

La Externalidad de la Autenticidad de la Información

Cuando la confianza se convierte en una superficie de ataque

Costos de verificación, suplantación institucional y la capa de confianza faltante de la Dimensión 4.5

Luis Ángel Rodríguez Alemán

AliPsé • Master en español v1.0 • Agosto de 2026

Tabla de Contenido

Resumen.....	4
Contribuciones del White Paper	4
1. La revolución digital inconclusa	5
1.1 Dimensión 4.5	5
1.2 Primero llegaron las externalidades positivas	5
1.3 La capa institucional faltante	5
2. Externalidades en la Dimensión 4.5	6
2.1 Una lente económica clásica.....	6
2.2 El carácter dual de la escala digital	6
2.3 Del daño del contenido al costo institucional	6
3. Viejos sesgos humanos, nuevas máquinas de amplificación	6
3.1 La negatividad precede a la tecnología digital	6

3.2 La IA generativa cambia la función de producción del engaño.....	7
3.3 Asimetría de Verificación de la Autenticidad	7
4. Autenticidad no es verdad.....	7
4.1 Una distinción necesaria	7
4.2 Cinco propiedades relacionadas	7
4.3 Superposición con ciberseguridad	8
5. La Externalidad de la Autenticidad de la Información.....	8
5.1 Definición.....	8
5.2 Frontera de la externalidad: la Carga del Incidente no es automáticamente una externalidad.....	8
5.3 Capas de costo	8
5.4 Relación con literatura existente	9
5.5 Economías de Verificación como condición comprobable	9
6. Suplantación institucional y Superficie de Ataque de la Autenticidad	9
6.1 Autoridad prestada	9
6.2 Arquitecturas de Confianza Prestada	9
6.3 El conjunto exploratorio de 43 casos	10
6.4 El caso SEC	10
7. El Modelo de Costos de Verificación.....	10
7.1 Propósito y conservadurismo	10
7.2 Observado, estimado, no observado	10
7.3 Resultados	10
7.4 AER	11
8. Del riesgo de autenticidad a la seguridad epistémica	11
8.1 Las instituciones requieren comunicación atribuida	11
8.2 El liar's dividend	11
8.3 Confianza eficientemente justificada	11
9. Más allá del daño financiero: externalidades políticas, sociales e institucionales	11
9.1 Deepfakes políticos	11
9.2 La confianza como stock	12
9.3 Externalidades políticas como agenda futura	12
10. Respuestas existentes — y la capa faltante.....	12
10.1 Detección	12
10.2 Provenance y Content Credentials	12
10.3 Watermarking	12
10.4 Provenance multicapa.....	13
10.5 Ciberseguridad y Superficie de Ataque de la Autenticidad	13

10.6 El stack emergente de autenticidad.....	13
11. Del derecho a la autenticidad de la información a la infraestructura de autenticidad.....	13
11.1 De norma a institución	13
11.2 Infraestructura de Autenticidad Digital	14
11.3 Autenticar Una Vez, Verificar Muchas.....	14
11.4 Registros de originales autoritativos	14
11.5 El derecho también protege al emisor	14
11.6 El caso SEC revisitado	14
11.7 Subproducción de autenticación	14
12. Completar la Dimensión 4.5.....	15
12.1 De señales de confianza a evidencia de confianza	15
12.2 Orquestación como usabilidad institucional.....	15
12.3 Universalidad y agnosticismo tecnológico	15
12.4 Una infraestructura adversarial.....	15
12.5 Privacidad e inclusión.....	15
12.6 Agentic AI y Autenticidad de la Autoridad Digital.....	15
12.7 La resiliencia requiere más que criptografía	15
12.8 Distribución + Confianza	16
13. Limitaciones y agenda de investigación	16
13.1 Limitaciones del dataset	16
13.2 Evidencia monetaria.....	16
13.3 Carga de verificación modelada	16
13.4 Exposición no es creencia	16
13.5 Costo de Verificación Evitado	16
13.6 Economías de Verificación como proposición comprobable.....	16
13.7 Authenticity Commons.....	16
13.8 Riesgos creados por la propia infraestructura	17
13.9 Falsabilidad	17
13.10 Afirmación apropiada	17
14. Conclusión: completar la capa de confianza	17
Declaración del autor.....	18
Referencias	19

Resumen

Las redes digitales han generado extraordinarias externalidades positivas al reducir el costo de la comunicación, la coordinación, la difusión del conocimiento, la participación y el intercambio económico. Este white paper denomina Dimensión 4.5 a esa capa digital de la vida social y económica: el ciberespacio o espacio digital, superpuesto a la geografía física, en el que la información puede desplazarse globalmente con un costo marginal cercano a cero. La misma arquitectura también genera costos y externalidades negativas cuando no es posible establecer de manera eficiente el origen, la identidad, la integridad o la autoridad de la información. La inteligencia artificial generativa intensifica este problema al abaratar radicalmente la producción de evidencia sintética plausible, mientras los costos de verificarla permanecen distribuidos entre receptores, organizaciones, plataformas y la sociedad.

El trabajo integra literatura sobre externalidades, verificación costosa, economía de la ciberseguridad, desinformación, “provenance” (origen, proveniencia) y autenticidad digital, y propone la Externalidad de la Autenticidad de la Información (EAI) como marco para analizar las cargas que recaen más allá del originador de información engañosa o no autorizada cuando su origen y autoridad no pueden establecerse eficientemente. Desarrolla además un Modelo de Costos de Verificación (VCM, por sus siglas en inglés) y deriva una condición comprobable bajo la cual la autenticación positiva reutilizable podría generar Economías de Verificación. Finalmente, plantea la lógica institucional y técnica de una capa interoperable de verificación y orquestación capaz de volver prácticamente ejercitable el derecho a la autenticidad de la información.

Un conjunto exploratorio de 43 observaciones de suplantación institucional ilustra la emergente Superficie de Ataque de la Autenticidad. El VCM separa exposición bruta a transferencias fraudulentas, pérdida directa neta realizada, Carga del Incidente, carga de autenticidad externalizada y costo neto de bienestar social. Bajo supuestos deliberadamente conservadores, el conjunto contiene aproximadamente US\$27.2 millones de exposición bruta documentada, aproximadamente US\$26.2 millones como piso de pérdida directa neta realizada y alrededor de US\$35,000 de carga modelada de verificación, respuesta, remediación y ecosistema. Estas magnitudes importan por razones distintas: las pérdidas y costos internos pueden justificar privadamente la inversión en autenticidad, mientras que los costos trasladados a clientes, contrapartes, plataformas, autoridades, mercados o sociedad crean un caso institucional y social adicional. Por ello, Carga del Incidente y EAI son lentes complementarios, no medidas rivales.

Palabras clave: autenticidad de la información; externalidades; costos de verificación; suplantación institucional; inteligencia artificial generativa; confianza digital; provenance; infraestructura de autenticidad digital; ciberseguridad; Dimensión 4.5.

Contribuciones del White Paper

- Define la Externalidad de la Autenticidad de la Información (EAI) como la carga impuesta más allá del originador de información engañosa o no autorizada cuando no puede establecerse eficientemente su origen y autoridad, distinguiéndola de los costos internalizados por víctimas u organizaciones que responden al incidente.
- Distingue autenticidad de verdad y separa origen, provenance, integridad, identidad y autoridad como propiedades relacionadas pero no equivalentes.
- Desarrolla una Asimetría de Verificación específica de la autenticidad: el engaño puede producirse y distribuirse a bajo costo, mientras los costos posteriores de autenticación, verificación, respuesta y remediación escalan entre receptores e instituciones.
- Construye un conjunto exploratorio de 43 casos de suplantación institucional y un Modelo de Costos de Verificación consistente con una lectura de bienestar económico.
- Deriva una condición comprobable para posibles Economías de Verificación: por encima de cierto umbral, la autenticación positiva reutilizable puede resultar socialmente más barata que la verificación descentralizada receptor por receptor.
- Enmarca la autenticidad digital como una ampliación de la superficie protegida por ciberseguridad y confianza institucional, no como una categoría competidora.

- Propone una Infraestructura de Autenticidad Digital interoperable en la que captura autenticada, marcado de generación por IA, provenance, watermarking, detección, registros y verificación/orquestación operen como capas complementarias.
- Extiende el derecho a la autenticidad de la información desde una afirmación normativa hacia un problema de diseño institucional: cómo volverlo fácil de ejercer sin crear una autoridad centralizada sobre la verdad.

1. La revolución digital inconclusa

1.1 Dimensión 4.5

Internet no se limitó a hacer más rápida la comunicación existente. Creó una capa operativa adicional de la civilización humana. En este paper, Dimensión 4.5 es el nombre que damos al ciberespacio o espacio digital como una dimensión de la actividad humana superpuesta al mundo físico. El nombre es deliberadamente metafórico; no pretende ser una afirmación de física ni de matemáticas. La actividad digital no está sometida a las tres dimensiones materiales del espacio de la misma forma que los objetos físicos: la información puede estar disponible simultáneamente desde lugares distantes, copiarse sin recorrer la distancia física en el sentido ordinario y coordinar actividad humana a través de la geografía con una fricción marginal extraordinariamente baja. Sin embargo, el ciberespacio no escapa a la cuarta dimensión, el tiempo: los eventos digitales ocurren en secuencia, tienen duración y dejan historia temporal. Por ello, un trabajo anterior llamó a esta capa civilizatoria Dimensión 4.5: una dimensión adicional a las tres espaciales más el tiempo, pero deliberadamente “no del todo” una quinta dimensión física (Rodríguez Alemán 2016).

A lo largo del documento, el término funciona como una abreviatura del dominio digitalmente mediado en el que personas, empresas, gobiernos, mercados, comunidades y, crecientemente, agentes autónomos de software se comunican, transaccionan, coordinan, crean y ejercen autoridad. La Dimensión 4.5 no es completamente física ni está separada de la realidad física; es una capa civilizatoria cuya economía distintiva surge, en parte, porque información, señales de identidad, transacciones y relaciones sociales pueden moverse con una fricción espacial radicalmente menor.

Sus logros económicos y sociales son sustanciales. Los bienes y servicios digitales generan excedentes que las cuentas nacionales tradicionales subestiman; las plataformas reducen costos de búsqueda y coordinación; y las redes permiten formas de colaboración antes prohibitivamente costosas (Brynjolfsson et al. 2019; World Bank 2016). La misma infraestructura ha facilitado colaboración científica, educación, trabajo remoto, emprendimiento, redes de diáspora, movilización política y comunicación durante crisis. Por ello, una agenda de autenticidad no debe partir del pesimismo tecnológico: la Dimensión 4.5 es un logro civilizatorio extraordinario.

1.2 Primero llegaron las externalidades positivas

Muchas externalidades definitorias de las redes digitales son positivas. Quien publica conocimiento útil puede crear valor para receptores muy superior a su propia compensación; cada nuevo participante puede aumentar el valor de una red para los demás; la infraestructura pública digital puede reducir el costo de relacionarse con el Estado; y los estándares abiertos pueden habilitar innovación por actores que no participaron en su creación. Estos beneficios no son periféricos: explican por qué la digitalización se volvió fundacional.

La movilización política ofrece otro ejemplo. Las comunicaciones en red han permitido a ciudadanos coordinarse rápidamente, documentar acontecimientos y superar cuellos de botella informativos que antes favorecían a instituciones centralizadas. Esto no implica que la movilización digital sea siempre democrática o benigna; muestra que la misma arquitectura de baja fricción puede producir externalidades sociales positivas cuando amplía voz, acceso y coordinación.

1.3 La capa institucional faltante

El problema es que la infraestructura para mover información maduró más rápido que la infraestructura para demostrar de dónde proviene. Optimizamos distribución antes de desarrollar mecanismos equivalentes para

origen, autoridad e integridad. Durante años la brecha fue parcialmente tolerable porque fabricar falsificaciones convincentes era costoso, el acceso a distribución masiva era limitado y el contexto social aportaba señales informales de confianza. La IA generativa modifica ese equilibrio.

El siguiente reto institucional no consiste en revertir la Dimensión 4.5, sino en completarla. La pregunta central es cómo preservar sus beneficios de distribución abierta y bajo costo al mismo tiempo que se reduce la incertidumbre sobre quién dijo qué, qué contenido es auténticamente atribuible a una fuente y qué actor estaba autorizado para hablar o actuar en determinado momento.

2. Externalidades en la Dimensión 4.5

2.1 Una lente económica clásica

Una externalidad aparece cuando una decisión impone beneficios o costos sobre terceros que no quedan plenamente reflejados en el cálculo privado de quien toma esa decisión. Pigou (1920) puso el foco en la divergencia entre costos privados y sociales; Coase (1960) mostró que derechos, costos de transacción y estructura institucional determinan cómo se internalizan esos efectos. La economía digital multiplica situaciones en las que una acción de costo muy bajo para un actor produce efectos distribuidos sobre muchos otros.

2.2 El carácter dual de la escala digital

La escala digital amplifica tanto beneficios como daños. El mismo costo marginal cercano a cero que permite que conocimiento, software o coordinación se propaguen rápidamente permite que spam, fraude, suplantación y contenido manipulado alcancen grandes poblaciones. La cuestión no es la escala en sí, sino qué costos y beneficios quedan internalizados y cuáles son desplazados hacia terceros.

Esta dualidad es central para la autenticidad. Un atacante puede generar una pieza de contenido una sola vez; miles de receptores pueden tener que decidir independientemente si es auténtica. Una institución puede publicar una sola advertencia; clientes, periodistas, bancos, plataformas y autoridades pueden invertir tiempo y recursos adicionales en reconstruir qué ocurrió. La arquitectura distribuye el costo de resolver la incertidumbre.

2.3 Del daño del contenido al costo institucional

Buena parte de la discusión pública sobre desinformación se concentra en si un mensaje es falso, persuasivo o dañino. Este paper desplaza parcialmente la pregunta hacia otra dimensión: ¿qué costos aparecen simplemente porque no puede establecerse de manera eficiente si el mensaje proviene de quien aparenta provenir? Aun antes de resolver su verdad sustantiva, la incertidumbre de origen o autoridad puede activar verificación, investigación, contención, comunicación pública, remediación, litigio, prevención futura y pérdida directa.

3. Viejos sesgos humanos, nuevas máquinas de amplificación

3.1 La negatividad precede a la tecnología digital

La preferencia humana por información negativa no nació con las redes sociales. La literatura psicológica documenta desde hace décadas que los eventos negativos suelen tener mayor peso cognitivo y emocional que los positivos (Baumeister et al. 2001; Rozin and Royzman 2001). Las plataformas digitales no inventaron este sesgo, pero pueden amplificarlo porque sus mecanismos de selección y distribución premian contenido que captura atención.

La evidencia contemporánea es consistente con esta interpretación. Robertson et al. (2023), en decenas de miles de experimentos aleatorizados con cientos de millones de impresiones, encontraron que cada palabra negativa adicional en un titular aumentaba aproximadamente 2.3% la tasa de clics. Brady et al. (2017) observaron mayor difusión de lenguaje moral-emocional; Rathje, Van Bavel, and van der Linden (2021) documentaron el fuerte efecto

de animosidad hacia grupos externos sobre engagement. Son viejos sesgos humanos operando dentro de máquinas de amplificación nuevas.

3.2 La IA generativa cambia la función de producción del engaño

La IA generativa no crea el incentivo para engañar, pero altera su función de producción. Voz, imagen, video, documentos y conversaciones plausibles pueden producirse con menor costo, mayor personalización y más velocidad. La literatura sobre persuasión asistida por modelos también muestra que, bajo determinadas condiciones experimentales, sistemas personalizados pueden superar a persuadores humanos (Salvi et al. 2025; Goldstein et al. 2024).

El resultado económico importante no es simplemente “más fake news”. Es una reducción del costo marginal de producir evidencia plausible de identidad, presencia o autoridad. Cuando producir plausibilidad se vuelve barato pero verificarla sigue requiriendo atención humana, controles institucionales o análisis técnico, cambia la distribución de costos.

3.3 Asimetría de Verificación de la Autenticidad

Llamamos Asimetría de Verificación de la Autenticidad a esta distribución: un actor puede producir y distribuir una señal engañosa a bajo costo mientras múltiples receptores deben incurrir individualmente en costos para determinar si la atribución es auténtica. El término debe distinguirse de otros usos de “verification asymmetry” en literatura de IA referidos a diferencias computacionales entre generación y verificación.

La asimetría no requiere que el engaño tenga éxito. Un intento fallido puede generar costos reales de verificación, respuesta y prevención. En consecuencia, medir únicamente pérdidas monetarias realizadas subestima el costo institucional de la incertidumbre de autenticidad.

4. Autenticidad no es verdad

4.1 Una distinción necesaria

Una afirmación puede ser auténtica y falsa: una persona real puede expresar honestamente una creencia equivocada. También puede ser verdadera pero no auténticamente atribuible a quien supuestamente la dijo. La autenticidad responde principalmente a preguntas de origen, identidad, integridad y autoridad; la verdad responde al contenido sustantivo. Confundir ambas categorías conduciría a una infraestructura peligrosa que pretendiera decidir qué debe creerse.

Por ello, una arquitectura de autenticidad no debe convertirse en una autoridad centralizada sobre la verdad. Su función es producir evidencia verificable sobre procedencia y autoridad, dejando la evaluación factual, normativa o política a los procesos e instituciones correspondientes.

4.2 Cinco propiedades relacionadas

Tabla 1. Propiedades de autenticidad y sus preguntas de verificación

Propiedad	Pregunta
Origen	¿De dónde provino este artefacto?
Provenance	¿Qué le ocurrió a lo largo de su ciclo de vida?
Integridad	¿Fue alterado el artefacto o la afirmación relevante?
Identidad	¿Qué persona, organización, dispositivo o agente está asociado?
Autoridad	¿Ese actor estaba autorizado para hablar o actuar en esa capacidad y en ese momento?

Estas propiedades se superponen, pero no son intercambiables. En este paper usamos además tres términos en una secuencia deliberadamente precisa. Autenticación es el establecimiento de un vínculo defendible entre un artefacto, afirmación, identidad o estado de autoridad y la fuente o actor al que se atribuye. Verificación es la comprobación posterior de ese vínculo y de su evidencia en otro punto de decisión. Orquestación es el descubrimiento, combinación, validación y presentación de evidencia heterogénea de autenticidad —por ejemplo

credenciales de provenance, watermarks, registros, señales de identidad o autorización y resultados de detectores— para que la verificación siga siendo utilizable entre tecnologías y canales.

Un archivo criptográficamente íntegro puede contener una afirmación falsa; una voz auténtica puede reproducirse sin autorización; una cuenta oficial real puede ser comprometida y emitir un mensaje institucional no autorizado; y una imagen sintética puede estar plenamente autorizada por su creador. De ahí una proposición central: AI-generated is not the same as unauthorized, and human-generated is not the same as authentic.

4.3 Superposición con ciberseguridad

La autenticidad digital se superpone de manera sustancial con ciberseguridad, pero no es idéntica a ella. La ciberseguridad protege sistemas, identidades, datos, accesos y operaciones; la autenticidad amplía la superficie a la atribución de mensajes, artefactos y actos digitales. Para CISOs y CIOs, esto debe entenderse como una expansión de la superficie que ya protegen, no como una disciplina rival.

5. La Externalidad de la Autenticidad de la Información

5.1 Definición

Proponemos la Externalidad de la Autenticidad de la Información (EAI) como marco para analizar las cargas impuestas más allá del originador de información engañosa o no autorizada cuando no pueden establecerse eficientemente su origen, integridad, identidad o autoridad. La EAI es una parte de un problema económico más amplio. Personas y organizaciones también pueden soportar directamente Cargas del Incidente: transferencias fraudulentas, verificación interna, investigación, remediación, disrupción operativa, efectos reputacionales y prevención incremental. Estos costos privados importan por sí mismos y pueden justificar inversión en controles de autenticidad aun cuando no se mida ninguna externalidad social.

5.2 Frontera de la externalidad: la Carga del Incidente no es automáticamente una externalidad

Para fines de medición distinguimos cuatro incidencias: (1) costo de recursos del atacante; (2) pérdida directa de la víctima y respuesta interna; (3) costo de terceros o del ecosistema soportado por receptores, contrapartes, plataformas, agencias públicas u otros actores; y (4) destrucción o pérdida social más amplia. Las categorías (2) a (4) pueden constituir Cargas del Incidente económicamente relevantes. La categoría (2) es especialmente importante para decisiones privadas de adopción: una empresa no necesita una justificación de externalidad social para invertir en reducir sus propias pérdidas esperadas. Las categorías (3) y (4) revelan el valor institucional o público adicional que puede producir la misma inversión.

5.3 Capas de costo

$$E_A = C_V + C_R + C_O + C_L + C_P + C_F + C_{Rep} + C_E$$

C_V representa verificación; C_R respuesta e investigación institucional; C_O disrupción operativa; C_L recursos legales o regulatorios; C_P comunicación y remediación pública; C_F prevención futura o seguridad incremental; C_Rep costo reputacional; y C_E verificación del ecosistema o receptores. Son categorías operativas, no clasificaciones automáticas de externalidad: su naturaleza económica depende de quién soporta el costo.

$$B_I = L_D + E_A$$

La Carga del Incidente B_I combina la pérdida directa neta realizada L_D con cargas modeladas de verificación, respuesta, remediación y ecosistema. Es una magnitud descriptiva del incidente, no sinónimo de EAI ni de costo neto de bienestar social.

$$W_S = C_A + E_A + D_L$$

El costo neto de bienestar W_S requiere distinguir transferencias de destrucción real de recursos. C_A es costo del atacante y D_L la porción de pérdida que representa destrucción o deadweight loss. Por ello, una transferencia fraudulenta no debe contarse automáticamente como pérdida social neta.

5.4 Relación con literatura existente

La EAI es adyacente, pero no reducible, a varias literaturas establecidas. La economía de ciberseguridad estudia desalineación de incentivos y externalidades de seguridad (Anderson and Moore 2006). Rao and Reiley (2012) muestran cómo abuso informativo de bajo costo puede imponer costos sociales muy superiores al ingreso del atacante en spam. Los modelos de verificación costosa estudian cómo los incentivos de verificación moldean entornos informativos (Kattwinkel and Knoepfle 2023; Cisternas and Vásquez 2026), y la literatura de certificación muestra cómo pequeños costos o errores de verificación pueden destruir excedente bajo asimetría de información (Creane, Jeitschko, and Sim 2022). La contribución aquí es integrar estos mecanismos alrededor de incertidumbre sobre origen y autoridad, medir su incidencia y explorar cómo la autenticación positiva reutilizable puede modificar la economía marginal de verificar.

5.5 Economías de Verificación como condición comprobable

Si cada receptor resuelve independientemente la autenticidad, el costo agregado puede crecer con el número de puntos de decisión. Una autenticación positiva reutilizable podría cambiar esa estructura. Sea C_{Aut} el costo fijo de autenticar un artefacto, afirmación o estado de autoridad; C_{v0} el costo esperado por receptor sin evidencia reutilizable; y C_{v1} el costo residual por receptor mientras la autenticación permanece válida.

$$C_{Aut} + N(C_{v1}) < N(C_{v0})$$

$$N^* = C_{Aut} / (C_{v0} - C_{v1})$$

N^* es el umbral por encima del cual autenticar una vez sería socialmente más barato que reconstruir confianza de forma descentralizada, siempre que $C_{v1} < C_{v0}$ y la autenticación siga vigente. Estas ecuaciones establecen una condición comprobable para posibles Economías de Verificación; no demuestran que ya existan en sistemas desplegados. Deben estimarse empíricamente costos de autenticación, verificación, vigencia, revocación, errores y adopción.

6. Suplantación institucional y Superficie de Ataque de la Autenticidad

6.1 Autoridad prestada

La suplantación institucional explota señales de confianza acumuladas por otros: rostro, voz, cargo, marca, dominio, sello, cuenta oficial, apariencia documental o contexto de canal. El atacante no crea autoridad desde cero; la toma prestada. Esto convierte reputación, identidad y canales de comunicación en activos atacables.

6.2 Arquitecturas de Confianza Prestada

Tabla 2. Arquitecturas de Confianza Prestada

Capa	Señal prestada	Ejemplo
Persona	Rostro, voz, nombre, cargo	CEO, gobernador, ministro, candidato
Institución	Marca, sello, dominio, identidad oficial	Banco, regulador, agencia gubernamental
Canal	Contexto de plataforma o cuenta	Cuenta social oficial, mensajería, videollamada
Medio	Realismo probatorio aparente	Audio, video, PDF, screenshot
Autoridad	Instrucción o endorsement	Orden de pago, anuncio de política, recomendación de inversión

Estas capas pueden combinarse. Un deepfake de un ejecutivo dentro de una videollamada aparentemente normal puede tomar prestadas simultáneamente identidad personal, contexto corporativo, canal y autoridad. Una cuenta gubernamental comprometida puede conservar autenticidad de canal pero perder autenticidad de mensaje.

6.3 El conjunto exploratorio de 43 casos

La base empírica contiene 43 observaciones documentadas de suplantación institucional que abarcan ejecutivos corporativos, autoridades públicas, instituciones financieras, figuras políticas, identidades mediáticas y estafas de inversión. Los casos fueron reunidos deliberadamente a partir de declaraciones oficiales, registros de enforcement, periodismo de alta calidad e investigación de amenazas bajo los criterios documentados en el Apéndice A; no constituyen una muestra de prevalencia. Dieciocho observaciones provienen de una misma investigación de Unit 42 y se tratan como cluster de fuente, no como 18 investigaciones empíricas independientes (Palo Alto Networks Unit 42 2024).

El conjunto no pretende estimar la frecuencia mundial del fenómeno. Su función es mostrar mecanismos recurrentes de confianza prestada, identificar capas de costo y probar si el VCM puede aplicarse de forma consistente a evidencia pública heterogénea.

6.4 El caso SEC

El caso de la U.S. Securities and Exchange Commission (SEC) es particularmente importante porque separa autenticidad de canal de autenticidad de contenido. El 9 de enero de 2024, una parte no autorizada obtuvo control de la cuenta auténtica @SECGov y publicó un anuncio falso sobre aprobación de productos spot Bitcoin ETF. La SEC declaró posteriormente que el contenido no había sido redactado ni creado por la Comisión y que el acceso resultó del control del número telefónico asociado a la cuenta (SEC 2024). La cuenta era real; la comunicación institucional no. Por ello, ciberseguridad de cuenta y autenticidad del contenido son controles complementarios.

7. El Modelo de Costos de Verificación

7.1 Propósito y conservadurismo

El VCM convierte el marco económico en una estructura auditable para medir tanto Carga del Incidente como EAI. No es un estado contable ni pretende capturar el costo social completo de la suplantación. Primero identifica cargas relevantes para víctimas y organizaciones —pérdida directa, verificación, respuesta, remediación y otros costos— y después analiza incidencia: cuáles permanecen internas y cuáles se desplazan hacia receptores, contrapartes, plataformas, instituciones públicas, mercados o sociedad.

7.2 Observado, estimado, no observado

Cada componente se clasifica como Observado (O), Estimado (E) o No observado/no monetizado (U). “Observado” exige evidencia específica del caso; “Estimado” utiliza un benchmark o supuesto explícito; “U” significa que la evidencia pública no permite monetización defendible. Un cero en el modelo mínimo no significa que el costo real haya sido cero.

7.3 Resultados

Tabla 3. Resultados de publicación del VCM

Medida	Valor de publicación v1.0	Interpretación
Piso de exposición bruta	≈ US\$27.2 millones	Capital documentado transferido o expuesto en cuatro casos; no es estimación de welfare loss.
Piso de pérdida directa neta L_D	≈ US\$26.2 millones	Piso de pérdida directa de víctimas después de tratamiento de recuperación.
Carga mínima modelada de verificación/respuesta	≈ US\$35,000	Carga conservadora del incidente; sólo la parte soportada fuera del decision-maker relevante es externalidad económica.
B_I mínimo modelado	≈ US\$26.2 millones	L_D + E_A; mezcla componentes observados y estimados.
W_S	No estimado	C_A y D_L generalmente no observados.
AER, robocall Biden	3.23×	Ratio mínimo exploratorio de un solo caso.

Los resultados monetarios están altamente concentrados. Cuatro casos explican todo el piso cuantificado de exposición y pérdida directa; 36 observaciones no tienen pérdida directa públicamente medible. La inferencia

correcta no es que esos casos causaron cero carga, sino que la evidencia pública no permite monetización defendible. Para organizaciones privadas, pérdida directa y respuesta interna siguen siendo dimensiones relevantes de riesgo e inversión; para policymakers, instituciones de desarrollo y actores del ecosistema, adquiere importancia adicional la incidencia de costos más allá de la víctima.

El workbook subyacente conserva cálculos no redondeados para reproducibilidad; el manuscrito redondea deliberadamente los outputs para no sugerir una precisión epistemológica mayor que la permitida por los supuestos. El registro completo de casos, jerarquía de fuentes, tratamiento de clusters, diccionario de variables, reglas O/E/U, convención FX, tratamiento de pérdidas y trazabilidad se documentan en los Apéndices A y B.

7.4 AER

$$AER = E_A / C_A$$

El Authenticity Externality Ratio (AER) compara la carga externalizada modelada con el costo de recursos del atacante cuando este último puede observarse. En el robocall que imitó la voz de Joe Biden en New Hampshire, el modelo mínimo produce $AER = 3.23$. Es una observación exploratoria, no una estimación poblacional. Su utilidad es mostrar que, al menos en ciertos casos, un gasto pequeño del atacante puede desencadenar costos distribuidos mayores.

8. Del riesgo de autenticidad a la seguridad epistémica

8.1 Las instituciones requieren comunicación atribuida

Las instituciones no funcionan sólo porque la información circule; funcionan porque determinadas comunicaciones pueden atribuirse a actores con autoridad. Órdenes de pago, anuncios regulatorios, instrucciones médicas, decisiones corporativas, comunicados públicos y declaraciones políticas dependen de que los receptores sepan quién habló y en qué capacidad. Cuando esa atribución se vuelve costosa o incierta, aparece una fricción institucional aun antes de discutir la verdad sustantiva.

8.2 El liar's dividend

Los deepfakes no sólo permiten fabricar evidencia; también permiten negar evidencia auténtica. Chesney and Citron (2019) describen este “liar's dividend”: la existencia generalizada de manipulación puede permitir que actores desacrediten material real alegando que es falso. La consecuencia es simétrica: la sociedad necesita detectar falsificaciones, pero también poder autenticar positivamente originales.

8.3 Confianza eficientemente justificada

El objetivo no es maximizar el escepticismo. Una sociedad en la que cada artefacto requiera análisis forense sería disfuncional. Tampoco se busca obediencia epistémica. Autenticación y verificación deben reducir incertidumbre innecesaria sobre origen y autoridad para que los receptores puedan dedicar su atención a evaluar el contenido sustantivo. El objetivo es confianza eficientemente justificada.

9. Más allá del daño financiero: externalidades políticas, sociales e institucionales

9.1 Deepfakes políticos

Los deepfakes políticos pueden imponer costos aun cuando no exista evidencia de que hayan cambiado un resultado electoral. Pueden obligar a campañas, medios, autoridades electorales y ciudadanos a verificar contenido, responder públicamente, documentar falsificaciones o decidir bajo incertidumbre. La literatura reciente aconseja prudencia frente a afirmaciones causales fuertes sobre persuasión o resultados electorales (Ching et al. 2025).

La elección de Eslovaquia de 2023 ilustra la diferencia entre contenido sintético documentado y causalidad electoral. De manera más amplia, un análisis de CETaS/Alan Turing Institute identificó 16 casos virales confirmados de desinformación habilitada por IA durante la elección británica y 11 en elecciones de la UE y Francia, sin encontrar evidencia de que hubieran afectado significativamente los resultados; sí documentó confusión de autenticidad, hostigamiento y riesgos para la integridad de sistemas democráticos (Stockwell 2024).

9.2 La confianza como stock

La confianza institucional puede entenderse como un stock acumulado lentamente y utilizado repetidamente en transacciones y comunicación. La suplantación explota ese stock sin pagar por crearlo. Si los receptores comienzan a exigir verificación adicional de cada comunicación, el costo no aparece únicamente en el incidente original: se manifiesta como fricción futura en todo el sistema.

9.3 Externalidades políticas como agenda futura

La evidencia actual no permite afirmar responsablemente que los deepfakes hayan determinado grandes resultados electorales. Este paper no lo hace. La medición rigurosa de costos políticos, confianza, participación y efectos acumulativos requiere un programa de investigación separado. La ausencia de causalidad electoral demostrada no equivale a ausencia de costos de autenticidad.

10. Respuestas existentes — y la capa faltante

10.1 Detección

La detección de deepfakes es necesaria, pero estructuralmente difícil. Una evaluación y survey cross-modality de 2026 encontró brechas persistentes de generalización out-of-distribution y vulnerabilidad sustancial frente a ataques adversariales white-box en los métodos evaluados (Nguyen-Le et al. 2026). El desempeño en benchmarks controlados no se traduce automáticamente en verificación robusta en el mundo real. La detección es una capa de defensa, no una solución institucional completa.

10.2 Provenance y Content Credentials

C2PA ofrece un estándar técnico abierto para adjuntar Content Credentials criptográficamente firmadas y tamper-evident a activos digitales y expresar afirmaciones verificables sobre su provenance. Puede vincular assertions a un activo, identificar signers, registrar acciones de creación o edición y permitir que sistemas compatibles validen si esas afirmaciones permanecen correctamente asociadas al activo. Es una capacidad fundamental.

Al mismo tiempo, C2PA evita deliberadamente emitir juicios de valor sobre si un provenance es “bueno” o “malo”, y su propia documentación explica que la información de provenance por sí sola no determina si el contenido representado es verdadero, exacto o factual (C2PA 2026a, 2026b). Las credenciales pueden estar ausentes, eliminarse o perder continuidad en workflows no compatibles; y provenance no resuelve por sí mismo toda pregunta de autoridad vigente, delegación revocada, estado de un registro o conflicto entre señales.

Por ello, verificación y orquestación no reemplazan C2PA. Consumen evidencia C2PA cuando existe, la combinan con otras señales cuando es necesario y presentan la evidencia resultante en el punto de decisión. El argumento del paper depende precisamente de tratar provenance como una capa fuerte de un sistema más amplio, no como un straw man.

10.3 Watermarking

Watermarks como SynthID insertan señales destinadas a sobrevivir ciertas transformaciones y ayudar a identificar contenido generado por sistemas participantes. Google DeepMind presenta SynthID como una herramienta complementaria, no como una “silver bullet” universal. Su cobertura depende del generador, del tipo de contenido y de la resistencia de la señal a transformaciones o intentos de remoción (Google DeepMind 2026).

10.4 Provenance multicapa

El enfoque de OpenAI de 2026 hace visible esta arquitectura emergente: conformidad C2PA, watermarking mediante SynthID y herramientas/API de verificación se combinan en lugar de tratarse como técnicas excluyentes. OpenAI describe C2PA como fundamento importante y reconoce que metadata puede perderse o eliminarse; sus herramientas de verificación consultan múltiples señales y evitan conclusiones definitivas cuando ninguna está presente (OpenAI 2026).

10.5 Ciberseguridad y Superficie de Ataque de la Autenticidad

La autenticidad amplía la superficie que la ciberseguridad debe proteger. Account takeover, credential theft, business email compromise, identity compromise y secure communications ya pertenecen al dominio del CISO. Deepfakes, suplantación institucional, mensajes no autorizados desde canales legítimos y abuso de autoridad digital extienden esa misma lógica hacia artefactos y comunicaciones. La respuesta adecuada es integración, no separación.

10.6 El stack emergente de autenticidad

Tabla 4. El stack emergente de autenticidad

Capa	Capacidad representativa	Función
Autenticidad de captura	Cámaras autenticadas / secure capture	Vincular evidencia con origen desde la creación.
Marcado de generación por IA	Disclosure del generador	Marcar contenido producido por sistemas participantes.
Provenance	C2PA / Content Credentials	Preservar assertions firmadas e historial de modificación.
Watermarking	SynthID y equivalentes	Proporcionar señales embebidas durables.
Detección	Detectores de deepfake/manipulación	Estimar patrones sintéticos o de manipulación.
Verificación y orquestación	Capa de verificación cross-signal	Convertir evidencia heterogénea en comprobaciones utilizables en el punto de decisión.

La capa faltante no es “otro detector”. En nuestra taxonomía, autenticación establece el vínculo; verificación comprueba esa evidencia posteriormente; y orquestación descubre, combina, valida y presenta evidencia heterogénea entre estándares, registros, watermarks, sistemas de identidad/autorización y detectores. El gap propuesto es, por tanto, una función de interoperabilidad y usabilidad por encima —no en sustitución— de los controles técnicos subyacentes.

11. Del derecho a la autenticidad de la información a la infraestructura de autenticidad

11.1 De norma a institución

El White Paper No. 1 formuló un argumento normativo: personas e instituciones deberían tener un derecho significativo a distinguir información genuinamente atribuible a una fuente de información falsamente atribuida a ella (Rodríguez Alemán 2026a). Este paper desarrolla un argumento diferente, económico e institucional: la incertidumbre de autenticidad puede generar costos de verificación, respuesta, remediación y ecosistema cuya incidencia se extiende más allá del actor que creó la incertidumbre. Ninguna proposición depende lógicamente de la otra. El derecho podría defenderse aunque las externalidades medidas fueran pequeñas, y una externalidad podría existir sin reconocer jurídicamente ese derecho.

Sin embargo, ambos argumentos convergen en un problema de diseño institucional. El argumento normativo pregunta si personas e instituciones deberían poder establecer atribución auténtica. El argumento económico pregunta si el costo de hacerlo está distribuido ineficientemente o se reconstruye repetidamente. Juntos motivan una pregunta práctica: ¿qué infraestructura puede volver ejercitable el derecho a la autenticidad y reducir carga de verificación sin centralizar autoridad sobre la verdad?

11.2 Infraestructura de Autenticidad Digital

Llamamos Infraestructura de Autenticidad Digital al conjunto interoperable de estándares, mecanismos de captura, credenciales, watermarks, registros, señales de identidad y autoridad, detectores y servicios de verificación/orquestación que permiten establecer y comprobar atribución digital. No se propone una base de datos universal ni una única tecnología. La infraestructura puede ser privada, pública, público-privada o construida sobre estándares abiertos.

11.3 Autenticar Una Vez, Verificar Muchas

El principio de diseño es Autenticar Una Vez, Verificar Muchas. Más precisamente, una fuente legítima establece autenticación para un artefacto, afirmación, identidad o estado de autoridad definido; los receptores pueden verificar repetidamente esa evidencia mientras permanezca válida. Revocación, expiración, modificación, compromiso de credenciales o cambios de autoridad delegada requieren nueva autenticación o evidencia actualizada. El principio trata de reutilizar evidencia válida, no de autenticar permanentemente.

Existe una analogía institucional útil con el principio “once-only” del gobierno digital. OECD (2026) explica que reutilizar información autoritativa requiere registros compartidos, mecanismos de intercambio, salvaguardas y procesos estandarizados de verificación para evitar que cada servicio reconstruya el mismo proceso. Es una analogía de economía institucional y reutilización; no afirmamos que OECD haya diseñado una arquitectura de autenticidad de contenido.

11.4 Registros de originales autoritativos

Una infraestructura madura puede permitir que organizaciones registren originales o assertions autoritativas y que terceros consulten posteriormente evidencia de correspondencia. El valor del registro no reside en declarar verdad, sino en responder una pregunta más acotada: “¿esto coincide con algo que la fuente registró o autorizó?”. Esta lógica puede complementar C2PA, watermarks y canales seguros.

11.5 El derecho también protege al emisor

La autenticidad no protege únicamente al receptor frente al engaño. También protege a la persona o institución frente a atribuciones falsas, alteraciones y comunicaciones no autorizadas. El derecho a la autenticidad incluye, por tanto, la posibilidad de que una fuente pueda decir de forma verificable: “esto sí es nuestro”, no sólo reaccionar después con “eso no fuimos nosotros”.

11.6 El caso SEC revisitado

El caso de la U.S. Securities and Exchange Commission (SEC) adquiere ahora una importancia adicional. La cuenta oficial fue comprometida; por tanto, la autenticidad del canal no bastó. Si cada comunicación institucional relevante contara además con evidencia positiva verificable de origen o autorización, un receptor podría comprobar el mensaje aun cuando el canal hubiese sido tomado. Ésta es la lógica de doble protección: ciberseguridad de la cuenta más autenticidad a nivel de contenido.

11.7 Subproducción de autenticación

Las organizaciones pueden subinvertir en autenticación cuando capturan sólo una parte del beneficio. Una institución evalúa el costo privado de autenticar, pero parte del beneficio puede recaer en clientes, ciudadanos, periodistas, contrapartes, plataformas y mercados que verifican más barato. En esos casos, la autenticación positiva puede generar una externalidad positiva: reduce costos de terceros que el emisor no necesariamente monetiza.

12. Completar la Dimensión 4.5

12.1 De señales de confianza a evidencia de confianza

Durante gran parte de la historia digital hemos confiado en señales: logos, dominios, cuentas verificadas, apariencia visual, familiaridad del canal, voz o rostro. La IA generativa abarata la imitación de muchas de esas señales. La transición necesaria es de señales de confianza a evidencia de confianza. La plausibilidad se vuelve barata; la prueba se vuelve valiosa. Aquí “prueba” significa evidencia verificable que respalda origen, integridad, identidad o autoridad, no prueba de que el contenido sustantivo sea verdadero.

12.2 Orquestación como usabilidad institucional

Incluso tecnologías fuertes de provenance crean valor social limitado si los receptores no pueden acceder, interpretar o combinar sus señales en el momento de decidir. Orquestación se define aquí como el descubrimiento, combinación, validación y presentación de evidencia heterogénea de autenticidad para que la verificación siga siendo utilizable entre tecnologías y canales. Puede incluir Content Credentials, watermarks, registros autoritativos, identidad y autorización, secure capture y outputs de detectores. No es un truth engine: debe hacer utilizable la evidencia sobre origen y autoridad, mostrar conflictos o señales faltantes y preservar incertidumbre cuando la evidencia no justifica una conclusión.

12.3 Universalidad y agnosticismo tecnológico

La infraestructura debe ser suficientemente universal para funcionar entre formatos, sectores, países y proveedores tecnológicos. Una capa de confianza dependiente de un solo generador de IA o plataforma tendría cobertura limitada y riesgo de lock-in. La orquestación debe ser tecnológicamente agnóstica: capaz de consumir evidencia de estándares y mecanismos distintos sin exigir que todos converjan en un único proveedor.

12.4 Una infraestructura adversarial

La autenticidad es un problema adversarial permanente. Cada nuevo mecanismo de provenance, watermarking, registro o detección crea incentivos para evasión, remoción, falsificación o abuso. La infraestructura debe asumir evolución continua, testing, investigación y defensa en profundidad. No existe una solución de una sola vez.

12.5 Privacidad e inclusión

Autenticidad no debe equivaler a identidad universal obligatoria. Muchas comunicaciones legítimas requieren pseudonimato, anonimato o protección de fuentes. El diseño debe autenticar lo necesario para una afirmación o autoridad específica sin convertir la infraestructura en un sistema general de vigilancia. También debe evitar excluir a personas o instituciones con menor capacidad tecnológica.

12.6 Agentic AI y Autenticidad de la Autoridad Digital

La siguiente frontera no será únicamente determinar quién creó un contenido, sino quién o qué está autorizado para actuar por quién, en qué capacidad y durante qué periodo. Los agentes de IA pueden enviar mensajes, negociar, comprar, aprobar o ejecutar acciones. Esto expande la Superficie de Ataque de la Autenticidad desde content authenticity hacia Digital Authority Authenticity.

La pregunta institucional futura será: “Who or what is authorized to act for whom, in what capacity, and at what time?” Esta frontera conecta autenticidad con machine identity, autorización delegada, ciberseguridad y gobernanza de agentes.

12.7 La resiliencia requiere más que criptografía

La criptografía es esencial, pero ninguna infraestructura de largo plazo debe asumir que los algoritmos actuales serán suficientes indefinidamente. La evolución hacia criptografía post-cuántica (PQC), revocación, rotación de credenciales, recuperación y mecanismos de actualización forma parte de la resiliencia. La confianza debe poder migrar sin destruir la continuidad institucional.

12.8 Distribución + Confianza

La primera fase de la revolución digital optimizó distribución. Completar la Dimensión 4.5 exige añadir confianza verificable sin destruir apertura, privacidad ni pluralismo. La ecuación institucional no es menos distribución; es Distribución + Confianza.

13. Limitaciones y agenda de investigación

13.1 Limitaciones del dataset

Los 43 casos son exploratorios, no representativos. Fueron seleccionados por disponibilidad y calidad de evidencia pública; no permiten estimar prevalencia. La concentración de 18 observaciones en una investigación de Unit 42 limita cualquier inferencia que suponga independencia. Observado no equivale a universo real: muchos incidentes no se reportan o carecen de evidencia pública.

13.2 Evidencia monetaria

La evidencia de pérdidas es escasa y concentrada. Sólo cuatro casos sostienen los pisos cuantificados. No es defendible extrapolar mecánicamente esas pérdidas a los 36 casos sin cifra pública. Gross exposure tampoco equivale automáticamente a pérdida neta ni a welfare loss.

13.3 Carga de verificación modelada

La carga modelada de verificación, respuesta, remediación y ecosistema es deliberadamente conservadora. No es un total observado de facturas y no todos sus componentes son automáticamente externalidades. La incidencia debe establecerse caso por caso. Varias capas plausibles —disrupción operativa, recursos legales, prevención futura y reputación— permanecen sin monetizar cuando no existe evidencia específica suficiente.

13.4 Exposición no es creencia

Exposición a un deepfake no implica engagement; engagement no implica creencia; creencia no implica acción; y acción no demuestra causalidad única. Esta cadena es especialmente importante en política. El framework debe evitar transformar visibilidad en impacto sin evidencia.

13.5 Costo de Verificación Evitado

$$AVC = VC0 - VC1$$

Una agenda futura puede medir Avoided Verification Cost (AVC): la diferencia entre costo esperado de verificar sin infraestructura reutilizable y costo después de introducirla. Este indicador permitiría evaluar pilotos y comparar mecanismos de autenticación en contextos reales.

13.6 Economías de Verificación como proposición comprobable

La condición de Economías de Verificación es empíricamente comprobable, no establecida por este dataset. Investigaciones futuras pueden estimar costo fijo de autenticación, tiempo de verificación antes y después, vigencia y revocación, costos de falsos positivos/negativos y tasas de adopción. La hipótesis debe rechazarse donde la autenticación reutilizable no reduzca el costo social esperado o donde C_{v1} no sea menor que C_{v0} .

13.7 Authenticity Commons

Otra línea de investigación es tratar la autenticidad como un commons: evidencia confiable de origen puede producir beneficios difusos que múltiples actores consumen sin agotar el recurso. El término no se reclama como original; se utiliza para explorar cuándo la autenticación positiva tiene características de bien público o infraestructura compartida.

13.8 Riesgos creados por la propia infraestructura

La infraestructura de autenticidad también puede crear riesgos: vigilancia, exclusión, concentración de poder, falsos positivos y negativos, compromiso de credenciales, dependencia de proveedores y abuso de registros. Por ello, la respuesta no debe ser una autoridad central de verdad. Estándares abiertos, interoperabilidad, minimización de datos, contestabilidad y pluralidad institucional son condiciones de diseño.

13.9 Falsabilidad

El framework debe poder fallar. La EAI sería empíricamente menos relevante si los costos externos resultaran triviales frente a los costos privados; las Economías de Verificación no se sostendrían si la evidencia reutilizable no redujera costo marginal; y la infraestructura propuesta perdería justificación si mecanismos existentes resolvieran de manera interoperable y suficientemente usable el problema de origen y autoridad.

13.10 Afirmación apropiada

La afirmación apropiada en esta etapa es limitada pero significativa: existe evidencia teórica y empírica suficiente para tratar la incertidumbre de autenticidad como una fuente potencialmente relevante de Cargas del Incidente y externalidades, justificar su medición sistemática y experimentar con infraestructura que reduzca el costo de establecer origen y autoridad sin sacrificar apertura, privacidad o pluralismo.

La pregunta de investigación que queda abierta es: ¿cuánto cuesta a la sociedad y a las organizaciones la incertidumbre de autenticidad, cómo se distribuyen esos costos y qué instituciones pueden reducirlos más eficientemente sin sacrificar apertura, privacidad o pluralismo?

14. Conclusión: completar la capa de confianza

La Dimensión 4.5 constituye uno de los grandes logros de la civilización contemporánea. Redujo dramáticamente la fricción de mover información, coordinar personas y realizar transacciones. Sus externalidades positivas son reales y extraordinarias. El problema no es que la revolución digital haya fracasado, sino que quedó institucionalmente incompleta.

La economía de la información cambió más rápido que las instituciones que sostienen atribución confiable. La IA generativa abarata la producción de evidencia sintética plausible, mientras autenticación, verificación, respuesta y remediación pueden distribuir cargas entre receptores, organizaciones, plataformas e instituciones públicas. Algunas permanecen privadas y afectan directamente a la persona u organización atacada; otras se desplazan hacia afuera y constituyen la Externalidad de la Autenticidad de la Información. Ambas importan. La Carga del Incidente explica por qué una empresa, banco, autoridad pública u otra organización puede invertir racionalmente en infraestructura de autenticidad para su propio beneficio. La EAI explica por qué la misma inversión puede crear valor adicional para clientes, contrapartes, mercados, instituciones públicas y sociedad.

El conjunto de 43 casos y el VCM no pretenden medir el costo global del problema. Su función es demostrar que la incertidumbre de autenticidad puede observarse como estructura de costos y que una parte puede modelarse de forma reproducible sin confundir transferencias fraudulentas con welfare loss. La siguiente etapa científica es mejorar evidencia, ampliar muestras y medir costos evitados mediante intervenciones reales.

El framework también aclara la frontera arquitectónica. AI-generated is not the same as unauthorized, and human-generated is not the same as authentic. La autenticidad digital amplía la superficie que ciberseguridad y sistemas institucionales de confianza deben proteger, mientras provenance, watermarking, detección, authenticated capture, identidad y autorización siguen siendo controles complementarios.

Esto desplaza el énfasis desde la detección negativa hacia la autenticación positiva. En lugar de preguntar únicamente si algo “parece falso”, las instituciones pueden establecer evidencia de que un artefacto, afirmación o estado de autoridad definido sí les pertenece y permitir que esa evidencia sea verificada repetidamente mientras siga vigente. La hipótesis de Economías de Verificación pregunta si esa reutilización puede reducir suficientemente el costo marginal esperado de verificar como para superar la reconstrucción descentralizada de confianza.

Esta perspectiva dual también aclara los incentivos de adopción. La infraestructura de autenticidad no necesita depender de una internalización altruista de una externalidad social. Una organización privada puede adoptarla porque la reducción esperada de fraude, investigación, remediación, disrupción operativa o daño reputacional supera su costo privado. Cuando la evidencia de confianza resultante también reduce verificación o riesgo para terceros, la misma inversión produce una externalidad positiva adicional. Adopción comercial e infraestructura de interés público pueden reforzarse sin compartir la misma función objetivo.

La arquitectura emergente es, por tanto, un stack: autenticidad de captura; marcado de generación por IA; provenance y Content Credentials; watermarking; detección; registros autoritativos; y una capa de verificación y orquestación que vuelve utilizables señales heterogéneas en el punto de decisión. Ninguna capa sustituye a las demás.

El White Paper No. 1 formuló la autenticidad de la información normativamente como un derecho. Este paper la formula de manera independiente como un problema económico e institucional de incidencia. Los argumentos no se prueban entre sí, pero convergen en el mismo reto de implementación: hacer que la atribución auténtica sea prácticamente ejercitable sin convertir la infraestructura de verificación en un árbitro centralizado de la verdad.

AliPsé representa un esfuerzo emergente por operacionalizar esta necesidad institucional. En lugar de intentar adjudicar la verdad, su papel propuesto es ofrecer una capa tecnológicamente agnóstica de verificación y orquestación mediante la cual puedan registrarse originales auténticos, preservarse evidencia de origen y evaluarse contenidos posteriores contra fuentes autenticadas. En este sentido, AliPsé no se concibe como alternativa a estándares de provenance, watermarking, detección de deepfakes, authenticated capture o controles de ciberseguridad, sino como un posible componente de una Infraestructura de Autenticidad Digital más amplia.

Este desequilibrio no implica que la revolución digital haya fracasado. Implica que permanece incompleta. Construimos la infraestructura para mover información. Ahora estamos comenzando a construir la infraestructura que permite a la información demostrar de dónde provino. Ésa es la capa de confianza faltante.

El desafío institucional pendiente de nuestro tiempo no es revertir la Dimensión 4.5.

Es completarla.

Declaración del autor

El autor es fundador de AliPsé, una iniciativa que desarrolla infraestructura de verificación y orquestación de autenticidad digital. AliPsé se discute en este white paper como una implementación emergente de la arquitectura institucional analizada. Esta relación debe ser considerada al evaluar las implicaciones aplicadas del documento. Las proposiciones teóricas, el dataset empírico, los supuestos del VCM y la evidencia externa se presentan con disclosure metodológico y de fuentes para que puedan evaluarse independientemente del desarrollo comercial de AliPsé.

Referencias

- Anderson, Ross, and Tyler Moore. 2006. “The Economics of Information Security.” *Science* 314 (5799): 610–613. <https://doi.org/10.1126/science.1130992>.
- Banca d’Italia. 2024a. “Warning: Deepfake Videos Circulating Online.” May 22, 2024. <https://www.bancaditalia.it/media/notizia/warning-deepfake-videos-circulating-online/>.
- Banca d’Italia. 2024b. “Warning: Fraudulent Deepfake Videos Online.” July 1, 2024. <https://www.bancaditalia.it/media/notizia/warning-fraudulent-deepfake-videos-online>.
- Banca d’Italia. 2026a. “Warning: Online Circulation of False and Fraudulent Content Unlawfully Using the Name and Image of Fabio Panetta, Governor of Banca d’Italia.” February 26, 2026.
- Banca d’Italia. 2026b. “Warning: Online Circulation of New False and Fraudulent Content Unlawfully Using the Name and Image of Panetta, Governor of Banca d’Italia.” July 8, 2026.
- Baumeister, Roy F., Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs. 2001. “Bad Is Stronger Than Good.” *Review of General Psychology* 5 (4): 323–370. <https://doi.org/10.1037/1089-2680.5.4.323>.
- Brady, William J., Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. Van Bavel. 2017. “Emotion Shapes the Diffusion of Moralized Content in Social Networks.” *Proceedings of the National Academy of Sciences* 114 (28): 7313–7318. <https://doi.org/10.1073/pnas.1618923114>.
- Brynjolfsson, Erik, Avinash Collis, W. Erwin Diewert, Felix Eggers, and Kevin J. Fox. 2019. “GDP-B: Accounting for the Value of New and Free Goods in the Digital Economy.” NBER Working Paper 25695. <https://doi.org/10.3386/w25695>.
- C2PA. 2026a. “C2PA and Content Credentials Explainer.” Coalition for Content Provenance and Authenticity. <https://c2pa.org/specifications/specifications/2.2/explainer/Explainer.html>.
- C2PA. 2026b. “Guiding Principles for C2PA Designs and Specifications.” Coalition for Content Provenance and Authenticity. <https://c2pa.org/principles/>.
- Chesney, Robert, and Danielle Keats Citron. 2019. “Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security.” *California Law Review* 107: 1753–1820.
- Ching, Didier, John Twomey, Matthew P. Aylett, Michael Quayle, Conor Linehan, and Gillian Murphy. 2025. “Can Deepfakes Manipulate Us? Assessing the Evidence via a Critical Scoping Review.” *PLOS ONE* 20 (5): e0320124. <https://doi.org/10.1371/journal.pone.0320124>.
- Cisternas, Gonzalo, and Jorge Vásquez. 2026. “Misinformation in Social Media: The Role of Verification Incentives.” Federal Reserve Bank of New York Staff Report 1028, revised July 2026. https://www.newyorkfed.org/research/staff_reports/sr1028.
- Coase, Ronald H. 1960. “The Problem of Social Cost.” *Journal of Law and Economics* 3: 1–44. <https://doi.org/10.1086/466560>.
- Creane, Anthony, Thomas D. Jeitschko, and Kyoungbo Sim. 2022. “Welfare Effects of Product Certification under Latent Adverse Selection.” *International Journal of Industrial Organization* 81: 102826. <https://doi.org/10.1016/j.ijindorg.2022.102826>.
- Federal Communications Commission. 2024. “FCC Fines Man Behind Election Interference Scheme \$6 Million for Sending Illegal Robocalls That Used Deepfake Generative AI Technology.” September 26, 2024. <https://docs.fcc.gov/public/attachments/DOC-405811A1.pdf>.
- Federal Reserve Board. 2025. “Foreign Exchange Rates—G.5A Annual.” January 6, 2025. <https://www.federalreserve.gov/releases/g5a/20250106/>.
- Goldstein, Josh A., Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. 2024. “How Persuasive Is AI-Generated Propaganda?” *PNAS Nexus* 3 (2): pgae034. <https://doi.org/10.1093/pnasnexus/pgae034>.

- Google DeepMind. 2026. “SynthID: A Tool to Watermark and Identify Content Generated through AI.” <https://deepmind.google/models/synthid/>.
- Hong Kong Government. 2024. “LCQ9: Combating Frauds Involving Deepfake.” June 26, 2024. <https://www.info.gov.hk/gia/general/202406/26/P2024062600192.htm>.
- Kattwinkel, Deniz, and Jan Knoepfle. 2023. “Costless Information and Costly Verification: A Case for Transparency.” *Journal of Political Economy* 131 (2): 504–548. <https://doi.org/10.1086/721618>.
- Leica Camera AG. 2026. “Content Credentials.” <https://leica-camera.com/en-int/photography/content-credentials>.
- Merlino, Luca P., Paolo Pin, and Nicole Tabasso. 2023. “Debunking Rumors in Networks.” *American Economic Journal: Microeconomics* 15 (1): 467–496. <https://doi.org/10.1257/mic.20200403>.
- Nguyen-Le, Hong-Hanh, Van-Tuan Tran, Dinh-Thuc Nguyen, and Nhlen-An Le-Khac. 2026. “Deepfake Detection across Image, Video, and Audio: A Comprehensive Survey with Empirical Evaluation of Generalization and Robustness.” *Artificial Intelligence Review* 59: 203. <https://doi.org/10.1007/s10462-026-11608-4>.
- OECD. 2026. *Digital Government Outlook 2026*. Paris: OECD Publishing.
- OpenAI. 2026. “Advancing Content Provenance for a Safer, More Transparent AI Ecosystem.” May 19; updated July 31, 2026. <https://openai.com/index/advancing-content-provenance/>.
- Palo Alto Networks Unit 42. 2024. “The Emerging Dynamics of Deepfake Scam Campaigns on the Web.” <https://unit42.paloaltonetworks.com/dynamics-of-deepfake-scams/>.
- Pigou, A. C. 1920. *The Economics of Welfare*. London: Macmillan.
- Rao, Justin M., and David H. Reiley. 2012. “The Economics of Spam.” *Journal of Economic Perspectives* 26 (3): 87–110. <https://doi.org/10.1257/jep.26.3.87>.
- Rathje, Steve, Jay J. Van Bavel, and Sander van der Linden. 2021. “Out-group Animosity Drives Engagement on Social Media.” *Proceedings of the National Academy of Sciences* 118 (26): e2024292118. <https://doi.org/10.1073/pnas.2024292118>.
- Robertson, Claire E., Nicolas Pröllochs, Kaoru Schwarzenegger, Philip Pärnamets, Jay J. Van Bavel, and Stefan Feuerriegel. 2023. “Negativity Drives Online News Consumption.” *Nature Human Behaviour* 7: 812–822. <https://doi.org/10.1038/s41562-023-01538-4>.
- Rodríguez Alemán, Luis Ángel. 2016. “La informática en un mundo interdisciplinario.” In *Reflexiones de la Academia Mexicana de Informática a los 40 años de su fundación*. Universidad de Guadalajara. [Final page range to be verified before publication.]
- Rodríguez Alemán, Luis Ángel. 2026a. *The Right to Information Authenticity*. AliPisé White Paper No. 1. English edition.
- Rodríguez Alemán, Luis Ángel. 2026b. *El derecho a la autenticidad de la información*. AliPisé White Paper No. 1. Spanish edition.
- Rozin, Paul, and Edward B. Royzman. 2001. “Negativity Bias, Negativity Dominance, and Contagion.” *Personality and Social Psychology Review* 5 (4): 296–320. https://doi.org/10.1207/S15327957PSPR0504_2.
- Salvi, Francesco, Manoel Horta Ribeiro, Riccardo Gallotti, et al. 2025. “On the Conversational Persuasiveness of GPT-4.” *Nature Human Behaviour* 9: 1645–1653. <https://doi.org/10.1038/s41562-025-02194-6>.
- SEC (U.S. Securities and Exchange Commission). 2024. “Statement on Unauthorized Access to the SEC’s @SECGov X.com Account.” January 12, 2024. <https://www.sec.gov/newsroom/speeches-statements/gensler-x-account>.
- Stockwell, Sam. 2024. *AI-Enabled Influence Operations: Threat Analysis of the 2024 UK and European Elections*. Centre for Emerging Technology and Security, The Alan Turing Institute. September 19, 2024.

<https://cetas.turing.ac.uk/publications/ai-enabled-influence-operations-threat-analysis-2024-uk-and-european-elections>.

Tan, Yiheng, Xiyang Huang, and Wei Li. 2023. “Does Blockchain-Based Traceability System Guarantee Information Authenticity? An Evolutionary Game Approach.” *International Journal of Production Economics* 264: 108974. <https://doi.org/10.1016/j.ijpe.2023.108974>.

Vosoughi, Soroush, Deb Roy, and Sinan Aral. 2018. “The Spread of True and False News Online.” *Science* 359 (6380): 1146–1151. <https://doi.org/10.1126/science.aap9559>.

World Bank. 2016. *World Development Report 2016: Digital Dividends*. Washington, DC: World Bank.