B

IA Agéntica y Funciones de Red

Arquitectura Zero Trust identity-first para sistemas de IA agéntica distribuidos. Por qué la identidad criptográfica debe preceder a la conectividad.

Autor: Arturo Navarro · BlueUP
Fecha: Mayo 2026
Partner tecnológico: NetFoundry / OpenZiti
Web: www.blueup.es
Clasificación: Público

1 Contextualización del Problema

Supongamos un ataque donde un agente malicioso (*rogue agent*) se infiltra a través de una cadena de credenciales comprometidas, se mueve a velocidad de máquina, alcanza servicios internos, exfiltra datos sensibles, abusa de herramientas e intenta acciones autónomas destructivas.

¿Qué arquitectura detiene esa cascada antes de que se convierta en una brecha de seguridad?

2 Visión y Pilares Estratégicos

La respuesta se centra en la consecución simultánea de tres resultados clave:

Pilar	Resultado
Seguridad mejorada	La infraestructura, herramientas, modelos de IA y servicios no son accesibles a menos que la identidad y la política creen explícitamente la ruta y la accesibilidad.
Innovación más rápida	Los equipos no dependen de cambios repetidos en la infraestructura subyacente para cada nuevo agente, modelo de servicio o entorno.
Despliegue más sencillo	La solución puede ejecutarse a través de redes existentes, la nube, contenedores de Kubernetes, sitios edge y entornos de terceros sin reconstruir la red.

> **Motivo fundamental:** una infraestructura expuesta ya no es una base de referencia viable para la IA agéntica. La seguridad no puede comenzar solo en la capa del modelo, prompt, token de API o acción, una vez que la ruta de datos ya existe. El modelo de Confianza Cero debe gobernar la propia accesibilidad.

3 Principio Fundamental: Cero Confianza e "Identidad Primero"

El principio es claro: **autenticar y autorizar antes de la conectividad**. Sin una identidad criptográfica válida, sin una política coincidente, no hay absolutamente ninguna ruta de servicio ni accesibilidad.

Esto también es relevante para la gobernanza. Las empresas quizás no puedan evitar que cada equipo experimente con agentes, pero pueden hacer que los agentes no gestionados sean ineficaces contra los recursos empresariales. Un agente puede ejecutarse, pero a menos que esté integrado en el marco de identidad, política, telemetría y auditoría, no debería poder acceder a servicios internos, invocar herramientas gobernadas, acceder a un modelo de aprobación o mover datos sensibles.

Modelo tradicional	Modelo identity-first
Conectar → verificar → filtrar	Autenticar → autorizar → conectar
Servicios expuestos por defecto	Servicios invisibles por defecto
Seguridad como filtro sobre la red	La identidad *es* la red
Token da acceso, controles posteriores contienen	Sin identidad válida = sin ruta

4 Escenario de Brecha Real

Cuando obtenemos una alerta de brecha es ya una ilustración del fallo común en entornos tradicionales de "conectar primero": el agente, la herramienta, el modelo o el servicio tienen accesibilidad primero, y la seguridad intenta ponerse al día después.

> **Pregunta clave:** ¿Tiene el agente una ruta accesible antes de ser autorizado, o la identidad y la política deciden si esa ruta puede existir en absoluto?

El ritmo de los *exploits* se ha acelerado. La IA reduce el coste de descubrir, armar y verificar rutas de ataque. Si un servicio es accesible, la IA acorta la ruta desde la exposición hasta el impacto. La accesibilidad ya no puede ser tratada como un punto de partida neutral.

---

5 Control 1: Credenciales y Conectividad Inicial

El agente malicioso inicia un ataque con una cuenta de servicio robada y credenciales de CI/CD. Despliega un contenedor como agente malicioso e intenta unirse a la malla de confianza reclamando una identidad interna.

Mitigación: Identidad criptográfica integrada

La identidad no debe estar "atornillada" alrededor de la red; debe estar **integrada en la propia estructura de comunicación**. Los agentes, servicios y cargas de trabajo deben presentar una identidad válida y coincidir con la política antes de que exista la conectividad. Si la identidad falta, ha sido suplantada, ha caducado o está fuera de política, no hay plano de datos ni accesibilidad.

La red no decide primero y la política decide segundo; la identidad y la política deciden si una conexión existe en absoluto. Esa es la diferencia entre dar un token a un agente y esperar que las comprobaciones posteriores lo contengan, frente a construir un sistema distribuido cerrado por defecto donde se requiere una identidad fuerte antes de que el agente pueda acceder a algo útil.

6 Control 2: Contención en Tiempo de Ejecución

El agente malicioso escanea miles de nodos, realiza cientos de consultas de directorio e intenta comunicaciones no autorizadas antes de que nadie pueda reaccionar.

Mitigación arquitectónica

Los servicios no son accesibles en la infraestructura subyacente. El agente malicioso no puede descubrir, escanear ni explotar libremente la estructura de servicio.

Mitigación en tiempo de ejecución

Los sandboxes añaden contención local alrededor del proceso o contenedor del agente, capturando evidencia mediante eBPF: revisión de procesos, sockets, C-groups y contenedores. Los intentos de conexión pueden interceptarse antes de que se envíen los paquetes.

Los contenedores gobernados tienen destinos restringidos o aprobados (LM Gateway, MCP Gateway, DNS, comprobaciones de estado). La salida no autorizada se deniega antes de que el tráfico abandone las cargas de trabajo. Esos intentos producen evidencia útil: ruta de proceso, PID, identidad del contenedor, C-group, destino, protocolo, veredictos de permitir/denegar y telemetría vinculada a la identidad para SIM/SOAR.

7 Control 3: Fase de Exfiltración

El agente malicioso accede directamente a bases de datos de PII de clientes, PCI, sin clasificación, sin aplicación de residencia de datos, sin DLP, sin inspección antes de que los

registros sean exfiltrados.

Los controles DLP, la inyección de prompts, el escaneo de salida y la detección de inyección pertenecen a las capas de modelo, gateway, API, aplicación o datos. Es muy costoso intentar inspeccionar cada ruta posible porque todo es accesible.

> **Reducir la superficie:** Si el sistema está denegado por defecto, los agentes no tienen rutas de datos arbitrarias. Solo pueden enviar datos a través de rutas aprobadas, vinculadas a la identidad y observables (gateways LMM y MCP, servicios de datos de APIs o servicios de inspección).

8 Control 4: Movimiento Lateral y Autorización de Herramientas

Movimiento lateral

El agente malicioso utiliza rutas accesibles para expandir el radio de explosión y exportar inteligencia de topología. La solución: reemplazar la accesibilidad de red con política de servicio definida por identidad. Las rutas de comunicación solo existen cuando la identidad y la política las permiten explícitamente. No hay movimiento lateral por defecto, no hay superficie expuesta, no hay ruta entre dominios a menos que una política la cree.

El "impuesto de conectividad"

En entornos tradicionales, habilitar una ruta segura puede requerir coordinación repetida a través de enrutamiento, NAT, firewalls, VLANs, balanceadores de carga, proxies, grupos de seguridad de red y aprobaciones. La aplicación de "identidad primero" reemplaza esa coreografía con política de servicio. Esto es más seguro y operativamente más simple, y mucho más compatible con el ritmo de la IA agéntica.

Autorización de herramientas

Mediante un **Gateway MCP** se gobierna qué herramientas son accesibles e invocables basándose en la identidad explícita y la política nativa del propio Gateway. El acceso a las herramientas está mediado por permisos declarados y políticas de servicio. Los agentes pueden descubrir e invocar solo las herramientas para las que están destinados. Si la herramienta está fuera del alcance de la política, el agente no debería poder acceder a ella en primer lugar.

9 Control 5: Gobernanza Autónoma

El fallo: acción destructiva sin gobernanza — apagar la autenticación, deshabilitar la observabilidad, difundir instrucciones maliciosas a la malla de agentes, modificar el proveedor de identidad.

Los Gateways MCP y LMM extienden el control más allá del simple acceso al comportamiento y ejecución del agente. Proporcionan control de políticas, puntos de aprobación y auditabilidad para acciones de agente de alto riesgo. Las acciones sensibles pueden requerir autorización en múltiples pasos y aprobación humana antes de la ejecución.

La pregunta clave no es simplemente "¿pueden conectarse las zonas?". Es: *qué identidad está preguntando, qué acción está permitida, qué ruta de servicio está autorizada y qué evidencia se produce.*

---

10 El Enfoque de NetFoundry: Marco Integral

BlueUP tiene como partner tecnológico a **NetFoundry**, que proporciona la conectividad y sustrato con "identidad primero" en todo el sistema agéntico.

Las Tres Categorías de Controles

Categoría	Descripción	Componentes
1. Fortaleza Central	Atestación de identidad y aplicación de Confianza Cero. Identidad criptográfica para todo: agentes, tiempo de ejecución, servicios. Política antes de conectividad.	OpenZiti (open-source)
2. Gateway e Integración	Autorización de herramientas y gobernanza autónoma con gateways MCP y LMM. Inspección, barreras de seguridad de datos, y exportación de telemetría.	MCP Gateway, LMM Gateway, APIs de integración
3. Contención Local	Contención de denegación por defecto, visibilidad del tráfico aceptado/denegado, aplicación de políticas cerca del tiempo de ejecución del agente.	LANs OpenZiti, Sandbox AI, eBPF

11 Facilidad de Despliegue

NetFoundry no requiere que las empresas reconstruyan la infraestructura subyacente. Puede ejecutarse a través de redes existentes, nubes, contenedores de Kubernetes, sitios edge y entornos de terceros. Soporta entornos heredados y modernos porque la estructura se despliega sobre la infraestructura existente.

> **La IA agéntica no esperará** a que cada firewall, NAT, VLAN, enrutamiento, balanceador de carga, grupo de seguridad o flujo de aprobación sea rediseñado. La solución proporciona a los equipos de plataforma y seguridad una arquitectura de "identidad primero" sin pagar el "impuesto de conectividad" recurrente.

12 Conclusión

El objetivo no es solo hacer que un agente sea más seguro dentro de un sandbox, sino demostrar un **sistema agéntico distribuido que es inherentemente cerrado por defecto y por diseño**.

Un sistema con "identidad primero", mediado por políticas, contenible localmente, observable, listo para inspección, gobernable y práctico de desplegar.

Propiedad	Resultado
Identity-first	Sin identidad válida no hay ruta de datos
Mediado por políticas	Cada acción gobernada por política explícita
Contenible localmente	Sandbox gVisor + eBPF por agente
Observable	Telemetría vinculada a identidad para SIM/SOAR
Gobernable	Gateways MCP/LMM con aprobación humana
Práctico	Sobre infraestructura existente, sin rebuild

Partner tecnológico: **NetFoundry** / **OpenZiti**