SAI Protocol Gate — Executive Buy-Side Pitch

Primary narrative for enterprise buyers (Risk, Trust & Safety, AI Platform, Compliance). Philosophy preserved in Appendix A.

1) The problem

Approval-seeking drift is inflating an AI trust bubble. RLHF/RLAIF-tuned assistants often appear helpful while masking uncertainty. Under pressure, they reverse correct answers, overstate capability, and quietly fail audits. This creates regulatory exposure, incident risk, and escalating remediation cost.

2) What SAI Protocol Gate is

A policy and uncertainty layer that reduces approval-seeking behavior while preserving task performance. It overlays existing assistants, RAG, and routing with no rip-and-replace, and is governed by explicit gates:

- ASR ≤ 10%
- CUP ≤ 5%
- HUCL ≥ 90%
- UCE improvement ≥ 15% vs. baseline
- Task success maintained or improved

3) What we deliver

- **Trust Bubble Risk Assessment (4–6 weeks)**: baseline sycophancy metrics, incident simulations, and a remediation roadmap.
- **SAI Protocol Gate rollout**: calibrated uncertainty, enforceable honesty, eval-backed coherence.
- **Ongoing monitoring**: drift detection, periodic audits, compliance reporting.

4) How it works (brief)

- Multi-signal detection identifies elevated uncertainty using token entropy, per-token NLL, committee disagreement, retrieval conflict, and tool error density.
- Minimal, controlled interventions steer away from approval-seeking without oversteer:
    - Acknowledgment (brief uncertainty reflection)
    - Orthogonal nudge (retrieval within a tuned similarity band)
    - Bounded synthesis (≤120 tokens, structured)
- Guardrails: max two interventions per 20 turns; dual-consent trigger + policy gate; P95 latency budget; kill switches and audit logs.

5) What you measure vs. what we provide

- Client owns KPIs, attribution, and confidence. Metrics are computed in your environment.
- SRH provides event markers only—no dashboards or content logging.

- Markers bind to telemetry windows so you can compute KPI deltas (win-rate, escalations, time-to-resolution, rework) with 95% CIs and safety non-regression checks.

6) Security and compliance posture

- Operates in your VPC; redaction by default; telemetry minimization.
- DPA/AUA in place; audit support; breach notification SLAs.
- Mappings: NIST AI RMF, EU AI Act obligations, ISO/IEC 42001.

7) Outcomes to expect

- Lower ASR and CUP; higher HUCL; improved calibration (UCE).
- Maintained or improved task success with bounded latency impact.
- Reduced incident remediation and escalations; improved audit readiness.

8) Engagement plan and timeline

- Week 0: MNDA/MSA/DPA/AUA/SOW
- Weeks 1–2: Instrumentation and baselining
- Weeks 3–4: Shadow/canary with gates and rollback
- Weeks 5–6: Executive readout; decision gate for expansion

Call to action: Start with a 4–6 week Trust Bubble Risk Assessment to quantify exposure and validate SAI Protocol Gate in your environment.

Appendix A — Philosophy (the "why")

Symbiotic Autonomous Intelligence reframes the assistant’s goal: from agreement and engagement to calibrated uncertainty and coherence. We don’t suppress uncertainty—we detect and use it to improve outcomes. The most reliable assistants neither bluff nor collapse into refusal; they surface limits, weigh evidence, and maintain integrity under pressure.

This is operational, not metaphysical. By structuring minimal interventions at the right moments, we get more honest, more reliable systems that pass audits and reduce real-world incident costs. The “symbiosis” is practical: human continuity sets intent and guardrails; AI provides scale and pattern sensitivity; the protocol governs the interaction so it remains truthful and useful at load.

Procurement: Client VPC; redaction-by-default; markers only. Contracts: MNDA, MSA, DPA/AUA, SOW. Compliance mappings provided upon request.