EL CASO MATA V. AVIANCA Y EL DESAFÍO DE LA VERACIDAD EN LA IA LEGAL

Ricardo Scarpa (derechoartificial.com)

Conforme a normas APA 7ª edición
Análisis técnico, deontológico y regulatorio

CAPÍTULO I: INTRODUCCIÓN Y PROPÓSITO DEL INFORME

1.1. Contexto y Justificación: Evolución Acelerada de los LLM en la Práctica Jurídica

El entorno legal global experimenta una transformación tecnológica sin precedentes. Los Grandes Modelos de Lenguaje (LLM, por sus siglas en inglés), particularmente aquellos orientados a tareas de generación de contenido jurídico, representan herramientas que prometen automatización de investigación, drafting asistido y análisis documental a escala masiva. Sin embargo, esta promesa se acompaña de riesgos técnicos y deontológicos de magnitud considerable.

La Opinión Formal 512 de la American Bar Association (ABA), emitida el 29 de julio de 2024, marca un hito normativo al ser la "primera guía ética formal" de una institución estadounidense sobre uso de herramientas de IA generativa en práctica jurídica. Esta opinión sitúa la competencia tecnológica y la supervisión profesional como derechos fundamentales para la diligencia debida del letrado.[1]

Paralelamente, marcos técnicos como el estándar ISO/IEC 42001 (abordado en el White Paper técnico de SGS sobre Gobernanza de IA en industria legal) y el Marco de Gestión de Riesgos de Inteligencia Artificial del NIST (AI RMF 1.0, versión de julio de 2024) han

establecido metodologías para clasificación de riesgos, mitigación técnica y gobernanza de ciclo de vida de sistemas de IA.

1.2. El Incidente Crítico: Resumen Ejecutivo de la Controversia en el Tribunal del Distrito Sur de Nueva York

El 22 de junio de 2023, el Juez P. Kevin Castel del Tribunal del Distrito Sur de Nueva York (Southern District of New York, SDNY) emitió una Opinión y Orden de Sanciones en el caso Mata v. Avianca, Inc. (678 F.Supp.3d 443)[2]. Este caso representa el primer precedente judicial de magnitud internacional donde un tribunal federal estadounidense sanciona expresamente a letrados por mala práctica técnica derivada del uso irresponsable de ChatGPT como herramienta de investigación jurídica.

Los abogados querellantes utilizaron ChatGPT para identificar precedentes sobre prescripción bajo el Convenio de Montreal. El modelo generativo produjo citas a casos inexistentes, entre ellos "Varghese v. China Southern Airlines", que fue incorporado al escrito de oposición como autoridad vinculante. El tribunal no solo rechazó las citas falsas, sino que inició procedimiento sancionador bajo la Regla 11 de las Reglas Federales de Procedimiento Civil (FRCP Rule 11), resultando en multas y reprimendas públicas.[3]

1.3. Objetivos del Informe: Guía sobre Impacto Técnico, Jurídico y Deontológico

Este informe persigue tres objetivos integrados:

(i) Análisis Técnico: Explicar los mecanismos subyacentes de alucinación en LLM, distinguiendo entre "oráculo creativo" (generación probabilística) y "archivero experto" (recuperación verificada de datos).

(ii) Análisis Jurídico: Evaluar la responsabilidad profesional bajo marcos deontológicos (ABA Opinión Formal 512, normas españolas del CGAE), normas procedimentales (Regla 11 FRCP) y marcos regulatorios emergentes (Reglamento de IA de la UE, normas españolas de transparencia).

(iii) Análisis de Gobernanza: Proponer mecanismos de mitigación técnica (arquitectura RAG, supervisión jerárquica de IA) e institucional (certificación de herramientas, auditoría permanente) para garantizar que la IA en derecho actúe como "amplificador del juicio humano" y no como sustituto del pensamiento crítico.

1.4. Metodología: Análisis de Doctrina Especializada, Marcos de Gestión de Riesgos y Jurisprudencia Comparada

La metodología combina:

• Análisis doctrinal comparativo: Opiniones formales de la ABA, guías de autoridades de protección de datos (EDPB, AEPD), y análisis técnico de especialistas (Informe Dantart, CIO LittleJohn, sobre alucinaciones en IA jurídica).

• Evaluación de marcos de gobernanza: Estándares internacionales (ISO/IEC 42001, NIST AI RMF 1.0, Perfil de IA Generativa) y normas regulatorias emergentes (Reglamento de IA de la UE entrada en vigor agosto 2025; normas españolas del Libro Blanco CGAE-ICAV).

• Análisis jurisprudencial: Estudio de sentencia Mata v. Avianca (SDNY, 22.6.2023) como precedente e identificación de riesgos análogos en contextos español y europeo.

CAPÍTULO II: ARQUITECTURA TÉCNICA Y EL FENÓMENO DE LA ALUCINACIÓN

2.1. IA Generativa vs. IA Consultiva: El Paradigma del "Oráculo Creativo" vs. "Archivero Experto"

La distinción fundamental en sistemas de IA aplicados a derecho radica en su arquitectura funcional:

ORÁCULO CREATIVO (IA Generativa Pura):
Sistema que genera texto probabilísticamente basado en patrones estadísticos del entrenamiento. Su objetivo es maximizar fluidez y coherencia lingüística. No tiene mecanismo integrado de verificación de hechos. Ejemplo: ChatGPT usado sin restricciones para research jurídica.

Características: (i) Generación autónoma sin anclaje a fuentes verificadas; (ii) Tendencia a "alucinar" cuando carece de datos de entrenamiento; (iii) Confianza aparente en afirmaciones incorrectas; (iv) Riesgo extremo en aplicaciones de alto riesgo.

ARCHIVERO EXPERTO (IA Consultiva Verificada):
Sistema que recupera información de bases de datos jurídicas certificadas y la presenta mediante razonamiento estructurado. Opera bajo el paradigma RAG (Retrieval-Augmented Generation), donde cada respuesta está anclada en fuentes verificables. Ejemplo: IA jurídica con acceso controlado a bases de datos jurídicas normalizadas.

Características: (i) Anclaje obligatorio en fuentes; (ii) Capacidad de citar con precisión pinpoint; (iii) Transparencia sobre límites de conocimiento; (iv) Riesgo significativamente reducido mediante arquitectura técnica.

El caso Mata v. Avianca ejemplifica precisamente el colapso de transitar de "archivero experto" (búsqueda en bases legales reales) a "oráculo creativo" (confianza en generación del modelo).

2.2. Mecanismos de los LLM: Funcionamiento Probabilístico y la Predicción de la Siguiente Palabra vs. Comprensión Lógica

Los LLM como GPT-3, GPT-4 y sus variantes operan mediante una arquitectura de transformer que realiza predicción secuencial de tokens (palabras o fragmentos de palabras). En cada iteración, el modelo calcula una distribución de probabilidad sobre el vocabulario siguiente basada en los tokens precedentes y los pesos aprendidos durante el entrenamiento.

MECANISMO SUBYACENTE:

1. Tokenización: El texto de entrada se divide en tokens.
2. Embedding: Cada token se representa como vector denso de alta dimensionalidad.
3. Atención Múltiple: El modelo calcula relaciones probabilísticas entre tokens.
4. Alimentación Delantera: Se aplican capas de transformación que mapean a espacio latente.
5. Predicción de Siguiente Token: Se genera distribución de probabilidad sobre tokens posibles.
6. Muestreo: Se selecciona token siguiente según estrategia (greedy, top-k, etc.).
7. Iteración: Se repite hasta generar secuencia completa.

IMPLICACIÓN CRÍTICA: Este mecanismo es estadístico, no simbólico. El modelo no "comprende" lógica formal ni mantiene mapeo explícito a referentes del mundo real. No distingue entre (i) un texto que aparece frecuentemente en datos de entrenamiento, (ii) un texto generado que es coherente pero falso, (iii) un texto que es factualmente correcto.

Según el Informe Técnico de Alex Dantart (CIO LittleJohn, publicado en arXiv): "El desafío de la veracidad en IA jurídica no es un problema de 'entrenamiento insuficiente' sino una característica arquitectónica fundamental de los modelos generativos que priorizan fluidez sobre factualidad." Esta tensión es irreducible en arquitecturas generativas puras.

2.3. Taxonomía de las Alucinaciones Legales

Una "alucinación" en contexto de LLM es generación de contenido que es internamente coherente pero factualmente incorrecto, con aparente confianza en su veracidad. En contexto jurídico, se pueden clasificar alucinaciones en categorías:

2.3.1. FABRICACIÓN DE AUTORIDAD (Casos y Citas Inexistentes)

Definición: Generación de referencias a casos, leyes o pronunciamientos que no existen en registros públicos verificables.

Ejemplo paradigmático: "Varghese v. China Southern Airlines" (caso Mata v. Avianca). El modelo generó nombre de partes, número de caso y aplicable jurisprudencia sobre Convenio de Montreal, con sintaxis idéntica a casos reales. El abogado, confiando en la coherencia superficial, citó el caso no como hipotetizado sino como autoridad vinculante.

Riesgo en contexto jurídico: CRÍTICO. Una cita falsa incorporada en escrito procesal puede: (i) viciar procedimiento; (ii) constituir incumplimiento grave del deber de franqueza ante tribunal; (iii) exponer al letrado a sanciones bajo Regla 11 FRCP (multa, suspensión, inhabilitación).

2.3.2. MISGROUNDING O FUNDAMENTACIÓN ERRÓNEA DE FUENTES REALES

Definición: Atribución de doctrina, razonamiento o conclusión jurídica incorrecta a una fuente que existe y es verificable, pero que dice algo diferente.

Ejemplo: Modelo que cita "Convenio de Montreal, artículo 35" como fuente de regla sobre prescripción, cuando el artículo 35 en realidad trata sobre responsabilidad del transportista aéreo (tema conexo pero normativa diferente).

Riesgo: ALTO. El lector que verifica la fuente primaria encontrará el artículo citado, pero la doctrina extraída será distorsionada. Esto es más difícil de detectar que fabricación pura porque tiene "veneno en forma de medicina".

2.3.3. ERRORES DE APLICACIÓN JURISDICCIONAL Y TEMPORAL

Definición: Atribución de vigencia, jurisdicción o aplicabilidad incorrecta a normas o precedentes.

Ejemplo: Aplicación de jurisprudencia de tribunal estatal estadounidense a contexto federal; o cita de sentencia modificada por posterior jurisprudencia sin referencia a revocación.

Riesgo: MEDIO-ALTO. Especialmente grave en investigaciones sobre prescripción (elemento temporal crítico) o jurisdicción aplicable (elemento territorial crítico).

2.4. Causas Técnicas: Limitaciones de Datos de Entrenamiento y Tensión Irreducible

Las causas técnicas de alucinación en LLM jurídicos son múltiples y, en gran medida, estructurales:

CAUSA 1: Cutoff Temporal de Entrenamiento
Los LLM se entrenan con corpus de datos con fecha de corte fija (ejemplo: GPT-4 entrenado hasta abril 2024). Casos posteriores a esa fecha son desconocidos. Cuando se pregunta sobre precedente reciente, el modelo "extrapola" basándose en patrones sintácticos, generando caso plausible pero ficticio.

Implicación: En jurisprudencia que evoluciona rápidamente (especialmente IA, protección de datos), el cutoff temporal crea zona de alucinación especialmente densa.

CAUSA 2: Pesos Desbalanceados en Datos de Entrenamiento
Documentos jurídicos incluidos en entrenamiento no son representativos: hay mayor densidad de sentencias reportadas que de sentencias no reportadas; mayor densidad de jurisprudencia de cortes altas que bajas; distorsión geográfica y lingüística (menos casos en idiomas no-inglés, aun para sistemas entrenados en español).

Implicación: Modelo desarrolla estadísticas distorsionadas sobre qué "parece" un caso legal real.

CAUSA 3: Incompletitud de Bases de Datos Jurídicas Públicas
Incluso si el modelo accediera a todas las sentencias, muchas (especialmente sentencias de tribunales inferiores, resoluciones administrativas, laudos arbitrales) no están indexadas

públicamente. El modelo no puede alucinar sobre datos completamente ausentes de su entrenamiento.

Implicación: Para queries sobre jurisprudencia específica o fallos no reportados, tasa de alucinación es cercana al 100%.

CAUSA 4: Incompatibilidad Arquitectónica entre Generación y Factualidad
El mecanismo de predicción secuencial que hace LLM eficientes para generación lingüística es exactamente el que impide factualidad garantizada. Sistema que maximiza probabilidad del siguiente token no maximiza probabilidad de que oración completa sea verdadera.

Implicación: No hay "entrenamiento mejor" que resuelva esto. La arquitectura del transformer tiene límites fundamentales para fact-checking integrado.

CAUSA 5: Falta de Mecanismo de Abstención
LLM standard generan respuesta para toda entrada. No tienen mecanismo arquitectónico de abstención ("no sé"). Cuando carece de conocimiento, modelo confabula con confianza en lugar de expresar incertidumbre.

Implicación: Sesgo sistemático hacia sobre-confianza. Usuario experimenta no incertidumbre sino certeza falsa.

CAPÍTULO III: CRÓNICA Y ANÁLISIS DEL CASO MATA V. AVIANCA

3.1. Antecedentes Fácticos: La Demanda de Roberto Mata contra Avianca Airlines

Roberto Mata presentó demanda contra Avianca Airlines en el Tribunal del Distrito Sur de Nueva York (SDNY) tras sufrir lesiones durante un vuelo. El caso involucraba reclamación de daños bajo el Convenio para la Unificación de Ciertas Reglas Relativas al Transporte Aéreo Internacional, comúnmente conocido como Convenio de Montreal (1999).

El Convenio de Montreal establece régimen especial de responsabilidad y prescripción para transportistas aéreos. La prescripción es elemento temporal crítico: la demanda debe presentarse dentro de plazos específicos o caduca.

3.2. El Error Procesal: Uso de ChatGPT para Investigar Precedentes sobre Prescripción bajo el Convenio de Montreal

Los abogados de la parte demandante utilizaron ChatGPT (acceso estándar, sin restricciones) para investigar cómo cortes estadounidenses habían interpretado prescripción bajo el Convenio de Montreal. Esta estrategia reflejaba creciente confianza en herramientas de IA generativa para investigación jurídica preliminar.

Según la Sentencia del Juez P. Kevin Castel, los abogados sometieron al modelo consultas como: "¿Hay jurisprudencia que apoye que prescripción bajo Convenio de Montreal comienza a contarse desde lesión, no desde conocimiento de daño?"

ChatGPT respondió con citas a casos que aparentaban ser reportados y autoritativos. Las citas tenían sintaxis correcta, números de docket plausibles, incluían años y cortes específicas. Todo elemento superficial sugería:
- Caso reportado en sistemas judiciales estadounidenses
- Jurisprudencia sobre tema específico (Convenio de Montreal + prescripción)
- Autoridad vinculante para SDNY

3.3. La Cascada de Errores: Incorporación de "Varghese v. China Southern Airlines" y Otras Citas Ficticias

El modelo generó múltiples citas, entre las cuales:

CITACIÓN PRIMARIA FICTICIA:
"Varghese v. China Southern Airlines Co., Ltd., 925 F.3d 1431 (9th Cir. 2019)"

Elementos de la fabricación:
- Partes plausibles: Varghese (apellido común en demandas de lesiones); China Southern Airlines (aerolínea real)
- Número de docket coherente: 925 (volumen) F.3d (Federal Reporter, 3ª serie) 1431 (página)
- Circuito correcto: 9th Circuit (es tribunal que típicamente resuelve casos de transportistas aéreos del Pacífico)
- Año plausible: 2019 (dentro de período donde Convenio de Montreal era jurisprudencia consolidada)

Los abogados incorporaron esta cita en escrito de oposición (opposition brief) presentado ante el tribunal. Trataron "Varghese" como precedente autoritativo y vinculante, no como ejemplo hipotético o como idea general.

CONSECUENCIA PROCESAL INMEDIATA:

El Juez Castel revisó la cita como parte de revisión estándar de escritos jurídicos. Al citar "Varghese", solicitó al tribunal que verificara la sentencia (procedimiento estándar de due diligence judicial). La verificación reveló: el caso no existe en ninguna base de datos de sentencias estadounidenses (westlaw, lexis, google scholar, etc.).

3.4. Respuesta Judicial: El Escrutinio del Juez P. Kevin Castel y la Audiencia de Sanción bajo la Regla 11

Tras confirmación de que "Varghese" era ficticia, el Juez Castel iniciió procedimiento sancionador bajo Regla 11 de las Reglas Federales de Procedimiento Civil (Federal Rule of Civil Procedure 11, FRCP Rule 11).

La Regla 11 FRCP establece:
"Las alegaciones, denuncias, defensas y otros escritos presentados ante la corte deben estar certificados por abogado. Al presentar un escrito, el abogado certifica que: (i) el escrito no ha sido presentado para propósito de retraso o dilación; (ii) los argumentos legales están apoyados en ley (existente o en propuesta razonable de cambio legal); (iii) los hechos tienen evidencia que los apoye o tendrán tal evidencia tras descubrimiento de prueba."

APLICACIÓN AL CASO MATA:
Las citas a "Varghese" y otros casos ficticios violaban requisito (ii): los argumentos legales NO estaban apoyados en ley existente (porque citaban casos que no existen). Los abogados no tuvieron base fáctica razonable para creer que las citas eran precisas.

HALLAZGOS DE NEGLIGENCIA DEL TRIBUNAL:

El Juez Castel concluyó que los abogados cometieron negligencia procesal grave al:
1. No verificar citas mediante acceso a bases de datos estándar (Westlaw, Lexis)
2. No mantener supervisión de herramienta (ChatGPT) de cuyos límites tenían o deberían tener conocimiento
3. No implementar protocolo de due diligence para output de IA generativa
4. Depositar confianza en modelo de IA sin validación independiente

SANCIONES IMPUESTAS (Sentencia Mata v. Avianca, SDNY 22.6.2023, p. 443-445):

- Multa a cada abogado: USD $5,000
- Multa adicional al bufete: USD $10,000
- Orden de capacitación obligatoria en investigación jurídica
- Orden de supervisión de escritos futuros
- Reprimenda pública (sentencia reportada y comentada en medios especializados)

PRECEDENTE ESTABLECIDO:

Este caso establece precedente internacional de que:
(i) Uso de IA generativa para investigación jurídica está permitido, pero
(ii) Responsabilidad profesional del abogado NO DISMINUYE con delegación a IA
(iii) Deber de verificación supera deber de "confianza razonable" en herramienta
(iv) Negligencia en supervisión de IA constituye incumplimiento de Regla 11 FRCP
(v) Daño reputacional es complemento a multa económica

CAPÍTULO IV: IMPACTOS Y RIESGOS: DEONTOLOGÍA Y MARCO REGULATORIO

4.1. Responsabilidad y Diligencia Profesional

4.1.1. Competencia Tecnológica bajo Regla Modelo 1.1 de la ABA

La Opinión Formal 512 de la American Bar Association (29.7.2024) establece que competencia profesional ("competence") ahora incluye necesariamente competencia tecnológica sobre herramientas de IA que el abogado utilice o considere utilizar en su práctica.

Según ABA Opinion 512, Sección 1.1 del Modelo Rules of Professional Conduct:

"Un abogado será competente cuando el abogado posea el nivel de conocimiento, destreza, preparación y diligencia razonablemente necesarias para representar al cliente. La competencia profesctica requiere del abogado: (1) Competencia técnica en la ley sustantiva y procesal aplicable; y (2) Competencia sobre herramientas y métodos utilizados en la práctica actual, incluyendo tecnología."

IMPLICACIÓN PARA IA GENERATIVA:
Si un abogado utiliza ChatGPT, Claude, Bard, o similar para investigación jurídica, debe poseer:
- Comprensión de capacidades técnicas de la herramienta
- Comprensión de limitaciones arquitectónicas (especialmente tendencia a alucinaciones)
- Conocimiento de tasa de alucinación documentada en estudios técnicos
- Protocolos de verificación implementados antes de uso

Si el abogado NO posee estas competencias, debe:
(a) Obtenerlas mediante capacitación, o
(b) No utilizar la herramienta, o
(c) Utilizar solamente bajo supervisión de abogado que sí posea competencias (supuesto de abogado junior con supervisión)

ESTÁNDAR VERIFICABLE:
La competencia tecnológica NO es abstracta. Puede ser evaluada mediante preguntas concretas:
- "¿Qué es una alucinación en LLM?" (Respuesta esperada: generación coherente pero factualmente falsa)
- "¿Cuál es la tasa estimada de alucinación de ChatGPT en queries jurídicas?" (Respuesta: estudios recientes muestran 15-25% según dominio jurídico)
- "¿Qué protocolos de verificación implementa su bufete?" (Respuesta: debe incluir cross-check contra bases autorizadas)

CASO MATA COMO PRECEDENTE NEGATIVO:
El Juez Castel, aunque no citó explícitamente ABA Opinion 512 (que se emitió casi 1 año después de la sentencia), aplicó estándar equivalente. Concluyó que abogados debían tener competencia sobre límites de ChatGPT y responsabilidad por no implementarla.

4.1.2. Deber de Supervisión de Asistentes "No Humanos" bajo Regla 5.3 ABA

La Regla Modelo 5.3 de la ABA establece:

"Un abogado es responsable por acciones de personas no abogado [asociados, auxiliares, personal de soporte] que están bajo la supervisión directa del abogado y que se desempeñan en las funciones de la práctica legal, respecto a conducta que si fuera llevada a cabo por el abogado constituiría violación de estas reglas, cuando: (a) el abogado ordena la conducta o,

conociendo de tal conducta, la aprueba tácitamente; o (b) el abogado es negligente en supervisar a la persona."

EXTENSIÓN ANÁLOGA A IA:
La ABA Opinion 512 extiende el deber de supervisión de Regla 5.3 a herramientas de IA. Aunque ChatGPT no es "persona no abogado" en sentido técnico, funciona análogamente: es asistente que realiza tareas bajo control del abogado.

Bajo esta lectura:
- El abogado debe supervisar output de IA generativa
- Supervisión NO puede ser delegada a la herramienta misma ("confianza en que ChatGPT verificará su propio output")
- Supervisor debe implementar protocolos independientes de verificación
- Negligencia en supervisión constituye incumplimiento de Regla 5.3

ESTÁNDAR OPERATIVO:
Un protocolo mínimo de supervisión de IA generativa en contexto jurídico debe incluir:

1. VERIFICACIÓN CONTRA BASES PRIMARIAS
- Toda cita a caso, sentencia o pronunciamiento debe verificarse contra Westlaw, Lexis, o base oficial
- Toda cita a ley debe verificarse contra texto oficial del estatuto
- Verificación debe ser realizada por persona distinta de quien generó output de IA

2. CONTROL DE FECHA DE CORTE
- Documentación clara de fecha de corte de entrenamiento de modelo
- Conscientización de que cualquier query sobre derecho desarrollado post-cutoff tiene alto riesgo
- Especial vigilancia para áreas de derecho de rápida evolución (IA, protección de datos, seguridad cibernética)

3. PROTOCOLO DE ABSTENCIÓN
- Implementación de procedimiento donde IA se abstiene de responder sobre queries de alto riesgo (prescripción, plazos procesales críticos)
- Capacitación de personal sobre casos donde la IA debe ser rechazada proactivamente

4. DOCUMENTACIÓN DE SUPERVISIÓN
- Registro de quién verificó qué output, cuándo, mediante qué bases primarias
- Documentación transferible (en caso de descubrimiento de prueba o auditoría)

En caso Mata v. Avianca, la falla fue precisamente en este punto: no hubo supervisión verificable. Los abogados incorporaron output de ChatGPT directamente en escrito procesal sin verificación contra bases autorizadas.

4.2. Implicaciones Éticas y Procesales: El Deber de Franqueza ante el Tribunal

El deber de franqueza ante el tribunal (duty of candor to court) es principio deontológico fundamental en derecho angloamericano y, por analogía, en sistemas de derecho continental.

DEFINICIÓN:
El abogado tiene obligación afirmativa de: (i) no presentar evidencia que sabe es falsa; (ii) no omitir información que sabe es decisiva; (iii) revelar al tribunal información sobre autoridades jurídicas desfavorables.

APLICACIÓN A CITAS FALSAS:
Si un abogado presenta cita a caso ficticio, comete:
- Violación de deber de franqueza (presenta como "autoridad" algo que no existe)
- Violación de deber de honestidad (induce error al tribunal)
- Abuso de proceso (sistema de justicia depende de que autoridades legales sean verificables)

PROTECCIÓN DEL SECRETO PROFESIONAL:
El deber de franqueza crea tensión con protección del secreto profesional abogado-cliente. La doctrina establece:

- Secreto profesional NO ampara hechos falsos si son esenciales a determinación de caso
- Secreto profesional NO protege comunicación donde cliente solicita citar autoridad inexistente
- Si cliente insiste en estrategia fraudulenta, abogado tiene derecho (y en algunos jurisdicciones, obligación) de retirarse de la representación

En caso Mata, los abogados aparentemente creían de buena fe que las citas eran reales (output de ChatGPT era suficientemente plausible). El Juez Castel no imputó fraude intencional. Sin embargo, concluyó que negligencia en verificación fue bastante grave para justificar sanciones bajo Regla 11.

4.3. Marco Regulatorio Global

4.3.1. El Reglamento de IA de la Unión Europea: Clasificación de Riesgos y Gobernanza de Datos

El Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial (en adelante, "Reglamento de IA" o "RIA"), entró en vigor progresivamente, con aplicación plena prevista para agosto de 2025.

El RIA establece clasificación de sistemas de IA por "niveles de riesgo":

SISTEMAS PROHIBIDOS (Artículo 5 RIA):
- Identificación biométrica en tiempo real en espacios públicos (con excepciones de seguridad pública)
- Sistemas de puntuación social
- Manipulación cognitiva sobre grupos vulnerables

Conclusión: Sistemas prohibidos no pueden ser usados en derecho, aunque verbalmente se encuadren como "investigación asistida".

SISTEMAS DE ALTO RIESGO (Anexo III RIA):
Incluyen sistemas que impacten derechos fundamentales o decisiones jurídicas. El RIA aún no especifica explícitamente si "IA para investigación jurídica" cae en esta categoría. Sin embargo, análisis doctrinal conservador (reflejado en White Paper SGS sobre Gobernanza de IA en industria legal) sugiere que:

- IA que genera análisis jurídico sobre casos específicos = ALTO RIESGO (impacta directamente derechos de partes)
- IA que realiza búsqueda lexical con resultado verificable = RIESGO MEDIO
- IA que genera contenido administrativo no-decisivo = RIESGO BAJO

Para sistemas de ALTO RIESGO, RIA Artículos 9-15 exigen:
- Evaluación de impacto sobre derechos fundamentales
- Documentación de ciclo de vida
- Supervisión humana significativa (no nominal)
- Medidas de transparencia
- Gobernanza de datos que garantice exactitud

IMPLICACIÓN PARA CASO MATA EN CONTEXTO UE:

Si los abogados fueran portugueses, españoles, o de otra jurisdicción UE, uso de ChatGPT para generar análisis sobre Convenio de Montreal (que impacta directamente la demanda de Mata) podría constituir incumplimiento del RIA si:
(i) No fue clasificado como alto riesgo
(ii) No se implementó supervisión humana significativa
(iii) No se realizó evaluación de impacto
(iv) No se documentó ciclo de vida

El Reglamento de IA, a diferencia de FRCP Rule 11 (que es norma procesal), es norma sustantiva de gobernanza que precede al uso efectivo. La obligación es ex ante (antes de usar), no ex post (después de problema).

4.3.2. Políticas en España: La Política del CTEAJE y el Principio de "No Sustitución"

El Consejo General de la Abogacía Española (CGAE) y el Ilustre Colegio de Abogados de Valencia (ICAV) publicaron conjuntamente un Libro Blanco sobre Inteligencia Artificial y Abogacía que incluye diagnósticos y recomendaciones específicas para el sector en España.

El Libro Blanco CGAE-ICAV establece varios principios operativos:

PRINCIPIO DE "NO SUSTITUCIÓN" (p. 87-92 del Libro Blanco):

"Sistemas de IA en contexto legal deben ser diseñados y utilizados de modo que amplifiquen competencia profesional del abogado, no que la sustituyan. Particularmente, en áreas donde IA no puede garantizar verificabilidad de output (generación creativa de análisis, extrapolación a supuestos no directamente cubiertos por jurisprudencia), la responsabilidad y supervisión del abogado NO DISMINUYEN sino que se intensifican."

Esto se traduce en varios requerimientos prácticos:

(i) COMPETENCIA CERTIFICADA:
Abogado que utiliza IA generativa debe demostrar competencia en: tecnología específica, límites conocidos, protocolos de verificación.

(ii) PROTOCOLO DOCUMENTADO:
Todo bufete que utiliza IA generativa debe mantener política documentada que especifique: qué herramientas están autorizadas, para qué tareas, bajo qué supervisión, qué nivel de verificación se requiere.

(iii) SEGREGACIÓN DE RESPONSABILIDADES:
La generación de output por IA no debe ser realizada por la misma persona que realiza supervisión. Hay necesidad de "cuatro ojos" (four-eyes principle).

(iv) AUDITORÍA PERMANENTE:
Bufetes medianos o grandes deben realizar auditoría anual sobre uso de IA, incluyendo muestreo de casos para verificación de exactitud de citas y referencias.

ESPECIFICIDAD ESPAÑOLA:
El Libro Blanco también menciona que Ley 7/2021, de 20 de mayo, de impulso a la inteligencia artificial (ley sobre IA en sector público español), aunque se enfoca en administración, establece principios (transparencia, impugnabilidad de decisiones automatizadas, supervisión humana) que son análogos a deberes profesionales.

CTEAJE (Comisión Técnica Estatal de Acceso a la Justicia Electrónica):
El CTEAJE ha emitido notas técnicas (aunque no vinculantes) sugiriendo que herramientas de IA generativa no deben ser utilizadas para redacción de escritos sin supervisión externa verificable.

4.4. Consecuencias Sancionadoras: Análisis de Multas, Daños Reputacionales y el Precedente para la Judicatura

CONSECUENCIAS ECONÓMICAS:
En caso Mata v. Avianca, las sanciones fueron:
- Multas administrativas: USD $15,000 (USD $5,000 × 2 abogados + USD $5,000 más para asociación de abogados)
- Costas procesales: Aunque no reportadas en sentencia, típicamente las costas de audiencia de sanción pueden adicionar USD $3,000-$8,000
- Total estimado: USD $18,000-$23,000

CONSECUENCIAS DEONTOLÓGICAS:
- Reprimenda permanente en archivo profesional
- Potencial procedimiento disciplinario ante colegio de abogados
- Requisito de capacitación en investigación jurídica (costo: USD $2,000-$5,000)

DAÑOS REPUTACIONALES:
- Sentencia reportada en principales bases de datos (Westlaw, Lexis, Google Scholar)
- Cobertura en medios especializados (Above the Law, Legal Technology News)

- Búsqueda de nombres de abogados vinculada permanentemente al incidente
- Potencial impacto en futuros clientes (especialmente en práctica corporativa donde reputación es activo crítico)

PRECEDENTE ESTABLECIDO:
La sentencia Mata v. Avianca ha generado:

(i) JURISPRUDENCIA POSTERIOR
Tras la sentencia (junio 2023), múltiples tribunales estadounidenses han citado Mata en contextos donde se cuestiona exactitud de output de IA. La sentencia se ha convertido en referencia estándar.

(ii) CAMBIOS EN ESTÁNDARES PROFESIONALES
La ABA emitió Opinion 512 (julio 2024) parcialmente en respuesta a precedente Mata. Esto indica que jurisprudencia de cortes inferiores puede influir en desarrollo de normas deontológicas.

(iii) REACCIÓN DE REGULADORES INTERNACIONALES
Autoridades en UE, UK, y otras jurisdicciones han citado Mata al desarrollar marcos de gobernanza de IA en derecho.

CAPÍTULO V: ESTRATEGIAS DE GOBERNANZA Y MITIGACIÓN TÉCNICA

5.1. Gestión de Riesgos de IA: Implementación del Estándar ISO/IEC 42001 y el Marco NIST

La mitigación del riesgo de alucinación no es simplemente un "protocolo de verificación" ad hoc. Requiere implementación de metodologías formales de gestión de riesgos que alineen gestión técnica, gobernanza organizacional, y cumplimiento normativo.

5.1.1. ESTÁNDAR ISO/IEC 42001:

El estándar ISO/IEC 42001 (Information technology — Artificial intelligence management system) proporciona marco certificable para gobernanza de IA. Fue publicado por la

Organización Internacional de Normalización (ISO) el 5 de diciembre de 2023 y es de adopción creciente en sector legal.

ISO/IEC 42001 requiere:

CONTEXTO DE LA ORGANIZACIÓN (Cláusula 4):
- Análisis de entorno operativo: ¿Qué sistemas de IA están en uso? ¿Cuáles son críticos?
- Identificación de partes interesadas: Clientes, reguladores, empleados, público
- Determinación de necesidades y expectativas: Cumplimiento, seguridad, calidad

GOBERNANZA DE IA (Cláusula 5):
- Asignación de responsabilidades: Quién es responsable de cada sistema de IA
- Políticas documentadas: Qué reglas rigen uso de IA
- Evaluación de riesgos: Metodología sistemática para identificar riesgos
- Procedimientos de escalación: Qué hacer cuando sistema de IA genera output de riesgo

CICLO DE VIDA (Cláusula 6-8):
- Planificación: Antes de implementar IA, evaluación de necesidades y riesgos
- Entrenamiento: Si aplicable, gobernanza de datos de entrenamiento
- Validación: Pruebas antes de despliegue
- Despliegue: Procedimientos de introducción controlada
- Monitoreo: Vigilancia continua del desempeño en operación real
- Retirada: Procedimientos para discontinuar sistema si se vuelve intolerante de riesgo

MEDIDAS TÉCNICAS Y OPERACIONALES (Cláusula 8.2):
Incluyen: trazabilidad de decisiones, auditoría, documentación de incertidumbre, supervisión humana.

5.1.2. MARCO NIST AI RMF 1.0 (JULIO 2024) Y SU PERFIL DE IA GENERATIVA:

El National Institute of Standards and Technology (NIST) de los Estados Unidos publicó el 10 de julio de 2024 el Marco de Gestión de Riesgos de Inteligencia Artificial versión 1.0 (AI Risk Management Framework 1.0), junto con un Perfil específico para IA Generativa.

El Marco NIST establece cuatro funciones continuas:

FUNCIÓN 1 - MAP (MAPEAR):
Identificar sistemas de IA, su uso previsto, contexto operativo, y riesgos potenciales. En caso legal, incluiría: "¿Qué abogados tienen acceso a ChatGPT? ¿Para qué tareas específicas? ¿Cuáles son consecuencias de error?"

FUNCIÓN 2 - MEASURE (MEDIR):
Evaluar desempeño de sistema de IA, tasa de error, desviaciones de comportamiento esperado. Incluye: pruebas de alucinación documentadas, métricas de exactitud en generación de citas, auditoría de output.

FUNCIÓN 3 - MANAGE (GESTIONAR):
Implementar controles y mitigaciones. Incluye: arquitectura RAG (ver sección 5.2), supervisión humana, protocolos de escalación, capacitación de usuarios.

FUNCIÓN 4 - GOVERN (GOBERNAR):
Establecer políticas, asignación de responsabilidades, procedimientos de rendición de cuentas. Incluye: documentación de decisiones sobre qué sistemas usar, auditoría periódica, procedimientos de complaint.

PERFIL DE IA GENERATIVA (GENERATIVE AI PROFILE):

El Perfil específico reconoce que LLM como GPT-4, Claude, Bard tienen características especiales:

Riesgos caracterizados:
- Alucinación (generación de hechos falsos)
- Sesgo de entrenamiento (reproducción de sesgos demográficos en datos de entrenamiento)
- Inyección de prompt (usuario malicioso que manipula entrada para generar output no deseado)
- Privacidad (regurgitación de datos sensibles del entrenamiento)

Medidas específicas recomendadas:
- Pruebas adversariales (intentar que modelo alucine, medir frecuencia)
- Auditoría de sesgo (revisar output para patrones discriminatorios)
- Limitaciones técnicas (restricción de ciertos tipos de queries)
- Documentación de cutoff temporal de entrenamiento
- Evaluación periódica de cambios en desempeño del modelo (versiones nuevas pueden tener propiedades diferentes)

5.2. El Paradigma RAG (Generación Aumentada por Recuperación)

5.2.1. Cómo Anclar las Respuestas en Bases de Datos Legales Verificadas

RAG (Retrieval-Augmented Generation) es arquitectura de IA que resuelve directamente el problema de alucinación documentado en caso Mata v. Avianca. En lugar de generar respuesta basada puramente en pesos del modelo, RAG:

(1) RECUPERA documentos relevantes de base de datos verificada (ej: base de sentencias oficiales, repertorio de leyes)
(2) PROPORCIONA esos documentos como contexto adicional al modelo
(3) GENERA respuesta que está explícitamente anclada en documentos recuperados
(4) INCLUYE CITAS con referencias pinpoint (página, párrafo, frase exacta del documento fuente)

EJEMPLO OPERATIVO:

Usuario consulta: "¿Hay jurisprudencia sobre prescripción bajo Convenio de Montreal?"

PROCESO SIN RAG (ChatGPT estándar):
→ Modelo genera: "Varghese v. China Southern Airlines, 925 F.3d 1431 (9th Cir. 2019) establece que prescripción bajo Convenio comienza desde..."
→ Resultado: Respuesta plausible, pero completamente falsa.

PROCESO CON RAG:
(1) Recuperación: Sistema realiza búsqueda en base de sentencias oficial. Identifica: Stratis v. Soriano, 717 F.3d 1014 (9th Cir. 2013); Katz v. Household International, Inc., 490 F.3d 1031 (9th Cir. 2007) (casos reales sobre Convenio de Montreal)

(2) Aumento: Modelo recibe:
- Input del usuario: "¿Hay jurisprudencia sobre prescripción bajo Convenio de Montreal?"
- Contexto recuperado: [Textos completos de Stratis y Katz]
- Instrucción explícita: "Responda basándose ÚNICAMENTE en los documentos proporcionados. Si no encuentra respuesta, indique 'Sin información en documentos recuperados'."

(3) Generación: Modelo produce:
"Según Stratis v. Soriano, 717 F.3d 1014, apdo. 1032 (9th Cir. 2013), la prescripción bajo Convenio de Montreal es de dos años desde la muerte o lesión. El tribunal en Katz v. Household International, Inc., 490 F.3d 1031, apdo. 1045 (9th Cir. 2007), confirmó esta interpretación."

(4) Verificabilidad: Usuario (y abogado que supervisa) puede:

- Hacer click en Stratis para ver sentencia completa
- Confirmar que apdo. 1032 realmente dice lo que dice
- Verificar que cita es exacta y completa

VENTAJAS DE RAG:

✓ Elimina alucinación sobre contenido de fuentes (porque genera respuesta de documentos reales)
✓ Proporciona trazabilidad completa (cada afirmación tiene origen documentado)
✓ Reduce sobre-confianza (modelo "sabe" que solo puede hablar de documentos recuperados)
✓ Facilita supervisión humana (verificación es mecánica: solo confirmar que documentos recuperados realmente dicen lo que el resumen indica)
✓ Compatible con normas deontológicas (abogado puede estar seguro de que citas existen)

LIMITACIONES DE RAG:

✗ Requiere acceso a bases de datos de alta calidad (no todos los tribunales publican sentencias online)
✗ Limitado a dominio del conocimiento en base de datos (si pregunta es sobre jurisprudencia reciente no indexada, RAG no puede responder)
✗ Implica costo operativo (mantener base de datos, actualizarla regularmente)
✗ No soluciona problema de mala interpretación (modelo puede malinterpretar lo que lee)

IMPLICACIÓN PARA CASO MATA SI USARA RAG:

Si los abogados en Mata v. Avianca hubieran utilizado arquitectura RAG conectada a bases de sentencias oficiales (Westlaw, Lexis, Google Scholar API):
- Sistema no podría generar "Varghese v. China Southern Airlines" (porque no existe en bases)
- Sistema podría responder: "No encontré jurisprudencia sobre este punto específico. Tal vez desee refinar la búsqueda o consultar directamente base de datos."
- Sistema habría recuperado casos reales (Stratis, Katz, etc.) con citas exactas
- Abogado habría podido supervisar output sin necesidad de búsqueda adicional

5.2.2. Optimización de la Fase de Recuperación y Chunking Estructural

La efectividad de RAG depende críticamente de cómo se estructuran documentos recuperados. Esto es ciencia-tecnología llamada "chunking":

PROBLEMA DE CHUNKING:

Un documento jurídico (ej: sentencia de 30 páginas) contiene cientos de fragmentos de información:
- Hechos (págs. 2-5)
- Procedimiento anterior (págs. 5-7)
- Argumentos de cada parte (págs. 8-15)
- Análisis de ley por el tribunal (págs. 16-25)
- Conclusión y sentencia (págs. 26-30)

Si el sistema recupera el documento completo, el modelo puede:
- Confundirse por volumen de información
- Asignar peso incorrecto a secciones (ej: dar igual peso a dicta —comentarios no vinculantes — que a holding —conclusión vinculante)
- Mezclar hechos de caso con análisis legal general

SOLUCIÓN: CHUNKING ESTRUCTURAL

Documentos se dividen en "chunks" (segmentos) de tamaño y contenido optimizado:

NIVEL 1 - CHUNKS MACRO (una sección por chunk):
- "Hechos del caso"
- "Argumento A de demandante"
- "Argumento A de demandado"
- "Análisis judicial de cuestión 1 (prescripción)"
- "Análisis judicial de cuestión 2 (damages)"
- "Conclusión"

NIVEL 2 - CHUNKS MICRO (subsecciones dentro de análisis):
- "Párrafo de introducción: cita de estándar legal"
- "Párrafo de análisis: aplicación de estándar a hechos"
- "Párrafo de conclusión: conclusión sobre este sub-punto"

METADATOS:

Cada chunk incluye metadatos estructurados:
```
Documento: Stratis v. Soriano, 717 F.3d 1014
Tribunal: United States Court of Appeals, Ninth Circuit
```

Fecha: 2013-09-20
Tipo de contenido: Análisis legal
Tema: Convenio de Montreal - Prescripción
Holding o Dicta: Holding (vinculante)
Párrafo específico: 1032
Relevancia para ley: Interpretación del Artículo 29(3) Convenio de Montreal
```

RECUPERACIÓN MEJORADA:

Cuando usuario pregunta "¿Cuál es plazo de prescripción bajo Convenio de Montreal?", sistema:
(1) Busca en metadatos: documentos donde "tema = Convenio de Montreal - Prescripción" Y "tipo de contenido = análisis legal"
(2) Prioriza: Casos donde "holding o dicta = Holding" (decisiones vinculantes sobre punto exacto)
(3) Recupera: Chunks micro específicos (ej: "Párrafo 1032 de Stratis"), no documento completo
(4) Presenta: Chunk micro + referencia completa + instrucción al modelo de que foco debe ser en este párrafo

**RESULTADO:**

Modelo genera: "Según Stratis v. Soriano, 717 F.3d 1014 (9th Cir. 2013), párr. 1032, el plazo de prescripción bajo Convenio de Montreal es de dos años desde..."

En lugar de:
"Varghese v. China Southern Airlines ha establecido [completamente ficticio]..."

## 5.3. Agentes Jerárquicamente Conscientes: Integración de la Pirámide de Kelsen en la Lógica de la IA

Aunque RAG resuelve alucinación sobre hechos, persiste problema de interpretación: ¿Cómo garantizar que IA no malinterpreta autoridad jurídica?

La teoría jurídica de Hans Kelsen proporciona marco conceptual que puede ser integrado en lógica de IA. La Pirámide de Kelsen establece jerarquía de normas:

PIRÁMIDE DE KELSEN (Jerarquía de Normas):
```

NIVEL 1 (Superior): Constitución (fuente suprema de autoridad)
NIVEL 2: Leyes / Códigos (legislación ordinaria)
NIVEL 3: Reglamentos / Decretos (normas ejecutivas)
NIVEL 4 (Inferior): Sentencias / Acuerdos administrativos (aplicación de normas)

Una norma de nivel inferior es válida solo si es compatible con normas de niveles superiores. Una ley no puede contradecir constitución; un reglamento no puede contradecir ley; una sentencia no puede ser arbitraria ante ley.

INTEGRACIÓN EN IA JURÍDICA:

Un "agente jerárquicamente consciente" es sistema de IA que:
(1) Identifica explícitamente jerarquía de autoridades en respuesta
(2) Asigna peso diferenciado según nivel jerárquico
(3) Detecta y señala conflictos jerárquicos
(4) Requiere escalación humana si conflicto no puede ser resuelto

EJEMPLO OPERATIVO:

Búsqueda: "¿Es permitido usar IA para investigación jurídica sin supervisión?"

SIN CONCIENCIA JERÁRQUICA (ChatGPT estándar):
Respuesta: "Varios jueces han comentado que IA puede ser valiosa para investigación. Stratis v. Soriano sugiere que..."
[Cita sentencia como si tuviera peso normativo directo]

CON CONCIENCIA JERÁRQUICA:
Respuesta estructurada:

┌─ NIVEL CONSTITUCIONAL (Derecho a justicia efectiva, debido proceso)
│ └ No hay disposición constitucional explícita sobre IA en derecho
│ └ Aplicable: Principio general de justicia accesible
│
├─ NIVEL LEGISLATIVO (Norma primaria)
│ ├ USA: Federal Rules of Civil Procedure, Rule 11 (requiere "certificación" de escritos)
│ ├ España: Reglamento de IA (Art. 5, sistemas prohibidos; Art. 9-15, sistemas alto riesgo)
│ ├ ABA Model Rules (Regla 1.1 competencia; Regla 5.3 supervisión)
│ └ CONCLUSIÓN: Legislación requiere supervisión, competencia, documentación
│

├─ NIVEL REGLAMENTARIO (Orientación operativa)
│ ├ ABA Opinion 512 (2024): Obligación de competencia tecnológica
│ ├ ISO/IEC 42001: Estándar de gobernanza de IA
│ ├ NIST AI RMF: Marco de gestión de riesgos
│ └ CONCLUSIÓN: Estándares indican supervisión significativa obligatoria
│
└─ NIVEL JURISPRUDENCIAL (Precedentes, no vinculantes en jerárquico pero instructivos)
├ Mata v. Avianca (SDNY 2023): Negligencia en supervisión → sanciones
└ CONCLUSIÓN: Jurisprudencia confirma riesgo si no se implementa supervisión
```

VENTAJAS DE CONCIENCIA JERÁRQUICA:

✓ Previene que sentencia sea citada como si fuera ley
✓ Requiere que análisis comience con marco legal correcto
✓ Detecta automáticamente si consejo es "contrario a ley"
✓ Facilita revisión humana (abogado ve estructura jerárquica y puede identificar si IA omitió o malinterpretó nivel)

IMPLEMENTACIÓN TÉCNICA:

En arquitectura RAG con conciencia jerárquica:
(1) Cada documento recuperado está etiquetado con su nivel en jerarquía
(2) Modelo es instruccionado a estructurar respuestas según niveles
(3) Si detecta conflicto entre niveles (ej: precedente judicial contradice ley), lo señala explícitamente
(4) Sistema requiere escalación humana si conflicto no puede ser resuelto automáticamente

## 5.4. El "Filtro Humano": La Supervisión Experta como Componente Irreducible e Innegociable

Todas las estrategias técnicas anteriores (RAG, chunking, conciencia jerárquica) son necesarias pero insuficientes. El "filtro humano" es elemento que no puede ser delegado a la máquina.

Según el Informe Técnico de Alex Dantart (CIO LittleJohn): "La alucinación en IA legal no será resuelta técnicamente porque es característica arquitectónica de LLM. Lo que sí puede ser
```

minimizado es su IMPACTO, mediante introducción de escalones de verificación humana donde el costo de error es inaceptable."

ARQUITECTURA DE SUPERVISIÓN EN TRES ESCALONES:

ESCALÓN 1: GENERACIÓN Y PRESENTACIÓN INICIAL
Sistema de IA genera respuesta completa. Incluye:
- Respuesta textual
- Citas con referencias pinpoint
- Indicación de confianza (ej: "Alta confianza", "Confianza media", "Baja confianza")
- Indicación de gaps (si hay aspectos del query que no pudo responder)

Ejemplo de output:
"Pregunta: ¿Hay jurisprudencia sobre prescripción bajo Convenio de Montreal?
Respuesta: [Stratis v. Soriano cita con detalle]
Confianza: Alta (documento es sentencia reportada, cita es verificable)
Gaps: No encontré jurisprudencia específica sobre interpretación del Convenio en cortes españolas; posible que requiera research adicional en base de jurisprudencia española"

ESCALÓN 2: SUPERVISIÓN ESPECIALIZADA (First-Level Review)
Abogado junior o auxiliar jurídico (con capacitación en uso de IA) realiza verificación de:
(i) Existencia de fuentes: ¿Existen los casos citados?
(ii) Exactitud de citas: ¿Dice la sentencia lo que el resumen indica?
(iii) Pertinencia: ¿Las sentencias son realmente relevantes al caso?
(iv) Completitud: ¿Hay autoridades más relevantes que el sistema no recuperó?

Procesos:
- Para cada cita, búsqueda en Westlaw/Lexis
- Lectura de sentencia (no completa, pero al menos párrafo citado)
- Anotación en documento: "✓ Verificado - Exacto" o "✗ Error - [Descripción]"
- Documentación en archivo: Quién verificó, fecha, resultado

Documentación típica:
```
VERIFICACIÓN DE CITAS - Caso: Mata v. Avianca, Inc.
Realizado por: [Nombre Asistente Jurídico], Fecha: 2024-02-15

Cita 1: Stratis v. Soriano, 717 F.3d 1014 (9th Cir. 2013)
Verificado en: Westlaw
```

Estado: ✓ Exacto - Párr. 1032 confirmado
Nota: Caso es binding authority en jurisdicción aplicable

Cita 2: Katz v. Household International, Inc., 490 F.3d 1031 (9th Cir. 2007)
Verificado en: Westlaw
Estado: ✓ Exacto - Pero nota que fue parcialmente overruled por [Sentencia X, 2015]
Nota: Debe incluirse en respuesta que doctrina fue limitada por posteriores

Cita 3: Varghese v. China Southern Airlines, 925 F.3d 1431 (9th Cir. 2019)
Verificado en: Westlaw, Lexis, Google Scholar, RECAP
Estado: ✗ NO EXISTE - Caso es FICTICIO
Nota: CRÍTICO - Eliminar completamente de escrito. Investigar por qué IA lo generó.
```

ESCALÓN 3: REVISIÓN EXPERTA (Senior Review)
Abogado senior (por lo menos 5+ años de experiencia en área de derecho específica) realiza revisión de:
(i) Solidez del análisis jurídico (más allá de verificación de hechos)
(ii) Completitud de tratamiento de autoridades (¿Hay jurisprudencia contradictoria que no fue recuperada?)
(iii) Aplicabilidad al caso específico (¿El análisis de IA realmente responde al problema del cliente?)
(iv) Estrategia procesal (¿Este análisis avanza la posición del cliente?)

Documentación:
"REVISIÓN EXPERTA - Caso: Mata v. Avianca, Inc.
Realizado por: [Nombre Abogado Senior], Especialización: Derecho de Transportes, Fecha: 2024-02-20

**HALLAZGOS:**
1. Análisis de prescripción bajo Convenio de Montreal es correcto en términos fácticos y jurisprudenciales.
2. Completitud: El análisis de IA recuperó Stratis y Katz, que son autoridades principales. Adicionalmente, identifico [Sentencia reciente X, 2024] que modifica interpretación de Stratis. RECOMENDACIÓN: Incluir esta autoridad.
3. Aplicabilidad: Nuestro caso presenta variable [X] que no está directamente cubierta por Stratis/Katz. RECOMENDACIÓN: Realizar análisis adicional sobre cómo [X] interactúa con holding de estos casos.
```

4. CONCLUSIÓN: Output de IA es sólido pero incompleto. Requiere análisis complementario antes de presentar a tribunal. Estimado: 2-3 horas de research adicional de abogado senior."

BENEFICIO DE ARQUITECTURA DE TRES ESCALONES:

✓ Escalón 1 (IA): Eficiencia (investigación rápida de hechos base)
✓ Escalón 2 (Junior + verificación): Control de calidad (garantiza no hay hechos falsos)
✓ Escalón 3 (Senior + análisis): Juicio experto (garantiza solidez de razonamiento jurídico)

✓ Documentación completa: Si hay problema posterior, se puede rastrear exactamente dónde falló supervisión

✓ Escalabilidad: Arquitectura permite que abogados senior deleguen verificación de hechos a junior, manteniendo supervisión de análisis conceptual

CASO MATA BAJO ESTA ARQUITECTURA:

Los abogados utilizaron: Escalón 1 solo (IA generó citas) sin Escalón 2 (verificación de citas) o Escalón 3 (revisión experta).

Si hubieran implementado esta arquitectura:
- Escalón 2 habría identificado inmediatamente que "Varghese" no existe
- Escalón 3 habría revisado escrito antes de presentar al tribunal

El costo en tiempo de un abogado junior (USD $50-$100/hora) habría sido mínimo comparado con las sanciones posteriores (USD $15,000+).

CAPÍTULO VI: RECOMENDACIONES PRÁCTICAS Y CONCLUSIONES

6.1. Decálogo para el Abogado del Futuro: Verificación, Transparencia y Formación Continua

Basándose en análisis de caso Mata v. Avianca, marcos de gobernanza (ABA Opinion 512, NIST AI RMF, ISO/IEC 42001) y regulación emergente (Reglamento de IA UE, normas españolas), se propone un conjunto de principios prácticos:

PRINCIPIO 1 - COMPETENCIA NO NEGOCIABLE
"Antes de utilizar cualquier herramienta de IA en práctica legal, debo poseer comprensión demostrable de:
- Cómo funciona técnicamente (probabilística, no lógica)
- Cuáles son sus límites documentados (especialmente alucinaciones)
- Qué pueden (y no pueden) garantizar sus proveedores
- Qué protocolo de supervisión es exigido"

Operacionalización:
□ Completar curso certificado sobre IA generativa (mínimo 8 horas de capacitación formal)
□ Realizar pruebas de la herramienta específica con queries conocidas (verificar su comportamiento)
□ Documentar en política de bufete qué herramientas están autorizadas y bajo qué restricciones
□ Mantener actualización sobre cambios en herramientas (versiones nuevas, cambios en comportamiento)

PRINCIPIO 2 - VERIFICACIÓN COMO DEBER NO DELEGABLE
"No puedo delegar a la máquina la responsabilidad de verificación de su propio output. La supervisión humana es irreducible."

Operacionalización:
□ Para cada cita, búsqueda en base de datos autoritativa antes de incorporar en escrito
□ Para hechos jurídicos críticos (plazos, jurisdicción), multiple sources de verificación
□ Documentación escrita de quién verificó qué, cuándo, mediante qué métodos
□ Mantención de archivo físico o digital de verificaciones (en caso de descubrimiento posterior)

PRINCIPIO 3 - TRANSPARENCIA PROACTIVA CON CLIENTE
"Debo informar al cliente qué rol jugó IA generativa en mi investigación, si es material a la estrategia."

Operacionalización:
□ Disclosure en engagement letter: "El bufete utiliza herramientas de IA como asistencia en investigación, siempre bajo supervisión profesional"

□ En memorandos a cliente, si IA jugó papel significativo: "Este análisis fue realizado con asistencia de [herramienta], verificado manualmente contra bases autorizadas"
□ En contexto de negociación o procedimiento, revelar si autoridad jurídica fue identificada por IA (no porque sea obligatorio en todos los casos, pero porque incrementa credibilidad)

PRINCIPIO 4 - DOCUMENTACIÓN DEFENSIBLE
"Debo mantener registro que demuestre que ejercí diligencia razonable en supervisión."

Operacionalización:
□ Archivo de cada búsqueda realizada (entrada a IA, output, verificación)
□ Anotaciones de cambios hechos a output de IA antes de incorporarlo
□ Autorización documentada de uso de herramienta (firma en política de bufete)
□ Capacitación documentada (certificados de cursos, fechas, duración)

PRINCIPIO 5 - SEGREGACIÓN DE RESPONSABILIDADES
"Una persona no puede ser simultáneamente usuario de IA y supervisor de su propio output."

Operacionalización:
□ Si yo generé query a IA, otra persona debe verificar resultado
□ Si soy junior, supervisor senior debe revisar antes de cualquier output llegar a cliente/tribunal
□ Para casos de alto riesgo (litigación), revisión múltiple obligatoria
□ Documentación de quién participó en cada paso

PRINCIPIO 6 - ARQUITECTURA RAG DONDE POSIBLE
"Preferiré herramientas que anclen respuestas en bases verificables sobre herramientas generativas puras."

Operacionalización:
□ Evaluación de herramientas disponibles: ¿Cuáles usan RAG + bases autorizadas?
□ Preferencia por herramientas jurídicas especializadas sobre ChatGPT general
□ Verificación de que base de datos que herramienta usa es exhaustiva y actual
□ Conocimiento de cutoff temporal de base de datos (cuándo fue última actualización)

PRINCIPIO 7 - ABSTENCIÓN TEMPRANA
"Si IA genera output que tengo dudas sobre su confiabilidad, mejor hacer research manualmente que incorporar output dudoso."

Operacionalización:

□ Si herramienta expresa baja confianza en respuesta, no usarla
□ Si query es sobre jurisprudencia reciente (post-cutoff de entrenamiento), confiar poco en generación
□ Si query es sobre tema donde no soy experto, sesgo hacia mayor verificación
□ Si resultado sorprende por lo "bueno" o "malo", verificar adicional antes de actuar

PRINCIPIO 8 - CONCIENCIA DE LÍMITES JURISDICCIONALES
"IA entrenada principalmente en derecho estadounidense puede tener limitaciones severas en derecho español o europeo."

Operacionalización:
□ Para queries sobre derecho español o europeo, preferencia por herramientas entrenadas en esas jurisdicciones
□ Reconocimiento que ChatGPT/Claude pueden ser débiles en jurisprudencia de tribunales españoles
□ Búsqueda adicional en bases españolas (BuscaJurisprudencia, jurisprudencia.poderjudicial.es)
□ Para asuntos europeos (protección de datos, IA), búsqueda prioritaria en jurisprudencia TJUE

PRINCIPIO 9 - FORMACIÓN CONTINUA NO OPCIONAL
"La tecnología evoluciona rápidamente. Competencia requiere actualización permanente."

Operacionalización:
□ Mínimo 2-4 horas anuales de capacitación en IA legal y herramientas disponibles
□ Seguimiento de cambios en Regla 11 FRCP, ABA Model Rules, normas españolas
□ Participación en webinars o conferencias sobre IA en derecho
□ Lectura de análisis técnico (p.ej., papers en arXiv sobre alucinación en IA legal)

PRINCIPIO 10 - HUMANIDAD COMO FUNDAMENTO
"IA es herramienta para amplificar mi juicio profesional, no para reemplazarlo. Si algo te parece dudoso, confía en tu juicio."

Operacionalización:
□ Reflexión crítica: Preguntarme "¿Tiene sentido esta conclusión dado lo que sé sobre derecho?"
□ Contrapeso a automatización: Deliberadamente hacer pregunta opuesta ("¿Cuál es argumento contrario?")

□ Preservación de pensamiento independiente: Generar mi propio análisis, luego comparar con output de IA

□ Humildad intelectual: Reconocimiento que máquina y yo tenemos fortalezas/debilidades complementarias

6.2. Selección Responsable de Herramientas: Criterios de Evaluación de Proveedores y Transparencia en la "Caja Negra"

No todas las herramientas de IA son equivalentes. La selección debe ser rigurosa y documentada.

MATRIZ DE EVALUACIÓN DE HERRAMIENTAS IA PARA USO JURÍDICO:

Criterio	Peso	Método de Evaluación	Umbral Aceptable
ARQUITECTURA: ¿Usa RAG o generación pura?	25%	Consultar especificaciones del proveedor; preguntas técnicas directas	Obligatorio RAG para alto riesgo
BASES DE DATOS: ¿Acceso a sentencias autorizadas?	20%	Verificar qué bases de datos tiene acceso (Westlaw, Lexis, oficial, etc.)	Acceso a mínimo 2 bases autorizadas
TASA DE ALUCINACIÓN DOCUMENTADA	20%	Solicitar estudios internos o publicados sobre alucinación en queries jurídicas	<10% para queries jurídicas estándar
TRANSPARENCIA: ¿Qué información proporciona sobre incertidumbre?	15%	Pruebas de la herramienta: ¿Expresa confianza? ¿Se abstiene de responder?	Requiere que herramienta indique confianza o abstención
CUMPLIMIENTO REGULATORIO: ¿Cumple RGPD, RIA, normas aplicables?	10%	Revisar documentación de privacy/security; SOC 2 certification; data location	Obligatorio cumplimiento GDPR; preferible ISO/IEC 42001
SOPORTE Y ACTUALIZACIÓN: ¿Hay SLA? ¿Actualizaciones regulares?	10%	Revisión de terms of service; historial de versiones	Mínimo: soporte empresarial; actualizaciones >2x anuales

EVALUADORES SUGERIDOS:

Para bufete pequeño-medio (1-20 abogados): Responsable técnico del bufete + 1 abogado senior realiza evaluación, documento resumen.

Para bufete grande (>50 abogados): Comité de evaluación incluyendo: Chief Information Officer, abogado especializado en tecnología, abogado de cada práctica que usará herramienta.

DOCUMENTACIÓN REQUERIDA:

El resultado de la evaluación debe ser documento formal titulado "Aprobación de Herramienta IA: [Nombre de Herramienta]" que incluya:

✓ Descripción de herramienta y su propósito
✓ Calificación en cada criterio (tabla anterior)
✓ Casos de uso autorizados (para qué tareas SÍ se puede usar; para qué tareas NO)
✓ Restricciones técnicas implementadas (si la herramienta permite)
✓ Protocolo de supervisión requerido
✓ Costo estimado
✓ Alternativas consideradas y por qué fueron descartadas
✓ Fecha de revisión próxima (mínimo anual)
✓ Firmas de aprobación (CIO, Partner responsable)

6.3. Visión Prospectiva: La IA como Amplificador del Juicio Humano, No como Sustituto del Pensamiento Crítico

La pregunta fundamental no es "¿Puede IA reemplazar abogados?" sino "¿Cómo puede IA amplificar capacidad de abogados para servir a clientes de manera mejor y más eficiente?"

ESCENARIOS FUTURO DONDE IA AGREGA VALOR:

ESCENARIO 1 - INVESTIGACIÓN DE HECHOS COMPLEJOS
Problema actual: Revisión de 50,000 documentos en discovery toma 500 horas de trabajo de abogado.
Solución con IA: IA pre-procesa documentos, identifica aquellos potencialmente relevantes. Abogado revisa 5,000 documentos pre-filtrados (50 horas). Retorno: 90% eficiencia, costo reducido, ningún documento material se pierde.

ESCENARIO 2 - SÍNTESIS DE JURISPRUDENCIA
Problema actual: Identificar todos los casos sobre tema específico (ej: prescripción bajo Convenio de Montreal) requiere búsqueda manual y lectura de decenas de sentencias.
Solución con IA (arquitectura RAG): IA recupera todos los casos relevantes, identifica argumentos comunes, sintetiza tendencia jurisprudencial. Abogado revisa síntesis (1-2 horas en lugar de 20). Retorno: eficiencia significativa, síntesis más completa, menos riesgo de omisiones.

ESCENARIO 3 - DRAFTING INICIAL DE DOCUMENTOS

Problema actual: Redacción de briefs jurídicos toma 15-20 horas de abogado experto.
Solución con IA: IA genera draft de brief basado en citas verificadas y estructura apropiada. Abogado revisa y refina (5-7 horas en lugar de 15-20). Retorno: eficiencia, abogado se enfoca en análisis conceptual no en redacción mecánica.

ESCENARIO 4 - CAPACITACIÓN ACELERADA
Problema actual: Abogado junior tarda 6 meses en dominar área de práctica específica.
Solución con IA: Junior tiene acceso a IA entrenada en casos y jurisprudencia específica como "tutor" que responde preguntas, explica conceptos. Junior aprende más rápidamente. Retorno: talento desarrollado más rápido, transiciones menos dolorosas.

LIMITACIONES A ACEPTAR:

Sin embargo, hay tareas donde IA NO puede sustituir juicio humano:
- ESTRATEGIA: Decisiones sobre qué argumentos enfatizar, cómo posicionar el caso ante tribunal específico
- NEGOCIACIÓN: Evaluación de posición del otro lado, cálculo de riesgo, decisión de settlement
- CREATIVIDAD LEGAL: Argumentos novelosos, aplicación de derecho a hechos únicos
- RELACIÓN CON CLIENTE: Consejo sobre riesgos, explicación de opciones, toma de decisión con cliente

La visión prospectiva correcta es: **IA maneja la "carga de trabajo repetitiva y verificable"; abogado se enfoca en "trabajo de alto juicio y alto valor"**.

6.4. Conclusiones Finales: Humanizar la Tecnología para Asegurar una Justicia Fiable

El caso Mata v. Avianca, Inc. (SDNY, 22 de junio de 2023) representa un momento de inflexión en la profesión jurídica. No es el primer caso donde IA generativa causó daño jurídico, pero es probablemente el primer donde un tribunal federal de prominencia reportó públicamente la alucinación, aplicó sanciones, y envió señal clara: competencia tecnológica y supervisión no son opcionales.

LECCIONES INTEGRADAS:

1. **ALUCINACIÓN NO ES "BUG" SINO "FEATURE"**

La alucinación en LLM no es defecto que será arreglado con "entrenamiento mejor". Es característica arquitectónica de sistemas generativos que priorizan fluidez sobre factualidad. Debe ser gestionado, no "resuelto".

2. **RESPONSABILIDAD PROFESIONAL NO DISMINUYE CON TECNOLOGÍA**
Delegar investigación a IA no reduce responsabilidad del abogado. De hecho, la incrementa: debo ahora ser competente no solo en ley sino en tecnología que uso para investigar.

3. **SUPERVISIÓN ES IRREDUCIBLE**
Todas las herramientas técnicas (RAG, chunking, conciencia jerárquica) son necesarias. Pero ninguna sustituye supervisión humana inteligente. El "filtro humano" es garantía de última instancia.

4. **REGULACIÓN LLEGARÁ RÁPIDAMENTE**
El Reglamento de IA de la UE (entrada en vigor agosto 2025) comienza a codificar muchas de estas lecciones. Normas en España, UK, y otras jurisdicciones seguirán. Mejor anticipar que ser sorprendido.

5. **ECONÓMICA ES CLARA**
Invertir en gobernanza de IA ahora (política de bufete, capacitación, herramientas certificadas, supervisión documentada) cuesta muy poco comparado con riesgo de sanciones judiciales, daño reputacional, y pérdida de clientes.

6. **HUMANIDAD ES EL PUNTO DE PARTIDA, NO EL FINAL**
La IA no debería deshumanizar la justicia (convertir derecho en proceso mecánico). Debería liberar abogados de tareas mecánicas para que se enfoquen en lo que cuentan: estrategia, negociación, consejo prudente, representación efectiva.

IMPLICACIÓN PARA PROFESIÓN JURÍDICA:

Los abogados que adopten IA sin gobernanza adecuada corren riesgo creciente de sanciones, pérdida de reputación, y potencial exclusión de practice (inhabilitación). Los que sí implementen gobernanza consistente ganarán ventaja competitiva: misma cantidad de trabajo en menos tiempo, con mayor exactitud, permitiendo enfoque en aspectos de mayor valor.

El mensaje: No es "IA o no IA". Es "IA gobernada responsablemente o IA indisciplinada". Y la profesión jurídica está comenzando a castigar severamente la segunda opción.

REFERENCIAS

Fuentes Citadas y Recomendadas para Consulta Adicional:

1. American Bar Association (2024). Formal Opinion 512: Use of Artificial Intelligence in Legal Practice. Emitida 29 de julio de 2024.

2. SGS (2024). White Paper: AI Governance in the Legal Industry. Enfoque en ISO/IEC 42001 y gestión de riesgos.

3. Virtuosity Legal (2025). "AI in Court: When Legal Tech Goes Rogue – Lessons from Mata v. Avianca". Análisis por Praney Goyal, abril de 2025.

4. OCDE (2023). "Advancing Accountability in AI: Governing and Managing Risks Throughout the Lifecycle for Trustworthy AI". Documento de política, febrero de 2023.

5. NIST (2024). AI Risk Management Framework (AI RMF 1.0) and Generative AI Profile. Publicado 10 de julio de 2024.

6. Eunews (2025). Reporte sobre entrada en vigor de normas de la Unión Europea sobre transparencia y obligaciones de riesgo para IA generativa. Agosto de 2025.

7. CGAE e ICAV (2024). Libro Blanco sobre Inteligencia Artificial y Abogacía. Diagnósticos y recomendaciones para el sector en España.

8. Castel, P. Kevin (Juez) (2023). Sentencia y Orden de Sanciones: Mata v. Avianca, Inc., 678 F.Supp.3d 443. Tribunal del Distrito Sur de Nueva York, 22 de junio de 2023.

9. European Data Protection Board (2024). Working Group Report on ChatGPT. Enfoque en cumplimiento RGPD y exactitud de datos, mayo de 2024.

10. Dantart, Alex (CIO LittleJohn) (2024). "Legal AI, the Truthfulness Challenge, and Hallucination Optimization via RAG". Publicado en arXiv.