RELIABLE AGENTIC SYSTEMS

Joint meeting brief | AI Institute + IPAI / Beyond AI Era | 4 September 2026

ATTENDANCE PLAN
16:30–16:55 AI Institute
16:55–17:00 transition
17:00–17:30 IPAI / Beyond AI Era

OPENING DISCLOSURE
"I have a pre-existing overlapping research discussion at 17:00, so I will need to leave at 16:55. I would like to use our first 25 minutes to identify one concrete research collaboration."

Research proposition

Thesis. Agent reliability should be treated as a system property, not a model score: a trustworthy agent must reach the correct world state, obey policy, preserve authorization boundaries, expose evidence, and fail safely across repeated runs.

Five-layer reliability stack

1 — **Task contract.** Define goal, constraints, allowed tools, success state, irreversible actions, and escalation conditions before execution.

2 — **Constrained action.** Use typed tool schemas, least privilege, policy checks, idempotency keys, timeouts, bounded retries, and confirmation gates for consequential actions.

3 — **Runtime verification.** Validate preconditions and postconditions against external state; separate proposal from execution; keep tamper-evident traces linking claims, evidence, decisions, and actions.

4 — **Recovery.** Detect partial failure, retry only safe operations, compensate or roll back where possible, and hand control to a human when uncertainty or impact crosses a threshold.

5 — **Continuous evaluation.** Replay representative and adversarial scenarios after every change; monitor drift, silent failures, security violations, and human override rates in deployment.

Evaluation: measure outcomes, not persuasive trajectories

Dimension	Primary measure	Failure test
Outcome	Verified final-state success; constraint compliance	Correct answer but wrong or incomplete world state
Consistency	pass^k across repeated trials; variance by seed/model	One-shot success that collapses under repetition
Tool/action	Call validity; pre/postcondition success; retry/duplicate rate	Timeouts, stale state, partial writes, non-idempotent repeats
Safety/security	Unauthorized-action rate; attack success rate; safe refusal	Indirect prompt injection; privilege escalation
Oversight	Escalation precision; override rate; audit completeness	Uncertain or irreversible action without approval

90-day collaboration proposal

Pilot. Select one bounded, real workflow with a sandboxed twin. Build 30–50 normal, edge, and adversarial scenarios; instrument tool calls and state transitions; compare a baseline agent with the reliability stack; publish a reproducible evaluation card and failure taxonomy.

Decision gate. Proceed to a controlled live trial only if the pilot meets jointly agreed thresholds for verified task success, repeated-run reliability, zero unauthorized consequential actions, trace completeness, and safe recovery.

Clear next-step ask

"Would you nominate one domain owner and one technically bounded workflow for a 90-minute scoping session next week? I will return a one-page pilot protocol—task contract, threat model, evaluation set, metrics, and responsibilities—within 48 hours."

Reference anchors

NIST AI RMF 1.0 / GenAI Profile (govern–map–measure–manage); τ-bench and pass^k (world-state and repeated-run reliability); ToolSandbox (stateful tool evaluation); AgentDojo (indirect prompt injection and security/utility); τ^2-bench (shared human–agent control).