



Prospective Double Jeopardy for the Metacognitive Layer: Susceptibilities of Multi-Exit Networks and Input-Adaptive Deep Neural Networks to Sponge Poisonings/Attacks

+

Details



Contribution type: article

Title: Prospective Double Jeopardy for the Metacognitive Layer: Susceptibilities of Multi-Exit Networks and Input-Adaptive Deep Neural Networks to Sponge Poisonings/Attacks

Authors:

Steve Chan, Decision Engineering Analysis Laboratory, VTIRL, VT, United States

Keywords: Artificial Intelligence, Data Center, Infrastructure, Anomaly Detection, Multi-Exit Network, Input-Adaptive Deep Neural Network, Adaptive Neural Network, Metacognitive Layer

Abstract:

It is envisioned that a high efficacy Metacognitive Layer (MCL) for Artificial Intelligence (AI) Systems (AIS) might be able to reduce AI hallucinations, improve AI construct validity, further explainability, and enhance AI Decision-Making (DM), among other value-added propositions. MCL implementations are on the rise within the AI ecosystem, and promising performance gains have already been reported. In a number of these instances, MCL is deployed atop Deep Neural Networks (DNNs). As part of the Kahneman System 1 (Fast Thinking) and System 2 (Slow Thinking) paradigm, MCL might also leverage Multi-Exit Networks (MENs), which can provide opportunities for early exits for System 1 or continued processing (later exits) for System 2. Oftentimes, these MENs are also built atop DNNs. As to be expected, the promise of the MCL is also accompanied by prospective vulnerabilities at the MEN/DNN. In particular, adversarial actors are increasingly using energy-latency attacks, which are designed to compromise the efficacy of MEN/DNN and MCL by substantially increasing the energy consumption and/or response latency. Although there are mitigation mechanisms available and being researched, revisiting the prospective “Achilles heel” attack surface of the MCL seems prudent. Adaptive Neural Networks (AdNN), the subclass of Input-Adaptive DNN (IA-DNN), and MEN (as a specific type of IA-DNN) are explored, and specific susceptibilities to attack vectors, such as the Sponge Poisoning/Attack (SPA), are exposed.

Publication Date: June 7, 2026

Presented during:

 REASON 2026, The First International Conference on Reasoning and Decision-making in Intelligent Systems

Dates: June 7, 2026 to June 11, 2026

Location: Porto, Portugal

Venue:

Hotel Novotel Porto Gaia

Rua Martir Sao Sebastiao, Afurada,
4400-499 Vila Nova de Gaia

[Hotel website](#)

Copyright (c) DTR Society, 2026

Get The Latest News From Us

Delivered To Your Inbox

First name

Last name

Email

☐ I acknowledge and accept the conduct policies

Subscribe