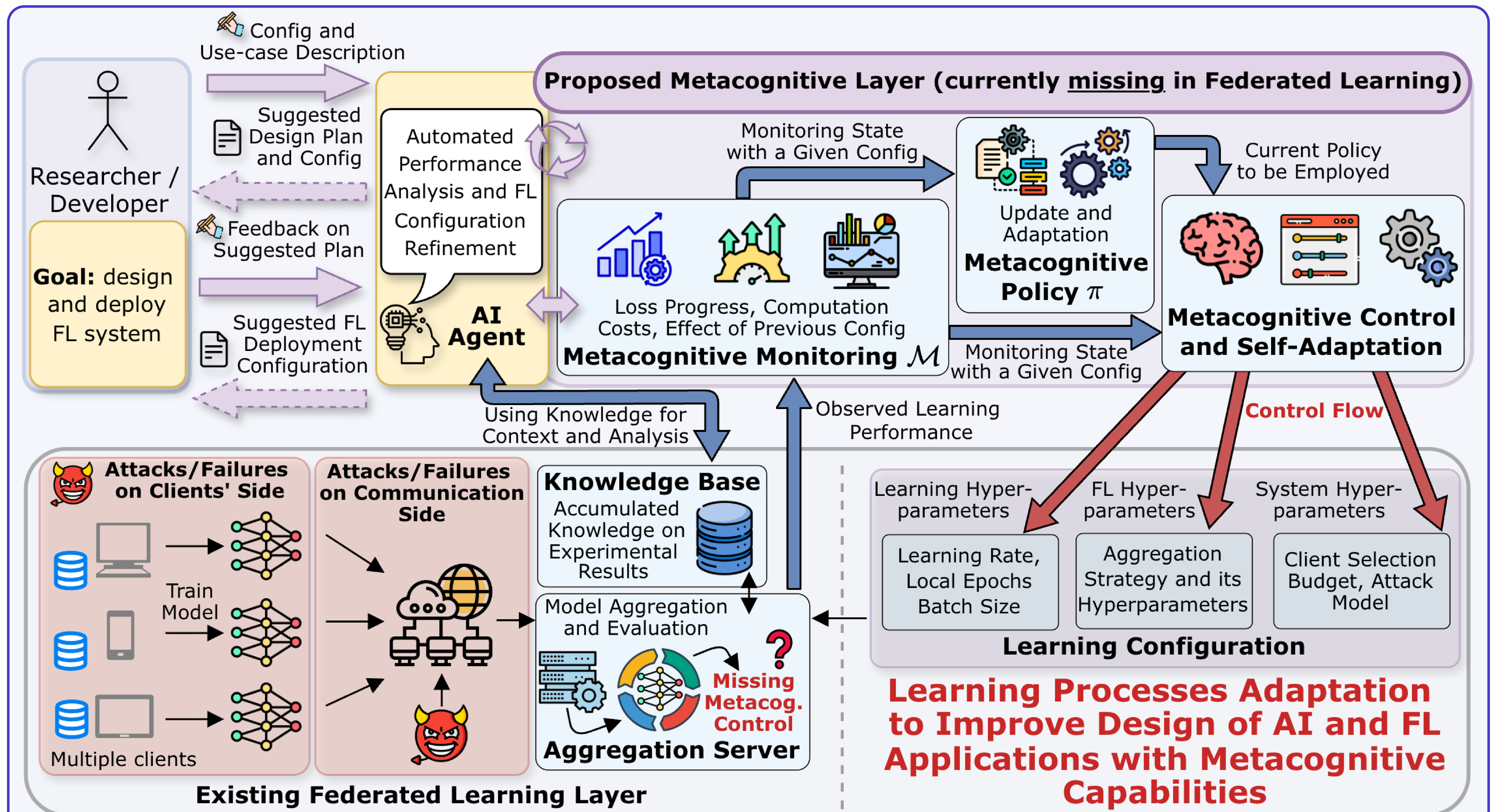

Figure 1 A metacognitive layer for FL, which is currently missing — monitoring, an adaptive policy π, and self-adaptive control atop the existing learning layer.

Challenges and alternatives

Alternatives. "Scale is all you need" [5] and post-hoc "generate-and-verify" pipelines raise accuracy but not efficiency or real-time robustness; and unlike early-exit or mixture-of-experts, a metacognitive controller is an explicit, learned, runtime-active layer.

Challenges — three open problems:

1. Cues and controllers at **low overhead** (compute and system complexity).
2. Translating cognitive-science concepts into **precise computational infrastructure**.
3. **Failures of metacognition**: detection may not be actionable [7], and controllers can misjudge competence or act globally-counterproductively.

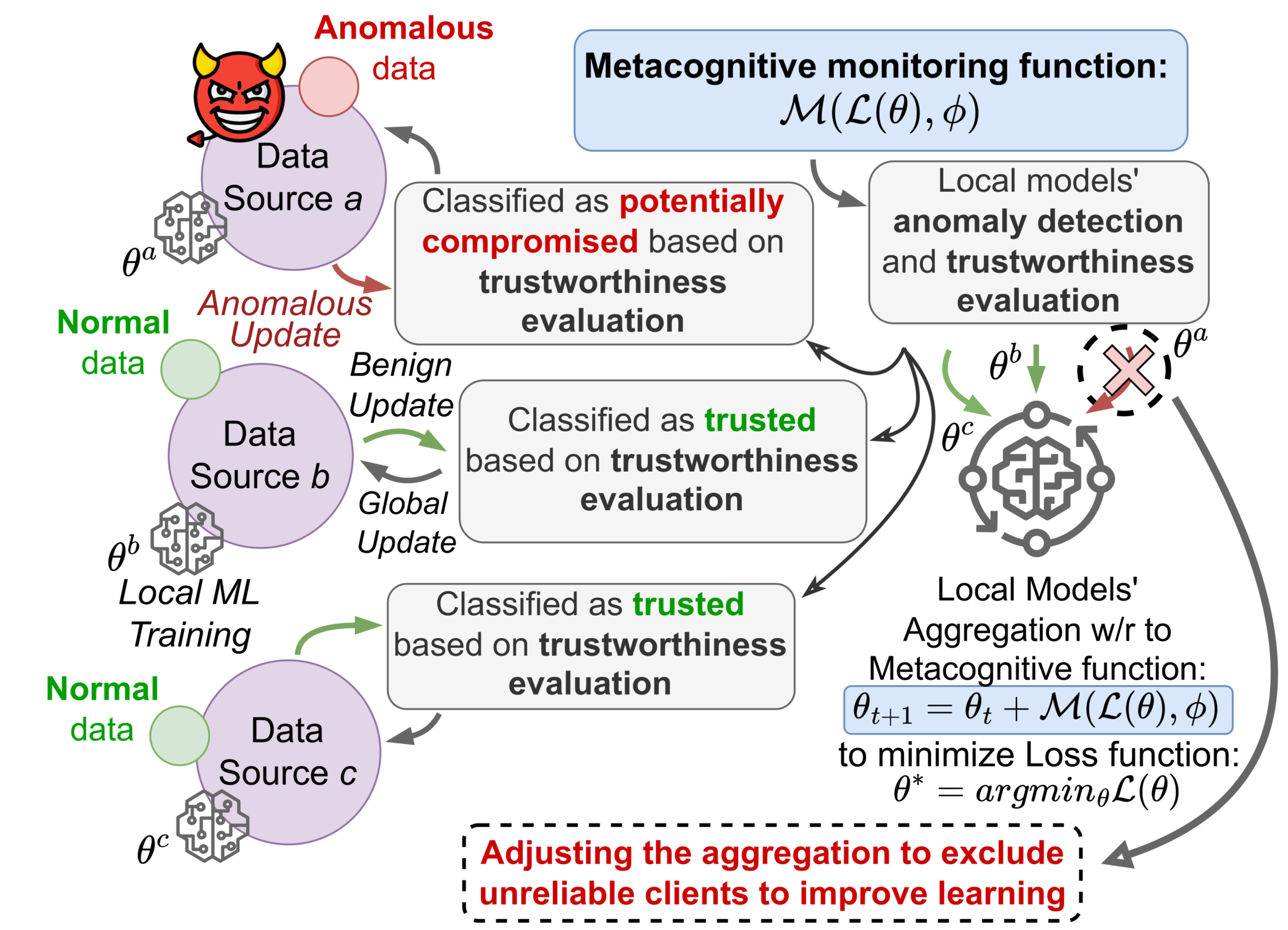

Figure 2. Monitoring $\mathcal{M}$ in FL: client-trust evaluation and selective aggregation exclude anomalous updates.

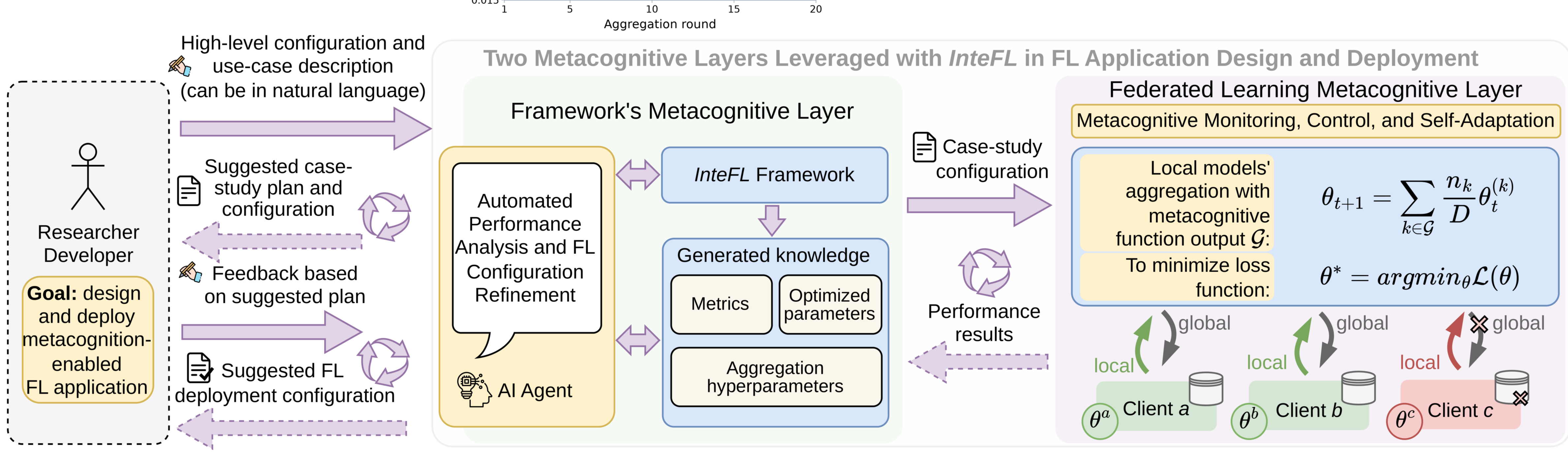

Figure 3. InteFL — an AI-assisted design layer (natural-language configuration plus a feedback loop) atop an FL metacognitive layer. The aggregation strategy plays the role of mode c. Architecture-agnostic (also BiomedBERT, GPT-2).

Call to action

As a design principle it yields AI that is at once more efficient, safer, and more secure:

- Adaptive resource use by input difficulty and priority
- Self-monitoring for anomalies and OOD inputs
- Robustness via introspection on adversarial inputs
- Recovery from concept drift by self-recalibration
- Auditable decisions that justify their reasoning

We therefore call on the community to **instantiate metacognitive control** — not just detection — and to build **shared benchmarks and open frameworks** for it (e.g., InteFL [1]).

Conclusion

We argue that communities from cognitive science to AI and ML should jointly develop metacognition and make it a **leading paradigm of AI design**.

Long posited as a key enabler of general intelligence, metacognition is — on our pragmatic view — a concrete paradigm for managing the growing computational cost of AI, and a practical path to efficient, robust, and secure systems, as **InteFL** [1] demonstrates for FL.

References

[1] D. Korobeinikov, R. Zatsarenko, S. Chuprov, A. Barea, and L. Reznik. IntelliFL Framework: Optimizing Federated Learning with Metacognition for Application Design and Deployment. IEEE Intelligent Systems, pp. 1–10, 2026. DOI: 10.1109/MIS.2026.3658072.

[2] T. O. Nelson and L. Narens. Metamemory: A Theoretical Framework and New Findings. The Psychology of Learning and Motivation, Vol. 26, pp. 125–173, Academic Press, 1990. DOI: 10.1016/S0079-7421(08)60053-5.

[3] R. Ackerman and V. A. Thompson. Meta-Reasoning: Monitoring and Control of Thinking and Reasoning. Trends in Cognitive Sciences, Vol. 21, No. 8, pp. 607–617, 2017. DOI: 10.1016/j.tics.2017.05.004.

[4] F. Lieder and T. L. Griffiths. Resource-Rational Analysis: Understanding Human Cognition as the Optimal Use of Limited Computational Resources. Behavioral and Brain Sciences, Vol. 43, Article e1, Cambridge University Press, 2020. DOI: 10.1017/S0140525X1900061X.

[5] R. S. Sutton. The Bitter Lesson. Incomplete Ideas (Self-published essay), Mar. 13, 2019. Available: http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

[6] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR Vol. 54, pp. 1273–1282, 2017. arXiv: 1602.05629.

[7] P. Shakarian, G. I. Simari, and N. D. Bastian. Probabilistic Foundations for Metacognition via Hybrid-AI. arXiv preprint arXiv:2502.05398, 2025.

Take-aways

- Metacognition is a practical design principle for efficient, robust, secure AI
- Adopting it lowers cost, broadens access, and unifies scattered efforts into cumulative science
- **Research tasks for the community:** build the theory, cues, and benchmarks for metacognitive control