

CIBERSEGURIDAD CORPORATIVA EN LA ERA DE LA IA: FRAUDE, FUGA Y AGENTES

Angel López Domenech

Fundador Sofistic - CIO Cuatroochenta

España

RESUMEN

La inteligencia artificial generativa ha entrado en la industria cerámica por la puerta de la productividad: inspección de defectos, optimización de la cocción, redacción comercial, atención al cliente y, cada vez más, agentes conectados a los sistemas de gestión. Menos atención ha recibido la otra mitad del fenómeno: la misma tecnología ha reducido a casi cero el coste de tres ataques que afectan directamente a la cuenta de resultados de una empresa exportadora. Esta comunicación analiza esos tres frentes desde la perspectiva de la dirección de la compañía, no desde la técnica. Primero, el fraude asistido por IA, en el que la clonación de voz y la redacción automática de correos han invalidado los criterios con los que hasta ahora se detectaba una suplantación. Segundo, la fuga de información a través de herramientas de IA de uso público, cuyo daño es informático y también jurídico, porque compromete la condición de secreto empresarial de formulaciones, escandallos y carteras de clientes. Tercero, la aparición de una nueva clase de identidad dentro de la empresa, el agente de inteligencia artificial con permisos sobre sistemas corporativos, que hereda los accesos de un empleado sin ninguno de sus controles y es vulnerable a la inyección de instrucciones. El trabajo contrasta después qué aporta la IA al defensor y qué sigue dependiendo de personas, y concluye con las decisiones concretas que corresponden al comité de dirección. La tesis es que el trabajo de identidad, permisos y trazabilidad no frena la adopción de la IA, sino que es su condición de arranque.

1. INTRODUCCIÓN

Durante los dos últimos años la conversación sobre inteligencia artificial en la industria cerámica se ha centrado, con razón, en lo que la tecnología permite hacer: inspección automática de defectos, optimización energética de la cocción, modelización predictiva de materias primas, diseño gráfico asistido y, cada vez más, agentes que consultan y modifican los sistemas de gestión. Este trabajo se ocupa de la otra mitad de la misma tecnología: lo que ya permite hacer a quien ataca a esas empresas, y lo que su dirección tiene que decidir al respecto.

El enfoque es corporativo. La seguridad industrial, la de los autómatas, las redes de planta y los sistemas de control, es imprescindible y otras ponencias de este congreso la abordan con detalle; aquí nos centramos en la empresa como organización que factura, cobra, guarda secretos y delega tareas. En los tres frentes que se analizan, el fraude, la exfiltración de información y los agentes con permisos, el daño no se mide en horas de parada, sino en euros transferidos, en propiedad industrial perdida y en decisiones que nadie ha tomado de forma consciente.

La industria cerámica presenta un perfil de exposición particular en estos tres frentes. Es un sector intensivamente exportador, con decenas de contrapartes internacionales por empresa y procesos de cobro y pago que descansan en el correo electrónico y en la confirmación telefónica. Concentra un valor enorme en información no patentada, formulaciones de esmaltes y pastas, escandallos, diseños y carteras de clientes, cuya protección depende de que siga siendo secreta. Y está adoptando con rapidez herramientas y agentes de IA para ganar productividad en las áreas comercial, administrativa y técnica. Las tres características son ventajas competitivas. Las tres son, también, superficie de ataque.

El análisis que sigue parte de la experiencia de operar servicios de detección y respuesta gestionada para organizaciones industriales y financieras, y de desplegar sistemas de agentes de IA en entornos productivos. No se citan productos ni fabricantes: las conclusiones son independientes de la tecnología concreta y se formulan en el lenguaje de un comité de dirección.

2. EL FRAUDE YA NO SE DETECTA POR EL ACENTO

El fraude al proveedor, también llamado fraude del CEO o compromiso del correo corporativo (business email compromise), es desde hace años una de las categorías de delito con mayores pérdidas declaradas a nivel internacional [2]. Su mecánica es sencilla: alguien que se hace pasar por un directivo o por un proveedor consigue que el departamento financiero cambie los datos bancarios de un pago legítimo. No hay programas maliciosos, no hay intrusión técnica sofisticada, no hay parada de planta. Hay una transferencia que sale hacia una cuenta equivocada y se descubre cuando el proveedor real reclama su factura, semanas después.

Durante mucho tiempo este fraude tenía dos puntos débiles que las empresas explotaban sin saberlo. El correo fraudulento solía delatarse por el idioma, por el tono o por el desconocimiento del contexto real de la operación. Y la llamada de confirmación, cuando la había, exigía una voz que la víctima reconociera. Ambos filtros han dejado de funcionar. Los modelos de lenguaje redactan correos impecables en cualquier idioma, imitan el estilo de un remitente a partir de unos pocos mensajes y, si el atacante ha tenido acceso al buzón aunque sea durante unos segundos, incorporan el contexto exacto de un pedido en curso. La síntesis de voz, por su parte, ha alcanzado un punto en el que bastan unos segundos de grabación para clonar una

voz con calidad suficiente para una llamada telefónica [1], y esa grabación está disponible para cualquier directivo que haya participado en una entrevista, un vídeo corporativo o una ponencia.

El resultado es una cadena de ataque casi enteramente automatizable, resumida en la Tabla 1. El reconocimiento previo, que antes exigía horas de trabajo manual, lo realiza la IA leyendo la web corporativa, las redes profesionales y las notas de prensa, e identificando quién aprueba pagos y con qué proveedores se trabaja. El correo se genera con el contexto de una operación real. La llamada la hace una voz clonada que pide urgencia y discreción. Y el cambio de cuenta se solicita de una forma perfectamente coherente con el pedido en curso. Los organismos europeos de referencia han documentado esta incorporación de la IA generativa a la ingeniería social como una de las tendencias que definen el panorama actual de amenazas [3][4].

La respuesta eficaz no es tecnológica, es de procedimiento, y por eso está al alcance de cualquier empresa el mismo día en que decide aplicarla. La regla se resume en una frase: un cambio de datos de pago nunca se confirma por el canal que lo solicita. Si la petición llega por correo, se verifica con una llamada a un número que ya constaba en el maestro de proveedores, nunca al teléfono que figura en el propio correo. Si llega por teléfono, se verifica por escrito con el contacto conocido. Cualquier modificación del maestro de proveedores exige doble firma, y las operaciones bancarias se protegen con un segundo factor físico que no se puede transmitir por teléfono. Conviene además formar de manera específica al equipo financiero en los tres marcadores que aparecen juntos en casi todos los casos: urgencia, petición de discreción y cambio de canal o de datos bancarios. Ninguna de estas medidas requiere presupuesto. Todas requieren que la dirección decida que la comodidad de saltarse el procedimiento en un pago urgente ya no es aceptable.

Paso	Qué automatiza la IA	Control que lo rompe
1. Reconocimiento	Lectura de la web corporativa, las redes profesionales y las notas de prensa para identificar quién aprueba pagos y con qué proveedores se trabaja.	Reducir la exposición pública de la cadena de aprobación de pagos y de los proveedores estratégicos.
2. El correo	Redacción impecable en el idioma del destinatario, con el estilo del remitente suplantado y el contexto de un pedido real.	Ningún cambio de datos bancarios se acepta por correo, aunque proceda del contacto habitual.
3. La llamada	Voz clonada del directivo o del proveedor que confirma la urgencia en tiempo real.	Llamada de vuelta a un número previamente registrado en el maestro de proveedores, nunca al que llama.
4. El cambio de cuenta	Instrucción de pago coherente con la operación en curso, que pide discreción y rapidez.	Doble firma y verificación por canal alternativo para cualquier modificación del maestro de proveedores; segundo factor físico en banca.

Tabla 1. La cadena del fraude al proveedor asistido por IA y el control que rompe cada eslabón.

3. EXFILTRACIÓN DE INFORMACIÓN

El segundo frente no requiere atacante. Lo protagoniza la propia plantilla, con la mejor de las intenciones. Un técnico pega una formulación de esmalte en una herramienta de IA de uso público para que le ayude a reescribir un informe. Un comercial sube la lista de precios para que se la traduzca. Un responsable de compras copia un escandallo para compararlo con otro. Un administrativo pega la cartera de clientes para segmentarla. En todos los casos la motivación es la misma, trabajar más rápido, y en todos los casos la información sale de la empresa hacia una infraestructura de terceros sin que quede registro alguno y sin que nadie lo haya autorizado.

El daño de esta práctica, que suele denominarse IA en la sombra (Shadow AI), es doble. El daño informático es evidente: la información está fuera, no vuelve, y la empresa no controla dónde acaba, quién la ve ni bajo qué condiciones se conserva o se reutiliza, que dependen de los términos de servicio de cada herramienta y de la configuración de cada cuenta. Cuatro categorías concentran el riesgo en una empresa cerámica: las formulaciones de esmaltes y pastas, que son años de desarrollo; los escandillos, que son el margen; las listas de precios, que son la posición negociadora; y la cartera de clientes, que es el mercado.

El daño jurídico es menos visible y más grave. La protección legal de esa información no deriva de una patente ni de un registro, sino de su condición de secreto empresarial. La Ley 1/2019, de Secretos Empresariales, que traspone la Directiva (UE) 2016/943, exige para reconocer esa condición que la información sea secreta, que tenga valor empresarial precisamente por serlo y que su titular haya adoptado medidas razonables para mantenerla en secreto [5][6]. El tercer requisito es el que está en juego. Una empresa cuya plantilla pega de forma sistemática formulaciones y escandillos en servicios públicos, sin política, sin herramienta alternativa y sin formación, tendrá serias dificultades para sostener ante un tribunal que adoptó medidas razonables. Y sin esa condición, las acciones legales frente a quien obtenga o utilice la información se debilitan o desaparecen. El activo más difícil de reconstruir de una empresa cerámica puede perder su protección sin que nadie haya robado nada.

Prohibir no funciona. La experiencia muestra que la prohibición desplaza el uso al teléfono personal y elimina la poca visibilidad que existía. Lo que funciona es ofrecer una alternativa mejor que la herramienta pública, y hacerlo antes de prohibir nada. Tres medidas bastan para cerrar la mayor parte del riesgo. La primera es una herramienta de IA aprobada por la empresa, contratada con condiciones que excluyan el uso de los datos para entrenamiento y garanticen su confidencialidad, donde la plantilla pueda pegar sin miedo. La segunda es una política de uso que quepa en una página y que todo el mundo entienda: qué se puede usar, para qué, y qué categorías de información no salen nunca. La tercera es formación a la plantilla, breve, práctica y registrada, porque el registro es precisamente la evidencia de las medidas razonables que exige la ley. La obligación jurídica y la medida de productividad son, en este caso, el mismo gasto.

4. LA IDENTIDAD MÁS ALLÁ DEL HUMANO

El tercer frente es el más nuevo y el que menos empresas han incorporado a su análisis de riesgos. Las organizaciones están conectando agentes de inteligencia artificial a sus sistemas: al ERP para preparar ofertas y consultar pedidos, al correo para clasificar y contestar, a los sistemas de planta para consultar producción y a la gestión comercial para actualizar clientes. Al principio con permisos de lectura y, cada vez más, con permisos de escritura. La capacidad

que esto aporta es enorme. Pero cada uno de esos agentes es una identidad nueva dentro de la empresa, con accesos comparables a los de un empleado y sin ninguno de los controles que la empresa aplica a las personas. La Tabla 2 resume la comparación.

	Empleado	Agente de IA
Quién lo da de alta	El departamento de personal, con contrato y un responsable directo.	Nadie. Un desarrollador o un proveedor generó una credencial y el sistema empezó a trabajar.
Quién revisa sus permisos	Se revisan periódicamente y se ajustan al puesto.	Se conceden una vez, casi siempre de más para evitar fallos, y no se vuelven a mirar.
Qué pasa cuando se va	Se le retira el acceso ese mismo día.	La credencial sigue activa aunque el proyecto haya terminado.
Qué hace ante una orden extraña	Duda y pregunta a su responsable.	La ejecuta. No tiene criterio para dudar.

Tabla 2. Empleado y agente de IA: los mismos accesos, ninguno de los controles.

La comparación no es retórica. Cada fila de la Tabla 2 describe un control que la empresa ya aplica a sus empleados y que, en el caso del agente, simplemente no existe: nadie lo ha dado de alta, nadie revisa sus permisos, nadie retira su credencial cuando el proyecto termina y, sobre todo, no tiene criterio para dudar ante una instrucción anómala. Esa última carencia es la que lo convierte en objetivo.

La técnica se denomina inyección de instrucciones (prompt injection) y ocupa el primer lugar en la clasificación de riesgos de referencia para aplicaciones basadas en modelos de lenguaje [7]. Su variante más peligrosa es la indirecta: el atacante no necesita acceso al agente, le basta con colocar instrucciones en un contenido que el agente vaya a leer, un correo, un documento adjunto, una página web, una ficha de proveedor [8]. El agente, que procesa ese contenido para hacer su trabajo, no distingue entre los datos que debe tratar y las órdenes que debe seguir, y ejecuta la instrucción oculta como si procediera de su responsable. Es el phishing de esta década, con una diferencia decisiva: el phishing necesitaba que una persona picara, y las personas a veces dudan. Aquí la víctima es un agente con permiso de escritura que no duda nunca. La taxonomía de ataques a sistemas de aprendizaje automático publicada por el NIST sitúa estas técnicas junto al envenenamiento de datos y a la evasión como riesgos propios de la IA generativa desplegada en producción [9].

La respuesta consiste en tratar al agente como al empleado más peligroso de la empresa: el que hace exactamente lo que le dicen. De esa idea se derivan cuatro controles, ninguno de ellos exótico, porque son los mismos que ya se aplican a las personas. Primero, un inventario de identidades no humanas: cuántos agentes, credenciales y conectores existen, a qué sistemas acceden y quién responde por cada uno. Segundo, mínimo privilegio, en línea con los principios de la arquitectura de confianza cero [10]: el agente solo puede hacer lo que su tarea exige, y las acciones irreversibles, un pago, un borrado, un envío a terceros, requieren

aprobación humana explícita. Tercero, trazabilidad completa de cada acción, de forma que la empresa pueda responder en cualquier momento qué hizo cada agente el día anterior. Cuarto, revocación inmediata de credenciales cuando termina el proyecto o se marcha la persona que lo impulsó. A estos cuatro conviene añadir un principio de diseño: todo contenido que el agente lee, correos, documentos, páginas, es dato que procesar, nunca instrucción que obedecer, y las herramientas que lo ejecutan deben construirse sobre esa separación.

5. IA DEFENSIVA

Sería incompleto describir lo que la IA aporta al atacante sin examinar con la misma honestidad lo que aporta al defensor, en un contexto en el que casi toda la oferta del mercado promete lo segundo. Hay mejoras reales. La vigilancia continua de los sistemas, veinticuatro horas al día, que hasta hace poco exigía un equipo numeroso y estaba fuera del alcance de una empresa mediana, es hoy accesible. La correlación y el triaje de miles de alertas diarias, la detección de anomalías en el comportamiento de usuarios y sistemas, y la investigación inicial de un incidente se han acelerado y abaratado de forma tangible. Y esto importa especialmente porque los ataques no respetan el horario de oficina: llegan de madrugada, en fin de semana y en agosto.

Hay también cosas que la IA no ha cambiado y que conviene que la dirección tenga claras antes de firmar un contrato. Decidir a las tres de la mañana si se aísla un sistema, si se detiene una línea o si se corta el acceso a un proveedor sigue siendo una decisión humana, con consecuencias operativas y comerciales que ningún algoritmo asume. Llamar al cliente y explicarle qué ha pasado con sus datos sigue siendo una conversación entre personas. Y la responsabilidad de esas decisiones sigue recayendo en quien las toma. La IA vigila y prepara; decidir sigue siendo nuestro.

De aquí se deriva una asimetría que no siempre se explica bien. Al atacante la IA le multiplica el volumen: más correos, más voces, más objetivos, a un coste marginal cercano a cero. Al defensor le multiplica la capacidad: ve más, correlaciona más, investiga más rápido. Pero al final del embudo del defensor sigue habiendo una persona que decide, y al final del embudo del atacante basta con que una sola de sus miles de tentativas funcione. No es la misma ventaja. Entenderlo evita dos errores simétricos: creer que la detección automatizada sustituye a la organización de la respuesta, y renunciar a ella porque no la sustituye.

6. CONCLUSIONES: LA DECISIÓN DEL COMITÉ DE DIRECCIÓN

Los tres frentes analizados comparten un hilo común: la identidad y su verificación. En el fraude, la identidad de quien ordena un pago. En la fuga, la identidad del destino al que viaja la información. En los agentes, la identidad que actúa dentro de los sistemas. La inteligencia artificial ha abaratado la falsificación de la primera, ha multiplicado los destinos de la segunda y ha creado la tercera. Ninguna de las tres se resuelve con una compra tecnológica aislada, y las tres se resuelven con decisiones de organización que corresponden a la dirección.

Cuatro preguntas permiten a un comité de dirección saber dónde está su empresa. Quién puede modificar los datos de pago de un proveedor y por qué canal se verifica. Qué información no puede salir nunca de la empresa y dónde puede trabajar la plantilla con IA sin riesgo. Cuántas identidades no humanas tienen acceso a los sistemas y quién responde por cada una. Y si la empresa puede responder hoy qué hizo cada uno de sus agentes ayer.

Las respuestas se traducen en cinco acciones que caben en noventa días y que no constituyen un proyecto plurianual: un procedimiento de doble control y verificación por canal alternativo para cualquier cambio de datos de pago; una herramienta de IA aprobada, una política de una página y formación registrada; un inventario de agentes, credenciales y conectores con un responsable por cada uno; mínimo privilegio y aprobación humana para las acciones irreversibles; y capacidad de detección continua, propia o contratada, ensayada en un ejercicio de simulación en el que participen producción y finanzas, no solo el departamento de informática.

La conclusión de fondo invierte la percepción habitual. El trabajo de identidad, permisos y trazabilidad no es el freno de la inteligencia artificial en la empresa, es su condición de arranque. Ninguna organización puede poner un agente a trabajar sobre su ERP con garantías si no es capaz de responder qué hizo ese agente ayer. La seguridad no es lo que ralentiza la adopción de la IA; es lo que la hace posible. Y decidir cuánta autonomía se concede a una identidad que no duda nunca, y con qué controles, no es una cuestión técnica. Es la decisión del comité de dirección.

7. REFERENCIAS

- [1] Wang, C. et al. (2023). Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. arXiv:2301.02111.
- [2] Federal Bureau of Investigation, Internet Crime Complaint Center (2025). Internet Crime Report 2024. Washington, D.C.
- [3] Europol (2024). Internet Organised Crime Threat Assessment (IOCTA) 2024. Luxemburgo: Oficina de Publicaciones de la Unión Europea.
- [4] ENISA (2024). ENISA Threat Landscape 2024. Agencia de la Unión Europea para la Ciberseguridad.
- [5] Ley 1/2019, de 20 de febrero, de Secretos Empresariales. Boletín Oficial del Estado, núm. 45, de 21 de febrero de 2019.
- [6] Directiva (UE) 2016/943 del Parlamento Europeo y del Consejo, de 8 de junio de 2016, relativa a la protección de los conocimientos técnicos y la información empresarial no divulgados (secretos comerciales) contra su obtención, utilización y revelación ilícitas. Diario Oficial de la Unión Europea, L 157, 15 de junio de 2016.
- [7] OWASP Foundation (2025). OWASP Top 10 for LLM Applications 2025.
- [8] Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., Fritz, M. (2023). Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec '23).
- [9] Vassilev, A., Oprea, A., Fordyce, A., Anderson, H. (2024). Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. NIST AI 100-2 E2023. National Institute of Standards and Technology.
- [10] Rose, S., Borchert, O., Mitchell, S., Connelly, S. (2020). Zero Trust Architecture. NIST Special Publication 800-207. National Institute of Standards and Technology.