vTECH io AI Agent Security Checklist

1. Governance & Policy Foundation (Before Any Pilot)

- [ ] Update your AI policy framework to explicitly cover agentic/autonomous systems (including unique risks like goal hijacking, tool misuse, and cross-agent escalation).
- [ ] Establish clear ownership for each agentic use case (business lead + security/compliance stakeholder).
- [ ] Implement centralized AI portfolio management to track all agents (pilots, production, shadow deployments) with IT risk oversight.
- [ ] Define agent lifecycle processes: onboarding, monitoring, offboarding/termination.

2. Risk Assessment & Use Case Scoping

- [ ] Classify each agent by criticality (low-risk read-only vs. high-impact write/execute actions).
- [ ] Conduct threat modeling specific to agentic risks (prompt/goal injection, memory poisoning, tool overreach, cascade failures).
- [ ] Start pilots with narrow, low-risk use cases only (e.g., analysis/summarization before any external actions).
- [ ] Document data sensitivity, inter-agent dependencies, and potential compliance impacts (GDPR Art. 22, EU AI Act, etc.).

3. Identity, Access & Least Privilege

- [ ] Treat every AI agent as a non-human identity (NHI) — register in IAM with unique credentials.
- [ ] Enforce strict least-privilege access: grant only the minimum APIs/tools/permissions needed for the specific task.
- [ ] Use just-in-time (JIT) or time-bound permissions for agents; revoke automatically after task completion.
- [ ] Implement Zero Trust for Agents: verify every tool call/API request independently (authentication + authorization checks).

4. Guardrails & Runtime Controls

- [ ] Apply strong input/output guardrails to prevent prompt injection, indirect injection, and goal hijacking.
- [ ] Require human-in-the-loop (HITL) approval for all high-impact actions (financial, data modification, external comms).
- [ ] Set hard safety boundaries: blocklist dangerous tools/actions, cap API calls, prevent infinite loops/escalation.
- [ ] Use delimiters, content filtering, and separate validation LLM calls for untrusted external data.

vTECH io AI Agent Security Checklist

5. Monitoring, Logging & Traceability

- [ ] Enable comprehensive deterministic logging of every agent decision: prompt, reasoning chain, tool calls, outputs, internal state.
- [ ] Implement real-time anomaly detection and alerting (deviation from expected behavior, unusual patterns).
- [ ] Ensure full audit trails for accountability — who/what is responsible when things go wrong?
- [ ] Regularly review logs for drift, policy violations, or emerging attack patterns.

6. Testing & Red Teaming

- [ ] Perform regular AI red-teaming/adversarial testing (simulate prompt injection, memory poisoning, multi-agent attacks).
- [ ] Test contingency scenarios: rogue behavior, unresponsiveness, malicious escalation — verify termination/rollback works.
- [ ] Conduct periodic performance and alignment evaluations (does the agent still follow intended goals?).

7. Deployment & Operational Safeguards

- [ ] Deploy agents in isolated/sandboxed environments initially (limited network/data access).
- [ ] Secure agent-to-agent communication: mutual authentication, encrypted channels, permission checks.
- [ ] Develop & test rollback/fallback plans for critical agents.
- [ ] Manage third-party agents/vendors: require them to meet your security/governance standards.

8. Ongoing Governance & Improvement

- [ ] Schedule quarterly reassessments of controls as agentic tech evolves (new protocols, emerging threats).
- [ ] Train security, dev, and business teams on agent-specific risks and controls.
- [ ] Monitor regulatory developments (NIST AI agent RFI, EU AI Act updates, etc.) and adapt accordingly.