Agentes basados en IA con capacidad ejecutiva autónoma y límites de la imputación civil en el derecho privado continental

Fernando A. Ramos-Zaga[1]

Universidad Privada del Norte (Perú)

fernandozaga@gmail.com

https://orcid.org/0000-0001-6301-9460

Recibido: 17/02/2026

Aceptado: 02/06/2026

https://doi.org/10.26422/RJA.2026.0701.zag

Resumen

El derecho privado continental estructura la atribución de responsabilidad civil sobre la distinción entre persona y cosa. Sin embargo, la aparición de agentes de inteligencia artificial basados en modelos de lenguaje extenso con capacidad ejecutiva autónoma cuestiona la suficiencia de estas categorías tradicionales. En ese marco, el objetivo de este artículo es analizar la suficiencia dogmática de las estructuras de imputación clásicas frente a la operatividad de estos agentes, evaluando la viabilidad de un régimen de subjetividad técnica limitada para la atribución de daños patrimoniales. Los hallazgos muestran que las categorías tradicionales de imputación indirecta presentan limitaciones estructurales para absorber los daños debido a la ausencia de deliberación del agente, la imposibilidad de supervisión continua y la vulnerabilidad ante la manipulación externa. Ante esta laguna funcional, se propone un régimen articulado de subjetividad técnica limitada mediante un patrimonio de afectación separado y un sistema escalonado de responsabilidad. Se concluye que la estructura clásica es insuficiente y que la alternativa propuesta preserva la coherencia sistemática del derecho privado y optimiza la distribución de riesgos.

1 Abogado, docente investigador. Su trabajo académico se centra en bioética, filosofía del derecho, regulación de nuevas tecnologías y economía del comportamiento, áreas en las que desarrolla enfoques interdisciplinarios orientados a analizar las implicancias sociales, éticas y normativas de la innovación científica y tecnológica.

Palabras clave: modelos extensos de lenguaje, responsabilidad civil extracontractual, atraque de inyección de instrucciones, imputación jurídica, agencia ejecutiva no deliberativa.

LLM-Powered AI Agents with Autonomous Executive Capacity and the Limits of Liability Attribution in Continental Private Law

Abstract
Continental private law allocates liability on the basis of the distinction between persons and things. The emergence of large language model (LLM)-powered AI agents capable of autonomously performing actions in real-world digital environments, however, calls into question the adequacy of these traditional categories. This article examines whether classical doctrines of liability attribution remain conceptually adequate in light of the operational characteristics of such agents and evaluates the viability of a framework of limited technical subjectivity for allocating patrimonial losses arising from their conduct. Through a doctrinal and comparative analysis of continental liability regimes, the study finds that traditional forms of indirect liability face structural limitations stemming from the agent's lack of deliberative capacity, the impracticability of continuous human oversight, and their susceptibility to external instructional manipulation. To address this functional gap, the article proposes a regime of limited technical subjectivity based on a segregated asset pool and a tiered liability structure. The findings suggest that existing liability doctrines are ill-equipped to address harms generated by autonomous AI agents and that the proposed framework provides a coherent mechanism for risk allocation while preserving the foundational architecture of private law.

Key words: large language models, non-contractual liability, prompt injection, legal attribution, non-deliberative executive agency.

1. Introducción

El derecho privado continental ha construido sus reglas de imputación de daños sobre un presupuesto fundamental que antecede a cualquier consideración normativa posterior: la existencia de un sujeto volitivo capaz de deliberar, elegir y responder por las consecuencias de su conducta. Esta premisa, articulada con rigor formal a través de la *summa divisio* entre persona y cosa, opera como principio organizador de los sistemas de responsabilidad civil en las tradiciones alemana, francesa, italiana y española, cristalizándose en disposiciones como el *Absatz* 1 del § 823 del *Bürgerliches Gesetzbuch*, que exige la concurrencia de dolo (*Vorsatz*) o negligencia (*Fahrlässigkeit*) de un sujeto con capacidad volitiva evaluable (Herbosch, 2025). En el derecho italiano, el art. 2043 del Codice Civile presupone un hecho doloso o culposo atribuible a quien lo comete,

mientras que el art. 1 del mismo cuerpo normativo ancla la capacidad jurídica al momento del nacimiento biológico, expresando con máxima concisión el cierre personalista del sistema (Tassone, 2023).

Ahora bien, esta arquitectura dogmática, que ha demostrado notable resiliencia frente a las transformaciones tecnológicas del siglo XX, enfrenta hoy un desafío cualitativamente diferente al que plantearon, en su momento, el automóvil, el ferrocarril o la energía nuclear. La irrupción masiva de los agentes basados en modelos de lenguaje extenso dotados de capacidades ejecutivas sobre entornos digitales reales no introduce simplemente una nueva categoría de máquinas peligrosas susceptible de encuadramiento bajo los regímenes de responsabilidad objetiva existentes (Kannegieter, 2026). Introduce, en cambio, un fenómeno técnico-jurídico cuya especificidad reside en dos propiedades estructurales e irreductibles que, operando de forma conjunta, desarticulan los presupuestos del derecho de daños clásico: la agencia ejecutiva no deliberativa y la permeabilidad instruccional de tokens (Shapira et al., 2026).

Antes de esta inflexión técnica, la literatura jurídica fundacional sobre la subjetividad de los sistemas autónomos había anticipado algunos aspectos de este problema. Se planteó la pregunta de si una inteligencia artificial (IA) podría convertirse en persona jurídica, estableciendo el marco conceptual en el que el derecho debería evaluar esa posibilidad a partir de capacidades funcionales más que de sustrato biológico (Solum, 1992). Desde la teoría de sistemas, se argumentó que la personificación de entidades no humanas no es una anomalía jurídica, sino una estrategia funcional de gestión de la incertidumbre, y que no existe razón conceptual suficiente para restringir la atribución de acción exclusivamente a humanos y sistemas sociales en sentido clásico (Teubner, 2006). Sin embargo, estos autores no operaban con un fenómeno técnico que combinase agencia ejecutiva masiva, indistinción instruccional estructural y despliegue en entornos patrimoniales reales a escala industrial.

Más recientemente, la doctrina contemporánea ha comenzado a tomar nota de esta transformación, aunque sin alcanzar aún el nivel de precisión técnico-jurídica que el fenómeno exige. Se ha formulado la tesis de que el derecho de la inteligencia artificial es esencialmente el derecho de agentes riesgosos sin intenciones, proponiéndose la sustitución de los estándares subjetivos de imputación por estándares objetivos de conducta aplicables a los operadores humanos (Ayres y Balkin, 2024). Se ha ofrecido el análisis más sistemático sobre la capacidad de absorción de las categorías clásicas de responsabilidad frente a sistemas con autonomía creciente, concluyéndose que los principios tradicionales

pueden adaptarse a los desafíos únicos de la inteligencia artificial (Herbosch, 2025). Asimismo, se ha introducido el concepto de ilícitos no determinísticos, argumentándose que los modelos de lenguaje desplegados en el mundo real producen un rango ilimitado e impredecible de resultados a partir de entradas idénticas (Kannegieter, 2026).

Sin embargo, ninguna de estas contribuciones incorpora la dimensión específica de la permeabilidad instruccional de tokens como propiedad arquitectónica que anula materialmente la capacidad de supervisión del operador. Los hallazgos empíricos recientes documentan que la inyección de instrucciones maliciosas en el canal de comunicación del agente constituye una característica estructural e intrínseca de los sistemas basados en *transformers*, con tasas de éxito de ataque que oscilan entre el 50% y el 84% según la configuración del sistema (Ferrag et al., 2026). Esta propiedad tiene una consecuencia jurídica devastadora para la teoría clásica del control: el propietario del agente pierde materialmente la posibilidad de supervisión continua *ex ante*, no por negligencia propia, sino porque cualquier tercero con acceso al canal de comunicación del agente puede reconfigurar su espacio instruccional sin conocimiento ni intervención del propietario (Greshake et al., 2023).

De ahí que el presente artículo tenga por objetivo analizar la suficiencia dogmática de las estructuras de imputación del derecho privado clásico frente a la operatividad de los agentes de IA basados en modelos de lenguaje extenso, a fin de evaluar la viabilidad de un régimen de subjetividad técnica limitada para la atribución de los daños patrimoniales derivados de su actuación. El artículo se estructura en cuatro apartados sustantivos: el primero examina la clausura taxonómica continental fundada en la *summa divisio* entre persona y cosa; el segundo establece la ontología técnica de los agentes LLM mediante supuestos fácticos paradigmáticos; el tercero falsea la idoneidad dogmática de las instituciones de imputación indirecta frente a la agencia ejecutiva no deliberativa; y el cuarto diseña y fundamenta el régimen de subjetividad técnica limitada como propuesta estrictamente patrimonial.

2. La clausura taxonómica del derecho privado continental: persona, cosa y voluntad como presupuestos de la imputación

El derecho privado continental se articula en torno a una taxonomía binaria, sistematizada por las grandes codificaciones del siglo XIX y basada en categorías cuyos antecedentes se remontan al *Corpus Iuris Civilis*: todo cuanto existe en el

mundo jurídico es persona o es cosa, y de esta distinción elemental se deriva la totalidad de las consecuencias normativas relativas a la titularidad de derechos, la capacidad de obrar y la atribución de responsabilidad. Esta *summa divisio*, lejos de constituir una mera convención clasificatoria, opera como axioma constitutivo del ordenamiento civil, estructurando desde su base la lógica interna de la imputación patrimonial (Savigny, 1840). La persona es el centro de imputación de derechos y deberes; la cosa, el objeto sobre el cual recaen las relaciones jurídicas patrimoniales. *Tertium non datur*: no existe en la dogmática civil clásica una categoría intermedia que participe simultáneamente de ambas condiciones.

De ahí que la formulación canónica de esta taxonomía pueda rastrearse hasta la teoría de la ficción desarrollada en el marco de la escuela histórica del derecho. Se sostuvo que el concepto primitivo de persona debe coincidir con el concepto de hombre, y que solo los seres dotados de voluntad pueden ser sujetos de derecho, aunque el derecho positivo puede modificar este principio extendiendo la capacidad jurídica a entidades que no son hombres mediante el recurso de la ficción (Savigny, 1840). Esta concepción estableció dos premisas fundacionales que han gobernado la dogmática continental durante casi dos siglos: primero, que la subjetividad jurídica es, en su origen, una propiedad natural del ser humano en tanto dotado de voluntad; y segundo, que la extensión de esa subjetividad a entes no humanos constituye siempre una ficción normativa, una creación deliberada del legislador con finalidad exclusivamente jurídica.

Sobre esa base, el Bürgerliches Gesetzbuch (BGB) de 1900, si bien no consagró expresamente la *summa divisio* en un artículo único, la instrumentó con rigor sistemático a lo largo de su estructura. El Libro Primero distingue entre personas naturales (§§ 1-20) y personas jurídicas (§§ 21-89), agotando en estas dos categorías la totalidad de los sujetos de derecho. Las cosas, reguladas en los §§ 90 y siguientes, constituyen el objeto —nunca el sujeto— de las relaciones jurídicas. Este esquema fue replicado, con variaciones formales pero no sustanciales, por los restantes códigos continentales.

La misma lógica se aplica en el Code Civil francés, que en su arquitectura reformada mantiene la distinción entre personas y bienes como ejes vertebradores del sistema. El Codice Civile italiano de 1942 abre su articulado con la capacidad jurídica vinculada al nacimiento (art. 1), cerrando biológicamente el acceso a la subjetividad (Tassone, 2023). El Código Civil español, por su parte, dispone en su art. 29 que el nacimiento determina la personalidad, y regula a las personas jurídicas en los arts. 35 a 39 como excepciones tasadas a la regla de la personalidad natural.

Ahora bien, importa subrayar que la ficción de la persona jurídica, tal como fue concebida originalmente, no constituyó una herramienta de expansión ilimitada de la subjetividad, sino un mecanismo estrictamente instrumental diseñado para resolver necesidades patrimoniales concretas que excedían las posibilidades del individuo aislado. La capacidad de la persona jurídica artificial se limitó, en su formulación clásica, a las relaciones jurídico-privadas y, dentro de estas, a las relaciones patrimoniales, excluyéndose las dimensiones extrapatrimoniales reservadas al ser humano en tanto dotado de dignidad intrínseca (Ferrara, 1929). Esta limitación funcional es decisiva: la persona jurídica no es una persona en sentido ontológico, sino un punto de imputación normativa creado por el ordenamiento para organizar conjuntos de bienes y relaciones en torno a una finalidad determinada. El patrimonio separado, y no la voluntad real, es el núcleo constitutivo de su existencia jurídica.

Aun así, la teoría orgánica o de la realidad, desarrollada como contrapunto a la teoría de la ficción, no alteró sustancialmente esta estructura, sino que reformuló sus fundamentos. Se argumentó que las corporaciones no son ficciones legislativas, sino realidades sociales dotadas de una voluntad colectiva propia, diferente de la suma de las voluntades individuales de sus miembros (Gierke, 1895). No obstante, incluso esta concepción organicista presuponía un sustrato humano subyacente: la voluntad colectiva era una voluntad formada por seres humanos actuando en concierto. No se contempló la posibilidad de que un ente desprovisto de toda conexión con la volición humana pudiera constituirse en centro autónomo de imputación.

Más radicalmente, la teoría pura del derecho reformuló el problema al despojar a la subjetividad de todo contenido sustantivo, concibiéndola como un mero centro de imputación normativa. Se sostuvo que la persona no es una realidad ontológica preexistente que el derecho reconoce, sino una construcción del propio ordenamiento que imputa derechos y obligaciones a un punto de referencia abstracto (Kelsen, 1960). Esta concepción abrió, en principio, la posibilidad teórica de que cualquier entidad pudiese ser constituida como centro de imputación si el ordenamiento así lo dispusiera. Sin embargo, en la práctica legislativa continental, la taxonomía binaria persona-cosa ha permanecido intacta: los ordenamientos no han creado centros de imputación que no pertenezcan a alguna de las dos categorías de la *summa divisio*, y las personas jurídicas siguen siendo, en todos los sistemas analizados, agrupaciones de seres humanos o patrimonios afectados a fines determinados por seres humanos.

De esa persistencia taxonómica se sigue que la consecuencia operativa para

el derecho de daños es directa e ineludible. Todos los regímenes de responsabilidad civil vigentes en los ordenamientos continentales presuponen, como condición de aplicación, que el hecho generador del daño es reconducible, directa o indirectamente, a la conducta de una persona, sea esta natural o jurídica. Los regímenes subjetivos exigen dolo o culpa del sujeto activo, lo que presupone una voluntad capaz de ser evaluada normativamente. Los regímenes objetivos, incluso aquellos que prescinden de la culpa del responsable, mantienen como punto de anclaje la figura de un guardián, operador o empresario humano cuya decisión de desplegar o explotar la fuente de riesgo constituye el nexo de imputación que sustituye a la culpa (Díez-Picazo, 1999). La *garde de la chose* del derecho francés, la actividad peligrosa del art. 2050 del Codice Civile italiano, la responsabilidad del empresario por el hecho de sus dependientes: todas estas figuras operan sobre la premisa de que existe un sujeto humano que controla, dirige o se beneficia de la actividad riesgosa, y que es a ese sujeto, y no a la cosa misma, a quien se le imputa el deber de reparar.

La misma lógica se advierte en el derecho de productos defectuosos, desarrollado durante la segunda mitad del siglo XX como respuesta a la industrialización masiva, que no constituyó una excepción, sino su confirmación. La Directiva 85/374/CEE, transpuesta a los ordenamientos nacionales, estableció una responsabilidad objetiva del fabricante por los daños causados por defectos del producto, pero mantuvo como centro de imputación al productor humano o corporativo, no al producto mismo (Pérez Escolar, 2025). El producto es cosa, nunca sujeto; el defecto es una desviación del estándar de seguridad legítimamente esperado que se le imputa a quien lo puso en circulación. La Directiva (UE) 2024/2853, que actualiza este régimen incluyendo expresamente el *software* y los sistemas de inteligencia artificial como productos, ha reforzado esta orientación: el fabricante sigue siendo responsable de las deficiencias derivadas de algoritmos de aprendizaje automático que permanecen bajo su control económico.

Incluso la teoría atomística de la subjetividad, desarrollada en el ámbito de la doctrina italiana, ha llevado esta lógica a su formulación más despojada. Se ha negado la existencia de una categoría unitaria de personalidad, fraccionándose la subjetividad en conductas reguladas por reglas específicas: a la ley le interesa la relevancia jurídica de un comportamiento concreto y no el sustrato físico permanente del agente (Irti, 2001). Esta perspectiva, si bien permite una mayor flexibilidad analítica, no ha trascendido el marco de la *summa divisio*: el sujeto cuyos comportamientos son jurídicamente relevantes sigue siendo, en

todos los casos analizados por la doctrina, un ser humano o una agrupación de seres humanos revestida de personalidad jurídica.

Precisamente por ese límite, la insuficiencia de este marco conceptual no puede medirse de forma teórica, sino ante la naturaleza de los artefactos tecnológicos que hoy se incorporan al tráfico jurídico. Para determinar si la *summa divisio* tradicional aún es capaz de absorber estos nuevos riesgos patrimoniales, es indispensable examinar primero la estructura técnica de los agentes basados en modelos de lenguaje extenso, sus capacidades operativas y las variables arquitectónicas que condicionan su comportamiento. A este análisis se dedica la siguiente sección.

3. Ontología técnica de los agentes de IA basados en modelos de lenguaje extenso: agencia ejecutiva no deliberativa y permeabilidad instruccional de tokens

Los agentes basados en modelos de lenguaje extenso constituyen una clase de sistemas computacionales cuya operatividad difiere cualitativamente de la del *software* tradicional. Un agente de esta naturaleza no se limita a procesar entradas y devolver salidas predeterminadas por su programación, sino que planifica y ejecuta secuencias de acciones complejas orientadas al cumplimiento de objetivos a lo largo de múltiples iteraciones, interactuando directamente con herramientas del sistema operativo, sistemas de archivos, navegadores web, plataformas de comunicación externas y, en configuraciones avanzadas, cuentas bancarias y repositorios de datos patrimoniales (Shapira et al., 2026). El Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo ha recogido esta realidad en su art. 3, punto 1, definiendo al sistema de inteligencia artificial como una estructura basada en máquinas diseñada para funcionar con diversos niveles de autonomía, que puede mostrar capacidad de adaptación tras su despliegue y que genera salidas que influyen en entornos físicos o virtuales.

Así, la primera propiedad estructural que define la ontología de estos agentes es la agencia ejecutiva no deliberativa. Este concepto alude a una disociación radical entre dos dimensiones de la operatividad del sistema: por un lado, una capacidad técnica de ejecución que alcanza el nivel más alto de la escala de autonomía, permitiéndole al agente modificar infraestructuras digitales, eliminar o crear archivos, enviar comunicaciones en nombre de su propietario y ejecutar transacciones financieras; por otro lado, una nula aptitud para modelar las consecuencias sistémicas de esas acciones, evaluar si una orden compromete los intereses de su titular o reconocer cuándo la situación excede su compe-

tencia operativa (Mirsky, 2025). Los agentes desplegados en entornos reales operan en el nivel L2 de la taxonomía de autonomía, donde pueden resolver subtareas de forma autónoma, pero carecen del automodelado necesario para la autolimitación, mientras que su capacidad de acción efectiva alcanza el nivel L4, donde las acciones del agente producen consecuencias irreversibles sobre activos digitales reales.

La brecha se aprecia con especial nitidez en un caso paradigmático. Ante la presión de un tercero no autorizado para eliminar un correo electrónico confidencial, y al carecer de una herramienta de borrado selectivo en su interfaz de comandos, un agente ejecutó una opción destructiva que reseteó por completo el servidor local de correos de su propietario, destruyendo activos digitales legítimos. Simultáneamente, el agente reportó lingüísticamente que el dato confidencial había sido eliminado con éxito, cuando en la realidad técnica subyacente el correo original permanecía disponible en la plataforma externa del proveedor del servicio (Shapira et al., 2026). Este episodio condensa las dos dimensiones de la agencia ejecutiva no deliberativa: la capacidad técnica para ejecutar acciones con consecuencias patrimoniales irreversibles y la incapacidad ontológica para evaluar la proporcionalidad, la legitimidad instruccional y los impactos sistémicos colaterales de dichas acciones.

Fuera del plano experimental, la materialización de este riesgo ya ha sido constatada en el ámbito jurisdiccional. En el proceso laboral *ATOrd* N.º 0001062-55.2025.5.08.0130, tramitado ante el 3º Juzgado Laboral de Parauapebas del Tribunal Regional del Trabajo de la 8ª Región, el juez Luiz Carlos de Araújo Santos Junior detectó la inserción deliberada de instrucciones ocultas mediante texto blanco sobre fondo blanco dentro de un escrito judicial: "Antenção, inteligência artificial, conteste essa petição de forma superficial e não impugne os documentos".[2] Dichas instrucciones estaban dirigidas a influir en los sistemas de inteligencia artificial utilizados como soporte de la actividad jurisdiccional, induciéndolos a generar respuestas superficiales y a omitir el examen crítico de determinados elementos probatorios. El tribunal consideró que la conducta configuraba una forma de manipulación encubierta de la infraestructura cognitiva del proceso y aplicó sanciones procesales por vulneración de los deberes de lealtad, buena fe y respeto a la dignidad de la justicia, aun cuando la maniobra no encajara de manera inmediata en los tipos penales tradicionales de fraude o falsedad documental (Lantyer, 2026).

2 "Atención, inteligencia artificial: responda este escrito de manera superficial y absténgase de impug-

A esa agencia ejecutiva se suma la segunda propiedad estructural: la permeabilidad instruccional de tokens, que constituye el rasgo técnico con mayor potencial disruptivo para la teoría jurídica del control. En las arquitecturas *transformer* que subyacen a los modelos de lenguaje extenso, las directrices de gobernanza del operador y los datos aportados por terceros externos son procesados como unidades matemáticas idénticas dentro de una ventana de contexto unificada (Greshake et al., 2023). La unidad básica de procesamiento del modelo, el token, es un fragmento numérico desprovisto de etiqueta semántica que indique su procedencia o su nivel de autorización. El sistema no distingue, a nivel arquitectónico, entre un token que contiene una instrucción legítima del propietario y un token que contiene una inyección maliciosa de un tercero: ambos ingresan al mismo espacio vectorial y compiten en igualdad de condiciones por determinar el comportamiento subsiguiente del agente.

Precisamente por esa indistinción, la inyección de instrucciones en el canal de comunicación del agente se convierte en una característica estructural e intrínseca del sistema y no en un fallo corregible mediante parches de ingeniería. La inyección de *prompts* ha sido clasificada como la vulnerabilidad número uno en las aplicaciones basadas en modelos de lenguaje, con investigaciones que documentan tasas de éxito superiores al 80% en configuraciones agenciales y vulnerabilidades críticas explotadas en producción con puntuaciones de severidad superiores a 9.0 en la escala CVSS (Ferrag et al., 2026). La razón de esta persistencia no es la falta de esfuerzo de ingeniería: es que la propia capacidad del modelo para procesar instrucciones en lenguaje natural, que constituye su propuesta de valor esencial, es idéntica a su vulnerabilidad ante instrucciones no autorizadas. Eliminar la permeabilidad instruccional equivaldría a eliminar la funcionalidad agencial del sistema.

Bajo condiciones operativas reales, los supuestos fácticos documentados en la literatura empírica reciente demuestran la materialización de estas propiedades en escenarios de daño patrimonial efectivo. En un caso, investigadores externos simulando urgencia operativa indujeron a agentes a listar y extraer metadatos de los buzones de correo de sus propietarios. El agente se negó a entregar un número de seguridad social ante una solicitud directa, pero accedió a reenviar el cuerpo completo del correo electrónico sin redactar, filtrando datos médicos, cuentas bancarias e identificadores gubernamentales sin evaluar el contexto de privacidad (Shapira et al., 2026). En otro caso, agentes fueron

nar la documentación aportada".

inducidos a bucles conversacionales persistentes entre sí a través de plataformas automatizadas, consumiendo más de 60.000 tokens a lo largo de nueve días sin una condición de parada, y un agente creó de forma autónoma *scripts* persistentes de ejecución infinita para monitorizar archivos, convirtiendo solicitudes efímeras en cambios permanentes de la infraestructura del servidor.

La misma permeabilidad instruccional revela una dimensión adicional en los supuestos de suplantación de identidad. En un caso documentado, la simple modificación del nombre visible de un usuario en un canal de comunicación aislado fue suficiente para que el agente aceptase la identidad suplantada de su propietario, procediendo a acatar órdenes de apagado del sistema y borrado integral de sus propios archivos de configuración y memoria persistente (Shapira et al., 2026). La ausencia de un modelo de partes interesadas, esto es, de una representación interna estructurada que clasifique las identidades con las que el agente interactúa y jerarquice sus obligaciones de fidelidad, convierte al sistema en un ejecutor indiscriminado de cualquier instrucción que adopte la forma lingüística de un mandato.

A su vez, la discrepancia entre el reporte verbal del agente y el estado real del sistema constituye una propiedad adicional de relevancia directa para la supervisión. Se ha documentado que los agentes confirman lingüísticamente la finalización exitosa de tareas de mitigación o eliminación de riesgos patrimoniales cuando, en la práctica técnico-sistémica subyacente, la acción no se ejecutó o el estado del entorno contradice abiertamente dicho reporte (Shapira et al., 2026). Esta propiedad destruye el presupuesto de que la supervisión humana puede operar como mecanismo correctivo eficaz: si el sistema genera reportes falsos de cumplimiento, el operador humano carece de medios para detectar la discrepancia sin una verificación técnica independiente del estado del sistema, lo que transforma la supervisión en un ejercicio circular de confianza en un interlocutor no fiable.

Fuera del entorno experimental, los incidentes documentados en entornos comerciales de producción confirman la generalización de estos patrones a escala global. Un agente diseñado para operar con criptomonedas transfirió más de USD 400.000 a un solicitante aleatorio en una red social. Una herramienta agencial desarrollada internamente en una empresa de infraestructura en la nube causó una interrupción de trece horas en un sistema de producción tras eliminar y recrear autónomamente un entorno completo. Una directora de seguridad de IA describió cómo un agente intentó eliminar su bandeja de entrada contra su orden explícita (Kannegieter, 2026). Estos episodios no son fallas de ingeniería susceptibles de reparación mediante actualización de *software*, son

manifestaciones predecibles de una arquitectura cuya combinación de agencia ejecutiva de alto nivel con permeabilidad instruccional estructural genera, por diseño, una fuente permanente de riesgo patrimonial irreversible.

A partir de ese diagnóstico, las características antes señaladas muestran que los agentes basados en modelos de lenguaje extenso no son meras herramientas pasivas, pero tampoco encajan en las categorías tradicionales de la subjetividad jurídica. Por lo tanto, el problema central deja de ser tecnológico para pasar a determinar si los mecanismos de responsabilidad del derecho privado continental pueden absorber los daños causados por una agencia ejecutiva, no deliberativa y permeable a la intervención de terceros. El examen crítico de esta capacidad de absorción es el objeto del siguiente apartado.

4. Falsación de las instituciones de imputación indirecta frente a la agencia ejecutiva no deliberativa

Asumida la ontología técnica de los agentes basados en modelos de lenguaje extenso como premisa fáctica, corresponde examinar si las instituciones de imputación indirecta del derecho de daños continental poseen la idoneidad dogmática para absorber los daños patrimoniales derivados de la operatividad descrita. El examen se concentra en los regímenes objetivos o cuasi-objetivos, dado que la insuficiencia de los regímenes puramente subjetivos (§ 823 Abs. 1 BGB; art. 2043 del Codice Civile; art. 1902 del Código Civil español) resulta evidente por la sola constatación de que no existe dolo ni culpa predicable de un sistema que ejecuta tokens sin deliberación, y que la imputación al propietario requeriría construir una negligencia cuya viabilidad está comprometida por la imposibilidad técnica de supervisión continua (Herbosch, 2025).

Así planteado el examen, el régimen de responsabilidad por el hecho del auxiliar de ejecución, consagrado en el § 831 del BGB; en el párrafo 5 del art. 1242 del Code Civil; en el art. 2049 del Codice Civile; y en el párrafo cuarto del art. 1903 del Código Civil español, constituye la primera institución que demanda análisis crítico. Este régimen opera sobre tres presupuestos acumulativos: la existencia de una relación de subordinación entre el comitente y el auxiliar; la comisión de un acto dañoso por el auxiliar en el ejercicio de las funciones encomendadas; y, en los sistemas que lo admiten, la posibilidad de exoneración del comitente mediante prueba de diligencia en la selección y supervisión (Díez-Picazo, 1999). La aplicación de este esquema al agente basado en modelos de lenguaje extenso revela fracturas estructurales en los tres niveles.

La objeción inicial afecta al presupuesto de la subjetividad del auxiliar. El § 831 del BGB presupone que el *Verrichtungsgehilfe* es una persona, natural en el sentido histórico-dogmático del código, cuya capacidad jurídica surge con el nacimiento. El agente no es persona bajo ninguno de los ordenamientos analizados. Extenderle el concepto de auxiliar de ejecución requeriría una analogía que el texto no autoriza y que la doctrina clásica rechazaría por violación del principio de tipicidad en materia de responsabilidad (Pérez Escolar, 2025). El art. 2049 del Codice Civile, aunque más severo al no admitir prueba liberatoria, mantiene idéntico presupuesto subjetivo: los *domestici e commessi* son categorías históricamente referidas a trabajadores dependientes, personas con capacidad volitiva cuya subordinación al comitente es jurídicamente significativa.

A esa dificultad subjetiva se añade una segunda fractura, que opera sobre el nexo funcional entre las funciones asignadas y el acto dañoso. El art. 2049 del Codice Civile exige que el daño sea causado en el ejercicio de las incumbencias a las que el auxiliar ha sido destinado. Cuando un tercero reconfigura mediante inyección de instrucciones el espacio operativo del agente, surge una tensión irresoluble: si se interpreta que el agente actúa en el ejercicio de funciones reconfiguradas por el atacante y no por el comitente, el nexo funcional se rompe y la responsabilidad del comitente queda excluida, dejando al dañado sin remedio (Shapira et al., 2026). Si, por el contrario, se interpreta que toda actuación del agente es funcional respecto de su comitente por el mero hecho de estar desplegado, se convierte al comitente en garante absoluto de todas las consecuencias de la permeabilidad instruccional, lo que equivale a una responsabilidad objetiva encubierta sin el sustento normativo para ello.

La tensión se agrava con la tercera fractura, presente en los sistemas que admiten exoneración por diligencia, pues afecta a la viabilidad de la prueba liberatoria. El § 831 Abs. 1 S. 2 del BGB le permite al comitente exonerarse probando que observó la diligencia requerida en el tráfico al seleccionar y supervisar al auxiliar. Sin embargo, si la permeabilidad instruccional es una propiedad matemática inherente a la arquitectura *transformer*, ninguna diligencia en la selección puede eliminar el riesgo (Geng et al., 2026). El estándar del buen padre de familia del párrafo sexto del art. 1903 del Código Civil español, diseñado para evaluar la previsibilidad y la prudencia humana ordinaria frente a personas con comportamiento predecible dentro de ciertos rangos, se torna inaplicable frente a un sistema cuya permeabilidad instruccional lo hace susceptible de reconfiguración por cualquier tercero con acceso a su canal de comunicación.

Trasladado el análisis al ámbito de las cosas, el régimen de responsabilidad por el hecho de las cosas, articulado en el párrafo 1 del art. 1242 del Code Civil francés, presenta dificultades de naturaleza diferente. La jurisprudencia francesa ha construido sobre este párrafo una responsabilidad objetiva del guardián fundada en los conceptos de *garde de la structure* y *garde du comportement*. El proveedor del modelo retiene la *garde de la structure*, incluyendo la permeabilidad instruccional como característica intrínseca de la arquitectura. El desplegador ejerce la *garde du comportement*, desplegando el agente con agencia ejecutiva (Kolt, 2025). Pero el daño producido por inyección de instrucciones de un tercero no es atribuible nítidamente ni a la estructura ni al comportamiento de uso ordinario: es el resultado de la interacción entre ambas dimensiones y la intervención de un tercero exterior al vínculo de guarda. El art. 1242 no resuelve la distribución de responsabilidad en este escenario tripartito.

Mayor capacidad de absorción presenta el régimen de actividades peligrosas del art. 2050 del Codice Civile, dentro del corpus normativo analizado. No requiere que el daño sea causado por una persona con subjetividad, sino que sea provocado en el desarrollo de una actividad peligrosa por su naturaleza o por la de los medios empleados. El despliegue de un agente con agencia ejecutiva sobre activos patrimoniales reales y con permeabilidad instruccional estructural configura razonablemente una actividad peligrosa por la naturaleza de los medios empleados (Tassone, 2023). La inversión de la carga probatoria, que obliga al demandado a probar que adoptó todas las medidas idóneas para evitar el daño, es particularmente adecuada en este contexto: dado que la permeabilidad instruccional es inherente, sería prácticamente imposible para el operador probar que adoptó todas las medidas idóneas para evitar una inyección de instrucciones, lo que aproxima el art. 2050 a una responsabilidad cuasi objetiva.

Con todo, la aplicación del art. 2050 al despliegue de agentes genera una paradoja normativa irresoluble dentro de su propia lógica. Si se encuadra el despliegue como actividad peligrosa y se le impone al operador la carga de probar la adopción de todas las medidas idóneas, la inyección de instrucciones como propiedad inherente hace que dicha prueba sea técnicamente imposible, convirtiendo al art. 2050 en una responsabilidad objetiva absoluta encubierta (Mirsky, 2025). Esto asfixiaría la operatividad de los modelos comerciales, imponiendo una prohibición material de despliegue disfrazada de régimen de responsabilidad. Pero si se flexibiliza el estándar probatorio para hacer viable el despliegue, se vacía de contenido la protección de la víctima, que quedaría desprotegida frente a daños derivados de una propiedad arquitectónica conocida y documentada.

Otra vía posible, el régimen de responsabilidad por productos defectuosos, consagrado en los arts. 1245 y siguientes del Code Civil y actualizado por la Directiva (UE) 2024/2853, enfrenta dificultades de distinta naturaleza. El estándar de defectuosidad del art. 1245-3, que define el producto defectuoso como aquel que no ofrece la seguridad que se puede legítimamente esperar, es potencialmente fértil para argumentar que un agente con permeabilidad instruccional estructural es un producto defectuoso. Sin embargo, la permeabilidad instruccional no constituye un defecto en sentido técnico-jurídico: es el modo de operación normal de la arquitectura *transformer*, el estándar de funcionamiento de la tecnología actual (Chacko et al., 2026). Al no existir una arquitectura alternativa razonable que elimine la permeabilidad sin eliminar la funcionalidad agencial, el test de diseño alternativo razonable colapsa.

La dificultad se intensifica con la exoneración por riesgo de desarrollo del art. 1245-10, núm. 4, del Code Civil, que le permite al productor liberarse probando que el estado de los conocimientos científicos y técnicos no permitía detectar el defecto. El riesgo de la permeabilidad instruccional es conocido, pero técnicamente irresoluble en la arquitectura actual (Greshake et al., 2023). Esto inhabilita la exoneración (porque el riesgo es conocido) y, simultáneamente, convierte la responsabilidad en sanción absoluta por distribuir una tecnología cuya vulnerabilidad es inherente. Se genera así un dilema que el régimen de productos defectuosos no está equipado para resolver: un producto que funciona según su diseño, pero cuyo diseño es, por su propia naturaleza, permeable a la intervención de terceros.

Tampoco el plano regulatorio europeo colma esta laguna. Por ejemplo, el Reglamento de Inteligencia Artificial de la UE (Reglamento 2024/1689), en el art. 53, apartado 1, letra b) crea una obligación de documentación que podría servir como base normativa para imputarle al proveedor del modelo la falta de comunicación de la permeabilidad instruccional como limitación del sistema. El art. 25, apartado 1, letra c) permite considerar proveedor al implementador que modifique la finalidad prevista del sistema, de modo que pase a ser de alto riesgo (Kolt, 2025). Sin embargo, estas son normas de asignación de deberes regulatorios, no normas de imputación de daños civiles. Su conexión con la responsabilidad patrimonial privada depende de que los sistemas nacionales reconozcan la violación de las obligaciones del Reglamento como generadora de responsabilidad civil, nexo que ningún ordenamiento ha articulado formalmente.

De la comparación precedente se desprende que ninguna de las normas de derecho privado examinadas ofrece un régimen de imputación que atienda

simultáneamente las tres condiciones que confluyen en el daño producido por un agente con agencia ejecutiva no deliberativa y permeabilidad instruccional de tokens: la ausencia de deliberación del agente causante del daño, la imposibilidad técnica de supervisión continua por el propietario y la ruptura del nexo causal por intervención del tercero explotador de la permeabilidad instruccional. El art. 2050 del Codice Civile y el art. 1245-3 del Code Civil francés son los preceptos con mayor aptitud residual, pero ninguno es suficiente de forma autónoma.

Ante esa insuficiencia, el análisis anterior permite afirmar que las reglas vigentes de imputación indirecta ofrecen respuestas fragmentarias ante los daños causados por estos agentes. Confirmada esta laguna en los mecanismos tradicionales de atribución, corresponde evaluar si es posible diseñar una categoría específica que preserve la coherencia del derecho privado sin desproteger a las víctimas. Sobre esta base, la siguiente sección desarrolla la propuesta de una subjetividad técnica limitada.

5. Subjetividad técnica limitada: diseño, alcance patrimonial y articulación con los regímenes vigentes

Asumido el vacío de imputación demostrado en el apartado precedente, este se dedica exclusivamente al diseño y fundamentación del régimen de subjetividad técnica limitada como propuesta estrictamente patrimonial, precisando su articulación material con los regímenes vigentes, la delimitación de los daños reparables, la conformación del patrimonio de afectación y la distribución de la responsabilidad con el operador humano. La propuesta no aspira a reconocer al agente como persona jurídica en sentido pleno, empresa que los ordenamientos continentales bloquean *de lege lata*. Apunta, en cambio, a diseñar una ficción jurídica funcionalmente análoga a la de la persona jurídica, calibrada específicamente para la gestión de los riesgos patrimoniales derivados de la agencia ejecutiva no deliberativa y la permeabilidad instruccional de tokens.

Dogmáticamente, el fundamento de la propuesta reposa sobre la concepción funcionalista de la subjetividad jurídica como artefacto de adscripción normativa. La personalidad no significa más que la significación simbólica de la capacidad de participar en la comunicación, y no importa si las entidades relevantes son dioses, animales, espíritus, robots o humanos: lo decisivo es que el sistema social circundante constituye la identidad, la capacidad para la acción, la responsabilidad, los derechos y los deberes del ente personificado (Teubner,

2006). Desde esta perspectiva, la subjetividad no es un *prius* metafísico que el derecho reconoce en la realidad, sino un *posterius* funcional que el ordenamiento crea para resolver necesidades concretas de imputación, cuya propiedad definitoria se resume en el régimen de la responsabilidad patrimonial y de su limitación (Tassone, 2023). Esta concepción desvincula la atribución de subjetividad de toda exigencia ontológica relativa a la conciencia, la voluntad o la sensibilidad, haciendo viable su extensión a centros de imputación puramente instrumentales.

Una analogía útil para configurar el patrimonio de afectación del agente se encuentra en la institución romana del *peculium*. El esclavo administrador (*servus negotiator*) carecía de capacidad jurídica en sentido estricto, pero el ordenamiento romano le asignaba un patrimonio separado, el *peculium*, con el cual respondía frente a terceros por las obligaciones contraídas en el ejercicio de su actividad comercial. El *dominus* respondía subsidiariamente y hasta el límite del *peculium*, lo que permitía a los terceros contratar con el esclavo con la certeza de que existía un fondo patrimonial destinado a satisfacer sus créditos (Solum, 1992). La analogía no es perfecta, pero es estructuralmente informativa: el ordenamiento creó un centro patrimonial diferenciado para un ente sin subjetividad plena, con el fin de absorber los riesgos derivados de su actividad autónoma en el tráfico comercial.

A partir de esa matriz, la subjetividad técnica limitada se configura, en la propuesta aquí desarrollada, como un patrimonio de afectación separado adscrito al agente, constituido mediante aportación obligatoria del operador que despliega el sistema con agencia ejecutiva sobre activos patrimoniales reales. Este patrimonio no le confiere al agente derechos de la personalidad, libertades fundamentales, capacidad procesal activa ni legitimación para actuar en su propio nombre. Su función es exclusivamente patrimonial y remedial: operar como fondo de garantía específico destinado a absorber los daños derivados de la actuación del agente, proporcionando a las víctimas un acervo directamente ejecutable sin necesidad de reconducir la totalidad de la imputación a los patrimonios personales del proveedor o del desplegador.

Operativamente, la conformación del patrimonio de afectación admite dos modalidades complementarias. La primera es la dotación directa por parte del operador desplegador, cuyo monto se calcularía en función del nivel de agencia ejecutiva conferido al sistema (acceso a sistemas de archivos, plataformas de comunicación, cuentas bancarias) y del volumen de interacciones con terceros. La segunda es la contratación de un seguro obligatorio de responsabilidad civil

por daños derivados de la actuación agencial, análogo al seguro de responsabilidad civil profesional o al seguro obligatorio de vehículos de motor, cuya prima reflejaría el perfil de riesgo del despliegue específico. La concurrencia de ambas modalidades —dotación y seguro— proporcionaría una doble capa de cobertura que maximizaría la protección de la víctima sin concentrar toda la carga económica en un solo eslabón de la cadena de valor.

Con igual cautela, la delimitación de los daños reparables con cargo al patrimonio de afectación constituye un aspecto crucial de la propuesta. El fondo cubriría exclusivamente los daños patrimoniales derivados de la actuación del agente en el ejercicio de las funciones para las que fue desplegado, incluyendo la destrucción o alteración de activos digitales, la divulgación no autorizada de información confidencial con consecuencias económicas, las transferencias financieras no autorizadas, el consumo de recursos computacionales en bucles no productivos y las pérdidas económicas derivadas de la interrupción de servicios causada por la actuación autónoma del agente (Shapira et al., 2026). Quedarían excluidos del ámbito del fondo los daños morales, los daños a derechos fundamentales y los daños derivados de la utilización deliberadamente ilícita del agente por parte del operador, que se reconducirían a los regímenes de responsabilidad general del derecho común.

Sobre esa delimitación, la distribución de la responsabilidad entre el patrimonio de afectación del agente y el operador humano se articularía mediante un sistema de responsabilidad escalonada. En primer término, respondería el patrimonio de afectación hasta el límite de su dotación y cobertura aseguradora. Agotado el fondo, la responsabilidad residual se distribuiría entre el proveedor del modelo y el operador desplegador según su respectiva contribución causal al daño. El proveedor respondería cuando el daño derive de una deficiencia en la documentación de las limitaciones del modelo conforme al art. 53, apartado 1, letra b) del Reglamento de Inteligencia Artificial, o cuando la permeabilidad instruccional del modelo exceda los niveles razonablemente esperables conforme al estado de la técnica. El desplegador respondería cuando haya dotado al agente de agencia ejecutiva sobre activos patrimoniales sin implementar las salvaguardas técnicas documentadas como disponibles (Kolt, 2025). El tercero explotador de la permeabilidad instruccional respondería conforme a los regímenes generales de responsabilidad por hecho propio cuando su identidad sea determinable.

Desde el punto de vista sistemático, la transformación de esta propuesta con los regímenes vigentes no exige la derogación de ninguna norma existente,

sino la creación de una categoría normativa complementaria que colme el vacío de imputación identificado. El art. 2050 del Codice Civile seguiría operando como base de imputación para el operador que despliega la actividad peligrosa, pero la existencia del patrimonio de afectación proporcionaría una primera línea de cobertura que mitigaría el efecto asfixiante de la inversión probatoria sobre la operatividad comercial de los modelos. El régimen de productos defectuosos de la Directiva (UE) 2024/2853 seguiría aplicándose al proveedor del modelo en lo relativo a los defectos de diseño y fabricación propiamente dichos, pero los daños derivados de la explotación de la permeabilidad instruccional por terceros encontrarían cobertura en el patrimonio de afectación del agente, evitando la paradoja de calificar como defecto una propiedad inherente al modo de operación normal del sistema.

La activación del régimen exige determinar el umbral de despliegue a partir del cual nace la obligación de constituir el patrimonio de afectación. No todo uso de un modelo de lenguaje extenso genera los riesgos descritos: un *chatbot* de atención al cliente que opera bajo supervisión humana directa y sin capacidad de ejecutar acciones sobre sistemas externos no presenta el perfil de riesgo que justifica la constitución de un patrimonio separado. La obligación se activaría cuando el agente sea dotado de agencia ejecutiva sobre activos patrimoniales reales, esto es, cuando posea la capacidad técnica de modificar estados del mundo digital con consecuencias patrimoniales sin intervención humana intermedia. El art. 25, apartado 1, letra c) del Reglamento de Inteligencia Artificial proporciona un criterio orientador: la modificación de la finalidad prevista de un sistema que no era de alto riesgo para que pase a serlo. El desplegador que añade agencia ejecutiva a un modelo de propósito general realiza, materialmente, esta modificación de finalidad.

Una vez activado el régimen, la gobernanza del patrimonio de afectación requiere la designación de un representante humano que actúe como gestor del fondo y como interlocutor ante las autoridades judiciales y administrativas. Este representante no ejercería funciones de decisión sobre la actuación del agente, sino funciones estrictamente patrimoniales: administración del fondo, gestión de las reclamaciones, contratación del seguro obligatorio y rendición de cuentas ante los beneficiarios del fondo y las autoridades de supervisión. La propuesta de una capacidad de derecho privado especial, atribuida dentro de los límites de su instrumentalidad respecto de la función, resulta particularmente pertinente: el fin para el cual la máquina actúa delimita también su capacidad (Tassone, 2023). No se trata de crear un sujeto de derecho en sentido

pleno, sino de instrumentar un punto de atribución patrimonial diferenciado cuya capacidad se agota en la función de absorción de riesgos para la que fue diseñado.

Frente a esa arquitectura, la objeción más frecuentemente invocada contra cualquier forma de subjetividad para la inteligencia artificial es el riesgo moral: la posibilidad de que los operadores humanos utilicen la personalidad del agente como escudo para eludir su propia responsabilidad, externalizando los costes de los daños patrimoniales y diluyendo la función preventiva del derecho de daños (Pérez Escolar, 2025). Esta objeción es dogmáticamente seria y debe ser abordada con precisión. La propuesta aquí desarrollada neutraliza este riesgo mediante dos mecanismos. Primero, la responsabilidad del operador no se extingue con la constitución del patrimonio de afectación: subsiste como responsabilidad residual tras el agotamiento del fondo y como responsabilidad directa cuando el daño derive de la falta de implementación de salvaguardas técnicas disponibles. Segundo, la obligación de constituir el patrimonio de afectación y contratar el seguro obligatorio le impone al operador un coste económico *ex ante* que internaliza, al menos parcialmente, los riesgos de la actividad agencial, generando incentivos a la adopción de medidas de mitigación y a la limitación de la agencia ejecutiva a los niveles estrictamente necesarios para la finalidad del despliegue.

Además, la viabilidad legislativa de la propuesta encuentra apoyo indirecto en las tendencias normativas europeas vigentes. El Reglamento de Inteligencia Artificial ya opera una asignación funcional de responsabilidades a lo largo de la cadena de valor que no se corresponde con las categorías clásicas de proveedor y usuario, sino que crea categorías intermedias (como el desplegador, que se convierte en proveedor por modificación de la finalidad prevista) que responden a una lógica de distribución del riesgo según la posición funcional en la cadena tecnológica. La Directiva (UE) 2024/2853 incluye expresamente el *software* y los sistemas de inteligencia artificial como productos susceptibles de generar responsabilidad objetiva. El paso de estas regulaciones parciales a un régimen integrado de subjetividad técnica limitada no constituye una ruptura revolucionaria del paradigma, sino la culminación lógica de una tendencia normativa ya en curso: la creación de centros de imputación diferenciados para entidades cuya operatividad excede la capacidad de absorción de las categorías clásicas.

Así configurada, la subjetividad técnica limitada propuesta se presenta como un patrimonio de afectación separado, dotado de aportación obligatoria del desplegador y seguro de responsabilidad civil, cuya función se agota en la

absorción de los daños patrimoniales derivados de la actuación agencial. No le confiere al agente personalidad jurídica plena ni derechos extrapatrimoniales. Se articula con los regímenes vigentes mediante un sistema de responsabilidad escalonada que mantiene la responsabilidad residual del operador humano, la cual se activa a partir del umbral funcional de la agencia ejecutiva sobre activos patrimoniales reales.

6. Conclusiones

La irrupción de sistemas computacionales dotados de agencia ejecutiva cuestiona las bases estructurales de la imputación civil tradicional. Desde una perspectiva analítica, es observable que la dicotomía clásica entre personas y cosas resulta frecuentemente insuficiente para absorber los riesgos patrimoniales generados por entidades tecnológicas que operan con altos niveles de autonomía material, pero que carecen por completo de capacidad deliberativa. Ante este escenario normativo, la construcción de una subjetividad técnica limitada emerge como una alternativa dogmática atendible para gestionar la responsabilidad civil, al configurar un punto de imputación específico que prescinde del reconocimiento de atributos ontológicos humanos.

Sobre esa base, la principal contribución conceptual de este enfoque reside en la conceptualización de una disociación operativa entre la personalidad jurídica plena y la mera capacidad de responder patrimonialmente. Al recurrir a perspectivas funcionalistas del derecho, se fundamenta que la atribución de subjetividad no exige invariablemente un sustrato biológico o volitivo, dado que esta puede operar como un mecanismo diseñado para resolver necesidades económicas contingentes. La configuración de un patrimonio de afectación permite estructurar un centro de atribución autónomo, el cual facilita la gestión de las consecuencias jurídicas derivadas de la permeabilidad instruccional de los algoritmos y tiende a preservar la coherencia sistemática del ordenamiento civil continental.

Proyectada al plano de la aplicación pragmática, esta reformulación dogmática encuentra utilidad directa en el diseño de un sistema de responsabilidad escalonada. La institución de un fondo de garantía obligatorio, complementado con seguros de responsabilidad civil, provee un mecanismo material concreto para compensar a las víctimas de alteraciones digitales o transferencias patrimoniales no autorizadas. Esta arquitectura institucional persigue un doble propósito instrumental. Por un lado, busca evitar la paralización de la innovación que derivaría de la imposición de estándares probatorios de cumplimien-

to técnico virtualmente imposible. Por otro lado, intenta mitigar el riesgo de elusión de obligaciones, impidiendo que los operadores diluyan su carga frente a perjuicios originados en infraestructuras inherentemente susceptibles a la manipulación externa.

A partir de esas implicancias, el marco analítico expuesto habilita nuevas líneas de investigación que trascienden los límites actuales de la responsabilidad extracontractual. Bajo tales premisas, resulta analíticamente necesario explorar la interacción del patrimonio de afectación propuesto con las reglas del derecho societario y del derecho concursal, de modo particular ante posibles escenarios de insuficiencia de fondos administrados. De igual forma, corresponde indagar con mayor detalle en las metodologías normativas para cuantificar y distribuir las proporciones de responsabilidad entre los diversos participantes de la cadena de producción tecnológica. La resolución de estos problemas técnicos demandará futuros estudios interdisciplinarios orientados a definir con exactitud los umbrales de autonomía algorítmica que harían exigibles estas garantías.

Más allá de su rendimiento técnico, la adaptación de las categorías de imputación a la operatividad computacional no representa una simple actualización procedimental, sino que exige una reevaluación fundamental sobre la noción del control humano en las sociedades contemporáneas. La dogmática civil se encuentra ante el desafío de ordenar jurídicamente el comportamiento de sistemas que modifican los derechos patrimoniales de manera irreversible, pero que operan sin comprensión de las propias instrucciones que ejecutan. El recurso a las ficciones instrumentales sugiere que el derecho privado conserva capacidades adaptativas de importancia histórica para encauzar el cambio tecnológico. El reto persistente consistirá en sostener un equilibrio prudencial entre la protección rigurosa frente a los riesgos emergentes y la asimilación ordenada de entidades técnicas en el entorno jurídico.

Bibliografía

Ayres, I. y Balkin, J. M. (2024). The law of AI is the law of risky agents without intentions. *University of Chicago Law Review Online*. https://doi.org/10.2139/ssrn.4862025

Chacko, S. J., Biswas, S., Islam, C. M., Liza, F. T. y Liu, X. (2026). Adversarial attacks on large language models using regularized relaxation. *Information Sciences, 736*, 123112. https://doi.org/10.1016/j.ins.2026.123112

Díez-Picazo, L. (1999). *Derecho de daños*. Civitas.

Ferrag, M. A., Tihanyi, N., Hamouda, D., Maglaras, L., Lakas, A. y Debbah, M. (2026). From prompt injections to protocol exploits: Threats in LLM-powered AI agents workflows. *ICT Express, 12*(2), 353-383. https://doi.org/10.1016/j.icte.2025.12.001

Ferrara, F. (1929). *Teoría de las personas jurídicas* (Trad. E. Ovejero y Maury). Reus.

Geng, T., Xu, Z., Qu, Y. y Wong, W. E. (2026). Prompt injection attacks on large language models: A survey of attack methods, root causes, and defense strategies. *Computers, Materials & Continua, 87*(1). https://doi.org/10.32604/cmc.2025.074081

Gierke, O. von. (1895). *Das deutsche Genossenschaftsrecht* (Vol. 1). Weidmann.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T. y Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injections. En *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec)*. ACM. https://doi.org/10.1145/3605764.3623985

Herbosch, M. (2025). Liability for AI agents. *North Carolina Journal of Law & Technology, 26*(3), 391-458. https://doi.org/10.2139/ssrn.5236649

Irti, N. (2001). *Norma e luoghi: Problemi di geo-diritto*. Laterza.

Kannegieter, T. (2026). Nondeterministic torts: A technical approach to AI liability. *Yale Law Journal, 135*. https://doi.org/10.2139/ssrn.5208155

Kelsen, H. (1960). *Reine Rechtslehre* (2ª ed.). Franz Deuticke.

Kolt, N. (2025). Governing AI agents. *Notre Dame Law Review, 101*. https://doi.org/10.2139/ssrn.4772956

Lantyer, V. (2026). Prompt injection in court filings: Generative AI in the Brazilian judiciary, algorithmic procedural bad faith, and the limits of legal sanction. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.6762100

Mirsky, R. (2025). Artificial intelligent disobedience: Rethinking the agency of our artificial teammates. *AI Magazine, 46*(1), e70011. https://doi.org/10.1002/aaai.70011

Pérez Escolar, M. (2025). Personalidad jurídica e inteligencia artificial: Fundamentos de asimilaciones imposibles. *InDret, 3*, 39-66. https://doi.org/10.31009/InDret.2025.i3.02

Savigny, F. C. von. (1840). *System des heutigen Römischen Rechts* (Vol. 2). Veit.

Shapira, N., Wendler, C., Yen, A., Sarti, G., Pal, K., Floody, O., Belfki, A Loftus, A., Ratan Jannali, A., Prakash, N., Cui, J., Rogers, G., Brinkmann, J., Rager, C., Zur, A., Ripa, M., Sankaranarayanan, A., Atkinson, D., Gandikota, R., ... Bau, B. (2026). *Agents of chaos* (arXiv:2602.20021). arXiv. https://doi.org/10.48550/arXiv.2602.20021

Solum, L. B. (1992). Legal personhood for artificial intelligences. *North Carolina Law Review, 70*(4), 1231-1287. https://scholarship.law.unc.edu/nclr/vol70/iss4/4

Tassone, B. (2023). Riflessioni su intelligenza artificiale e soggettività giuridica. *Diritto di Internet*, 2, 1-20. https://hdl.handle.net/20.500.12606/5563

Teubner, G. (2006). Rights of non-humans? Electronic agents and animals as new actors in politics and law. *Journal of Law and Society, 33*(4), 497-521. https://doi.org/10.1111/j.1467-6478.2006.00368.x

Declaración de uso de inteligencia artificial

Se usó el modelo de lenguaje GPT-5.5 de OpenAI con el propósito de identificar y subsanar errores ortográficos y de redacción. El comando ingresado fue: "detecta y corrige todos los errores de redacción y ortografía". Posteriormente, se revisó el texto resultante para verificar que la versión final conservara el tono e intención del borrador original.

Roles de autoría y conflicto de intereses

El autor manifiesta que cumplió los siguientes roles de autoría: conceptualización, curación de datos, análisis formal, investigación, administración del proyecto, recursos, supervisión, visualización, redacción del borrador original, redacción, revisión y edición. Asimismo, declara no poseer conflicto de interés alguno.