Dirección:

Fernando H. Llano Alonso

José Ignacio Solar Cayón

GLOSARIO BÁSICO DE INTELIGENCIA ARTIFICIAL JURÍDICA

Autores:

Miguel Ángel Adame Martínez
Miguel de Asís Pulido
Jesús Daniel Ayllón García
Francisco Bariffi
Valeria Barone
Moisés Barrio Andrés
Vicente Bellver Capella
Pilar Blanco-Morales Limones
Roger Campione
Javier Ignacio Carbo Rubiera
John Anthony Carlin Sánchez
Thomas Casadei
Pablo Contreras Vásquez
Luis Miguel Cordero Palomo
Jorge Crego Blanco
David Antonio Cuesta Bárcena
Ibán García del Blanco
Diego García Saiz
Domingo Gómez Pérez
Eduardo González Pérez
María Isabel González-Tapia
Daniel Jove Villares
Migle Laukyte
Fernando H. Llano Alonso
Patricia Llaque Gálvez
Ignacio Lledó Benito
José Andrés López de Fez
Sandro Luce
África María Morales Moreno
Javier Antonio Nisa Ávila
Juan Antonio Ortega Ramírez
Cristián Pérez García
Antonio Enrique Pérez Luño
Stefano Pietropaoli
Juli Ponce Solé
Miguel Ángel Presno Linera
Cristina San Miguel Caso
Álvaro Sánchez Bravo
María Olga Sánchez Martínez
Antonio José Sánchez Sáez
José Ignacio Solar Cayón
Carmen Vida Carrión
David Vila-Viñas
César Villegas Delgado
Pablo Viollier Bonvin
Diana Carolina Wisner Glusko

Laborum EDICIONES

OPEN ACCESS

Temas Derecho del Trabajo y de la Seguridad Social

Editorial

Francisco Ortiz Castillo
Director editorial

Consejo Editorial

José Luis Monereo Pérez, Catedrático de Derecho del Trabajo y de la Seguridad Social, *Universidad de Granada*
Presidente del Consejo Editorial

José Luján Alcaraz, Catedrático de Derecho del Trabajo y de la Seguridad Social, *Universidad de Murcia*

Faustino Cavas Martínez, Catedrático de Derecho del Trabajo y de la Seguridad Social, *Universidad de Murcia*

María Nieves Moreno Vida, Catedrática de Derecho del Trabajo y de la Seguridad Social, *Universidad de Granada*

Inmaculada Ballester Pastor, Catedrática de Derecho del Trabajo y de la Seguridad Social, *Universitat Jaume I*

Cristina Sánchez-Rodas Navarro, Catedrática de Derecho del Trabajo y de la Seguridad Social, *Universidad de Sevilla*

Guillermo Rodríguez Iniesta, Catedrático de Derecho del Trabajo y de la Seguridad Social, *Universidad de Murcia*
Director General de Publicaciones

Belén del Mar López Insua, Catedrática (acreditada) de Derecho del Trabajo y de la Seguridad Social, *Universidad de Granada*
Directora de la Colección Trabajos de Investigación

Consejo Científico

Juan Carlos Álvarez Cortés, Catedrático de Derecho del Trabajo y de la Seguridad Social, *Universidad de Málaga*

Alberto Arufe Varela, Catedrático de Derecho del Trabajo y de la Seguridad Social, *Universidad da Coruña*

Jaime Cabeza Pereiro, Catedrático de Derecho del Trabajo y de la Seguridad Social, *Universidad de Vigo*

María Teresa Díaz Aznarte, Catedrática de Derecho del Trabajo y de la Seguridad Social, *Universidad de Granada*

Juan José Fernández Domínguez, Catedrático de Derecho del Trabajo y de la Seguridad Social, *Universidad de Valladolid*

Marina Fernández Ramírez, Catedrática de Derecho del Trabajo y de la Seguridad Social, *Universidad de Málaga*

Juan Gorelli Hernández, Catedrático de Derecho del Trabajo y de la Seguridad Social, *Universidad de Huelva*

Juan Antonio Maldonado Molina, Catedrático de Derecho del Trabajo y de la Seguridad Social, *Universidad de Granada*

Jesús Martínez Girón, Catedrático de Derecho del Trabajo y de la Seguridad Social, *Universidad da Coruña*

Carolina Martínez Moreno, Catedrática de Derecho del Trabajo y de la Seguridad Social, *Universidad de Oviedo*

Jesús Mercader Uguina, Catedrático de Derecho del Trabajo y de la Seguridad Social, *Universidad Carlos III de Madrid*

Olimpia Molina Hermosilla, Catedrática de Derecho del Trabajo y de la Seguridad Social, *Universidad de Jaén*

Antonio Ojeda Avilés, Catedrático de Derecho del Trabajo y de la Seguridad Social, *Universidad de Sevilla*

Margarita Ramos Quintana, Catedrática de Derecho del Trabajo y de la Seguridad Social, *Universidad de La Laguna*

Pilar Rivas Vallejo, Catedrática de Derecho del Trabajo y de la Seguridad Social, *Universidad de Barcelona*

Susana Rodríguez Escanciano, Catedrática de Derecho del Trabajo y de la Seguridad Social, *Universidad de León*

Carmen Sáez Lara, Catedrática de Derecho del Trabajo y de la Seguridad Social, *Universidad de Córdoba*

Carmen Salcedo Beltrán, Catedrática de Derecho del Trabajo y de la Seguridad Social, *Universitat de València*

Antonio V. Sempere Navarro, Magistrado del Tribunal Supremo, Catedrático de Derecho del Trabajo y de la Seguridad Social

Ángel Luis de Val Tena, Catedrático de Derecho del Trabajo y de la Seguridad Social, *Universidad de Zaragoza*

Arántzazu Vicente Palacio, Catedrática de Derecho del Trabajo y de la Seguridad Social, *Universitat Jaume I*

Francisco Vila Tierno, Catedrático de Derecho del Trabajo y de la Seguridad Social, *Universidad de Málaga*

Dirección:
Fernando H. Llano Alonso
José Ignacio Solar Cayón

GLOSARIO BÁSICO DE INTELIGENCIA ARTIFICIAL JURÍDICA

Autores:

Miguel Ángel Adame Martínez
Miguel de Asís Pulido
Jesús Daniel Ayllón García
Francisco Bariffi
Valeria Barone
Moisés Barrio Andrés
Vicente Bellver Capella
Pilar Blanco-Morales Limones
Roger Campione
Javier Ignacio Carbo Rubiera
John Anthony Carlin Sánchez
Thomas Casadei
Pablo Contreras Vásquez
Luis Miguel Cordero Palomo
Jorge Crego Blanco
David Antonio Cuesta Bárcena
Ibán García del Blanco
Diego García Saiz
Domingo Gómez Pérez
Eduardo González Pérez
María Isabel González-Tapia
Daniel Jove Villares
Migle Laukyte
Fernando H. Llano Alonso
Patricia Llaque Gálvez
Ignacio Lledó Benito
José Andrés López de Fez
Sandro Luce
África María Morales Moreno
Javier Antonio Nisa Ávila
Juan Antonio Ortega Ramírez
Cristián Pérez García
Antonio Enrique Pérez Luño
Stefano Pietropaoli
Juli Ponce Solé
Miguel Ángel Presno Linera
Cristina San Miguel Caso
Álvaro Sánchez Bravo
María Olga Sánchez Martínez
Antonio José Sánchez Sáez
José Ignacio Solar Cayón
Carmen Vida Carrión
David Vila-Viñas
César Villegas Delgado
Pablo Viollier Bonvin
Diana Carolina Wisner Glusko

Laborum
EDICIONES

Esta publicación es parte del Proyecto de I+D+i “Inteligencia artificial jurídica y Estado de Derecho” [PID2022–139773OB-I00], financiado por MICIU/AEI/10.13039/501100011033 y por FEDER, UE, y del Proyecto de I+D+i DENORIA [PID2023-146621OB-C21].

Edita:
Ediciones Laborum, S.L.
Avda. Gutiérrez Mellado, 9 - Planta 3ª, Oficina 21
30008 Murcia
Tel.: 968 24 10 97
E-mail: laborum@laborum.es
www.laborum.es

ISBN Digital: 979-13-88025-85-3

ISBN Papel: 979-13-88025-84-6
D.L.: MU 1159-2026

© Copyright de la edición, Ediciones Laborum, 2026
© Copyright del texto, sus respectivos autores, 2026

Ediciones Laborum, S.L. no comparte necesariamente los criterios manifestados por los autores en el trabajo publicado.

La información contenida en esta publicación constituye únicamente, y salvo error u omisión involuntarios, la opinión de su autor con arreglo a su leal saber y entender, opinión que subordinan tanto a los criterios que la jurisprudencia establezca, como a cualquier otro criterio mejor fundado.

Ni el editor, ni el autor, pueden responsabilizarse de las consecuencias, favorables o desfavorables, de actuaciones basadas en las opiniones o informaciones contenidas en esta publicación.

Reconocimiento - No comercial - Sin obra derivada (BY-NC-ND): El material puede ser distribuido, copiado y exhibido por terceros si se muestra en los créditos. No se puede obtener ningún beneficio comercial. No se pueden realizar obras derivadas.

Tabla de Contenido

Prólogo 13

Primera Parte.
LA INTELIGENCIA ARTIFICIAL

I. IA y entorno tecnológico 23

Máquina pensante, *por I. Lledó Benito* 23

Inteligencia artificial estrecha (*Narrow AI* / *Weak AI*), *por J. A. Ortega Ramírez*. 24

Inteligencia artificial general (AGI), *por J. A. Ortega Ramírez*. 25

Superinteligencia Artificial (ASI), *por J. A. López de Fez* 25

Inteligencia artificial simbólica, *por M. de Asís Pulido* 27

Ontología, *por M. de Asís Pulido y C. Pérez García* 29

Agente de IA (*Agentic AI*), *por M. Laukyte* 29

Big Data, *por D. García Saiz* 31

– Minería de datos (*data mining*)

Blockchain, *por D. García Saiz* 33

Realidad virtual, *por P. Llaque Gálvez* 36

Metaverso, *por I. Lledó Benito*. 37

Gemelo digital, *por P. Llaque Gálvez*. 39

Neurotecnologías, *por P. Llaque Gálvez*. 40

Inteligencia Biológica Sintética (*Synthetic Biological Intelligence*, SBI), *por J. A. Nisa Ávila* 42

Computación biocibernética, *por J. A. Nisa Ávila* 44

Cómputo molecular, *por J. A. Nisa Ávila*. 45

Ciberseguridad, *por D. Gómez Pérez* 45

II. Fundamentos tecnológicos de la IA . . . 47
Algoritmo, *por D. García Saiz*. . . . 47
Algoritmo evolutivo, *por J. I. Carbo Rubiera y C. Pérez García* . . . 49
Algoritmo genético, *por J. I. Carbo Rubiera y C. Pérez García* . . . 51
Función de aptitud, *por J. I. Carbo Rubiera y C. Pérez García* . . . 52
Código fuente, *por D. García Saiz* . . . 52
– Código abierto
Aprendizaje automático (*Machine Learning*), *por M. de Asís Pulido* . . . 55
Aprendizaje supervisado, *por M. de Asís Pulido* . . . 56
Aprendizaje no supervisado, *por M. de Asís Pulido* . . . 58
Aprendizaje autosupervisado, *por M. de Asís Pulido* . . . 59
Aprendizaje por refuerzo, *por M. de Asís Pulido* . . . 60
Aprendizaje profundo (*Deep Learning*), *por M. de Asís Pulido* . . . 61
Redes neuronales, *por C. Vida Carrión* . . . 62
Redes neuronales convolucionales (*Convolutional Neuronal Network*, CNN), *por J. A. Nisa Ávila* . . . 63
Redes generativas antagónicas (*Generative Adversarial Network*, GAN), *por J. A. Nisa Ávila* . . . 64
Procesamiento del lenguaje natural *(Natural Language Processing*, NLP), *por C. Vida Carrión* . . . 64
Minería de argumentos (*Argument Mining*), *por C. Vida Carrión* . . . 64
Black Box, *por C. Pérez García*. . . . 65
White Box, *por C. Pérez García* . . . 66
Capa, *por J. Nisa Ávila* . . . 66
Nodo, *por J. A. Nisa Ávila* . . . 67
Pesos, *por J. A. Nisa Ávila*. . . . 68
Vectores de posición, *por J. A. Nisa Ávila* . . . 69
Autocodificador (*Autoencoder*), *por J. Nisa Ávila* . . . 69
Datos, *por D. Jove Villares* . . . 70
Datos personales, *por D. Jove Villares* . . . 72
Datos de entrenamiento, validación y prueba, *por D. Jove Villares* . . . 73
Datos de entrada (*inputs*) y de salida (*outputs*), *por D. Jove Villares* . . . 75
Datos sintéticos (*Synthetic data*), *por M. A. Adame Martínez* . . . 76
Metadato, *por J. A. Nisa Ávila* . . . 78
Web scraping, *por P. Viollier Bonvin* . . . 78

Servidor, *por J. A. Nisa Ávila* 80

III. IA generativa y LLMs 80

Modelo generativo / Sistema de IA generativa, *por J. A. Ortega Ramírez*. 80

Modelo fundacional (*Foundation Model*), *por J. A. Ortega Ramírez* 82

Modelo multimodal, *por J. A. Ortega Ramírez* 82

– Sistema multimodal

Modelo de lenguaje de gran tamaño (*Large Language Model*, LLM), *por E. González Pérez*. 82

– Modelo de lenguaje pequeño (*Small Language Model*, SLM)

– Arquitectura de transformadores

– Modelo de IA de uso general

– Modelo de IA específico de dominio

Modelo preentrenado, *por J. A. Ortega Ramírez*. 84

– Preentrenamiento

Ajuste fino (*Fine tuning*), *por E. González Pérez* 85

RAG (*Retrieval-Augmented Generation*), *por E. González Pérez* 87

Alineamiento posterior al entrenamiento (*Post-training Alignment*), *por J. A. Ortega Ramírez* 89

Uso de herramientas (*Tool Use*), *por J. A. Ortega Ramírez* 89

Agente con herramientas, *por J. A. Ortega Ramírez* 90

Llamada a funciones (*Function calling*), *por J. A. Ortega Ramírez* 91

Inferencia, *por J. A. Ortega Ramírez.* 92

Generación condicionada, *por J. A. Ortega Ramírez* 92

Multimodalidad generativa, *por J. A. Ortega Ramírez* 92

Sampling o Muestreo, *por J. A. Ortega Ramírez* 93

Temperatura, *por J. A. Ortega Ramírez*. 94

Top-K / Top-P, *por J. A. Ortega Ramírez.* 94

Memoria conversacional, *por J. A. Ortega Ramírez*. 95

Token, *por J. A. Ortega Ramírez* 96

– Tokenización

Ventana de contexto (*Context window*), *por M. A. Adame Martínez*. 97

Token de contexto, *por J. A. Ortega Ramírez* 98

Embeddings / Incrustaciones vectoriales, *por J. A. Ortega Ramírez* 99

Espacio latente, *por J. A. Ortega Ramírez* 99

Destilación de modelos, *por J. A. Ortega Ramírez* 99

Cuantización, *por J. A. Ortega Ramírez* 100
Orquestación, *por J. A. Ortega Ramírez* 100
Prompt, *por J. A. López de Fez* 101
– Ingeniería de *prompts*
Alucinación, *por C. Pérez García* 103
Razonamiento en cadena como fenómeno técnico, *por J. A. Ortega Ramírez.* 104
Salida estructurada, *por J. A. Ortega Ramírez* 105
Medios sintéticos (*Synthetic media*), *por J. A. Ortega Ramírez* 105
Procedencia del contenido (*Provenance*), *por J. A. Ortega Ramírez* 106

Segunda Parte.
ECOSISTEMA ÉTICO Y JURÍDICO DE LA IA

I. Ética de la IA 109
Gobernanza de la IA, *por C. Villegas Delgado* 109
Ética de la IA, *por F. H. Llano Alonso* 111
Principio de no malevolencia (*Non-maleficence*), *por F. H. Llano Alonso* 112
Principio de beneficencia (*Beneficence*), *por F. H. Llano Alonso* 114
Principio de justicia (*Justice / Fairness*), *por F. H. Llano Alonso* 115
Principio de autonomía (*Autonomy*), *por F. H. Llano Alonso* 117
Principio de explicabilidad (*Explicability*), *por F. H. Llano Alonso* 118
Justicia algorítmica, *por A. J. Sánchez Sáez* 120
Alineamiento de la IA (*AI Alignment*), *por F. H. Llano Alonso* 121
Alineamiento engañoso (*Deceptive Alignment*), *por J. A. López de Fez* 123
Guardarraíles, *por M. de Asís Pulido* 125
Ética por diseño (*Ethics-by-Design*), *por J. A. López de Fez* 126
Human-in-the-loop (HITL), *Human-on-the-loop* (HOTL), *Human-out-of-the-loop* (HOOTL), *por J. A. López de Fez* 128
Gubernamentalidad algorítmica o algocracia, *por S. Luce* 129
Transhumanismo, *por F. H. Llano Alonso.* 131
Posthumanismo, *por V. Bellver Capella* 133
Antropodatismo, *por L. M. Cordero Palomo* 134
Sesgo positivista, *por C. Pérez García* 136
Armas autónomas (*Lethal Autonomous Weapon Systems*), *por R. Campione* 138
Cuidados e inteligencia artificial, *por D. Vila-Viñas* 140
Nuevas tecnologías y doctrina social de la Iglesia, *por A. Sánchez Bravo* 141

Inteligencia artificial y doctrina social de la Iglesia, *por A. Sánchez Bravo* 143

II. Derecho de la IA 145

1. Conceptos doctrinales generales 145

Libertad informática, *por A. E. Pérez Luño* 145

Cumplimiento por diseño (*Compliance-by-Design*), *por J. A. López de Fez* 146

Decisión algorítmica, *por S. Pietropaoli* 148

Personalidad algorítmica, *por C. Pérez García* 150

Responsabilidad algorítmica, *por C. Pérez García* 151

Auditoría algorítmica, *por I. Lledó Benito* 152

Transparencia algorítmica, *por D. C. Wisner Glusko* 153

Brechas digitales, *por T. Casadei (trad. de C. Pérez García)* 155

Discriminación algorítmica, *por V. Barone* 157

Sesgo (*Bias*), *por D. C. Wisner Glusko* 159

Sesgo de automatización, *por D. C. Wisner Glusko* 160

Sesgo de confirmación, *por M. O. Sánchez Martínez* 162

Propiedad intelectual e IA, *por M. A. Adame Martínez* 164

Debido proceso tecnológico (*Technological due process*), *por J. I. Solar Cayón* 165

2. Conceptos legales 167

A) RIA y Legislación europea

Sistema de IA, *por M. Barrio Andrés* 167

Sistema de IA de alto riesgo, *por J. I. Solar Cayón* 169

Proveedor, *por J. I. Solar Cayón* 171

Responsable del despliegue, *por J. I. Solar Cayón* 173

Garantía del cumplimiento del Derecho, *por J. I. Solar Cayón* 174

– Autoridad garante del cumplimiento del Derecho

Reconocimiento facial, *por J. A. Carlin Sánchez* 175

Sistema de identificación biométrica, *por J. A. Carlin Sánchez* 176

Verificación biométrica, *por J. A. Carlin Sánchez* 177

Sistema de categorización biométrica, *por J. A. Carlin Sánchez* 177

Sistema de reconocimiento de emociones, *por J. A. Carlin Sánchez* 178

Evaluación de la conformidad, *por I. Lledó Benito* 179

Organismo de evaluación de la conformidad, *por I. Lledó Benito* 180

Autoridad notificante, *por I. Lledó Benito* 181

Organismo notificado, *por I. Lledó Benito* 182

Supervisión humana, *por F. H. Llano Alonso* . . . 183
Ultrasuplantación (*Deepfake*), *por J. D. Ayllón García* . . . 184
Sandbox o Espacio controlado de pruebas, *por P. Contreras Vásquez* . . . 186
Perfilado, *por F. Bariffi* . . . 188
Alfabetización digital, *por I. García del Blanco* . . . 190
Derecho a explicación, *por M. A. Adame Martínez* . . . 193
Marcado de agua (*Watermarking*) y Etiquetado (*Labelling*) de contenidos IA, *por M. A. Adame Martínez* . . . 194
Tutela judicial algorítmica transfronteriza, *por P. Blanco-Morales Limones* . . . 196
Lex fori digital, *por P. Blanco-Morales Limones* . . . 197
Identidad digital transfronteriza, *por P. Blanco-Morales Limones* . . . 199
Acceso algorítmico al Derecho extranjero, *por P. Blanco-Morales Limones* . . . 200
Razonamiento algorítmico transfronterizo, *por P. Blanco-Morales Limones* . . . 202
Pericial algorítmica transfronteriza, *por P. Blanco-Morales Limones* . . . 204
Interpretación algorítmica transfronteriza, *por P. Blanco-Morales Limones* . . . 206
Traducción algorítmica, *por P. Blanco-Morales Limones* . . . 207
Confianza mutua algorítmica, *por P. Blanco-Morales Limones* . . . 209
Asimetría cognitiva algorítmica, *por P. Blanco-Morales Limones* . . . 212
Contrato de IA transfronterizo, *por P. Blanco-Morales Limones* . . . 214
Privacidad algorítmica transfronteriza, *por P. Blanco-Morales Limones* . . . 216
B) Legislación española
Decisión administrativa automatizada, *por D. A. Cuesta Bárcena* . . . 218
Reserva de humanidad, *por J. Ponce Solé* . . . 220
Derechos digitales, *por D. A. Cuesta Bárcena* . . . 222
Actuaciones procesales automatizadas, *por C. San Miguel Caso* . . . 224
– Actuaciones proactivas
Actuaciones judiciales asistidas, *por C. San Miguel Caso* . . . 225

Tercera Parte.
LA IA JURÍDICA Y SUS EFECTOS

I. El Derecho en la era de la IA . . . 231
Internet y Derecho, *por A. E. Pérez Luño* . . . 231
Ciberciudadanía, *por A. E. Pérez Luño* . . . 232
Nuevas tecnologías, Derecho y libertades, *por A. E. Pérez Luño* . . . 234
Constitucionalismo digital, *por M. A. Presno Linera* . . . 235

Neuroderechos, *por F. H. Llano Alonso* . . . 237
Tecno-regulación, *por J. Crego Blanco* . . . 238
Derecho Cuántico, *por J. A. Nisa Ávila* . . . 240

II. Aplicación de la IA en el dominio jurídico . . . 240
Informática jurídica, *por A. E. Pérez Luño* . . . 240
Informática jurídica documental, *por A. E. Pérez Luño* . . . 241
Informática jurídica decisional, *por A. E. Pérez Luño* . . . 242
Informática jurídica de gestión, *por A. E. Pérez Luño* . . . 243
Legal Tech, *por M. Barrio Andrés* . . . 244
LawTech, *por J. I. Solar Cayón* . . . 246
RegTech, *por J. I. Solar Cayón* . . . 247
Inteligencia artificial jurídica, *por J. I. Solar Cayón* . . . 248
Sistemas expertos jurídicos, *por J. I. Solar Cayón* . . . 251
E-discovery (Descubrimiento electrónico), *por M. A. Adame Martínez* . . . 252
Codificación predictiva (*Predictive Coding* o *Technology-Assisted Review*), *por J. I. Solar Cayón* . . . 253
Sistemas de análisis predictivo, *por A. M. Morales Moreno* . . . 255
Sistema de evaluación de riesgo delictivo, *por M. I. González-Tapia* . . . 257
Sistemas de negociación y resolución automatizada de disputas, *por J. I. Solar Cayón* 259
Sistemas de lectura automática y análisis de contratos, *por J. I. Solar Cayón* . . . 260
Diligencia debida automatizada (*Automated due diligence*), *por M. A. Adame Martínez* 262
Modelo de lenguaje legal, *por J. I. Solar Cayón* . . . 263
Justicia digital, *por J. Nieva-Fenoll* . . . 265
Tribunal en línea, *por J. I. Solar Cayón* . . . 267
Codificación computacional del Derecho (*Law as Code*), *por J. I. Solar Cayón* . . . 268
Smart contract, *por R. Campione* . . . 270

III. Efectos de la IA jurídica . . . 272
Deber de competencia tecnológica, *por J. I. Solar Cayón* . . . 272
Comoditización de los servicios legales, *por J. I. Solar Cayón* . . . 274
– Desintermediación
Proveedor alternativo de servicios legales, *por J. I. Solar Cayón* . . . 275
Legal outsourcing, *por J. I. Solar Cayón* . . . 277
– *Legal process outsourcers (*LPO)
– *Offshoring*

Alternative Business Structure, *por J. I. Solar Cayón* 278
– *Incorporated Legal Practices*
Mercado jurídico virtual (*Online Legal Marketplace*), *por J. I. Solar Cayón.* 280
Litigación basada en datos (*Data-driven lawyering*), *por A. J. Sánchez Sáez* 282
Inversión de la carga de la prueba algorítmica, *por A. J. Sánchez Sáez* 283
Policía predictiva, *por C. San Miguel Caso* 285
Insurtech, *por I. Lledó Benito* 286
IV. Nuevos perfiles profesionales 287
Analista de procesos legales (*Legal Process Analyst*), *por J. I. Solar Cayón* 287
Director de tecnología legal (*Chief Legal Technology Officer*), *por J. I. Solar Cayón* . . 288
– Director de innovación legal (*Chief Legal Innovation Officer*)
Director de IA aplicada (*Director of Applied AI*, *Head of Machine Learning Practice, Head of Legal Technology Adoption*), *por J. I. Solar Cayón*. 289
Científico de datos jurídicos (*Legal Data Scientist*), *por J. I. Solar Cayón* 290
– Analista de datos jurídicos (*Legal Data Analyst*)
Ingeniero de conocimientos jurídicos (*Legal Knowledge Engineer*), *por J. I. Solar Cayón* 290
Ingeniero de *prompts* jurídicos (*Legal Prompt Engineer*), *por J. I. Solar Cayón* 292
Responsable de gestión de proyectos legales (*Legal Project Manager*), *por J. I. Solar Cayón* 292

Índice Analítico 295

Prólogo

Pocas veces en la larga historia del Derecho los juristas se han enfrentado a una transformación tan profunda y tan veloz como la que hoy provoca la IA. Otras revoluciones técnicas marcaron también un antes y un después en el mundo del Derecho: la imprenta, que multiplicó el acceso a las leyes; la codificación, que dio forma sistemática al ordenamiento; o la constitucionalización de los derechos, que situó a la persona en el centro del edificio jurídico. Ninguna de ellas, sin embargo, aunó la rapidez, la ubicuidad y la opacidad que caracterizan a la actual mutación tecnológica. El Derecho siempre ha respondido al cambio técnico con un cierto retraso, y ese desfase, que en otras épocas fue tolerable, amenaza ahora con convertirse en un riesgo estructural para la garantía de los derechos y para la propia inteligibilidad de las normas.

Lo que hace singular a esta tecnología no es solo su difusión vertiginosa, sino el modo en que desafía algunas de las categorías sobre las que el Derecho se ha asentado durante siglos. Los sistemas más avanzados de IA operan con frecuencia como verdaderas cajas negras, cuyos procesos internos resultan difíciles de reconstruir incluso para quienes los han diseñado, lo que pone en aprietos a un ordenamiento que exige motivar las decisiones y explicar sus razones. Estos sistemas exhiben, además, grados crecientes de autonomía, hasta el punto de adoptar comportamientos que sus creadores no previeron, lo que problematiza nociones tan arraigadas como la causalidad, la culpa o la imputación de responsabilidad. Y aprenden de datos que arrastran consigo las desigualdades del pasado, de manera que pueden perpetuar o incluso amplificar discriminaciones bajo una apariencia de neutralidad técnica. Ante semejantes desafíos, el Derecho no puede limitarse a aplicar sus viejas herramientas sin más; necesita primero comprender la naturaleza del fenómeno, y para comprenderlo necesita, antes que nada, nombrarlo con propiedad.

Quien se asoma por vez primera al mundo de la IA advierte enseguida una dificultad que es, antes que técnica, lingüística. Tecnólogos y juristas hablan de una misma realidad con palabras distintas, o bien emplean las mismas palabras atribuyéndoles significados diferentes. Términos que en el laboratorio designan operaciones matemáticas precisas adquieren, al cruzar el umbral de los tribunales y de los despachos, una resonancia normativa que no siempre se corresponde con su sentido original. Y a la inversa: categorías jurídicas decantadas durante siglos resultan de pronto insuficientes para nombrar fenómenos que carecen de precedente. De esa confusión nacen equívocos que no son meramente académicos, pues tienen consecuencias prácticas muy reales en la regulación, en la contratación y en la resolución de conflictos.

El presente *Glosario básico de inteligencia artificial jurídica* nace, precisamente, para tender un puente entre esos dos lenguajes. Se trata, hasta donde alcanzamos a saber, del primer trabajo de esta especialidad en la doctrina jurídica española, y con él hemos querido ofrecer a la comunidad jurídica un vocabulario común, razonado y accesible. No aspira esta obra a ser un catálogo exhaustivo de los términos que ha generado la convergencia entre el Derecho y las tecnologías digitales -empresa que sería tan inabarcable como efímera, a la vista de la velocidad con que el campo evoluciona-, sino a recoger y explicar las voces de IA jurídica más habituales, aquellas que con mayor frecuencia emplean por igual tecnólogos y juristas. Nombrar bien es la primera condición para regular bien y para juzgar con acierto; y un glosario que aclare qué significa, por ejemplo, un sistema de IA de alto riesgo -que no es una metáfora, sino una categoría jurídica de la que penden obligaciones precisas- presta a la práctica un servicio nada desdeñable.

No es indiferente que la obra haya adoptado la forma de un glosario, y no la de un tratado o un manual al uso. El género tiene virtudes propias que lo hacen especialmente idóneo para el momento que vivimos. Un glosario fija el significado de las voces, las ordena y las pone al alcance de la mano, de modo que puede consultarse con agilidad cuando la duda surge en mitad de un razonamiento, de un dictamen o de una vista. Frente a la exposición sistemática, que exige una lectura continuada, la entrada del glosario ofrece una respuesta acotada y suficiente, sin renunciar por ello al rigor. Y, sobre todo, democratiza el acceso al conocimiento: quien no es especialista encuentra en él una puerta de entrada, y quien lo es halla un instrumento de precisión terminológica. En un dominio en el que la imprecisión del lenguaje engendra con tanta facilidad la confusión del pensamiento, contar con definiciones claras y compartidas es ya, de suyo, una conquista.

Conviene insistir en que esta exigencia de precisión no obedece a un escrúpulo puramente académico. De la recta comprensión de los conceptos

dependen consecuencias muy concretas para las personas. Cuando una decisión administrativa se adopta de forma automatizada, cuando un sistema algorítmico interviene en la concesión de un crédito o en la valoración de un riesgo, cuando una herramienta predictiva orienta la actuación de la policía o la decisión de un tribunal, lo que está en juego son derechos fundamentales: la igualdad y la no discriminación, la tutela judicial efectiva, la protección de los datos personales o la propia presunción de inocencia. Distinguir con nitidez entre el perfilado y la decisión automatizada, entre el sesgo del dato y el sesgo del modelo, o entre la asistencia y la sustitución del juez por la máquina, no es un ejercicio de pedantería conceptual, sino el presupuesto mismo de una protección eficaz. Las palabras, en Derecho, nunca son inocentes.

Esta obra ha sido auspiciada por la Facultad de Derecho de la Universidad de Sevilla y se inscribe en una línea de trabajo que el Decanato viene impulsando desde hace algunos años, en favor de incorporar la IA a la formación jurídica y de preparar a nuestros estudiantes para ejercer como futuros profesionales del Derecho en la sociedad de las nuevas tecnologías. No se trata de una iniciativa aislada, sino de un eslabón más en una cadena de esfuerzos que han ido convirtiendo a esta Facultad en un punto de referencia dentro del panorama universitario español en lo que atañe al estudio del Derecho de la IA.

Conviene recordar, a este respecto, que la Facultad de Derecho de la Universidad de Sevilla ha sido pionera al ofrecer entre sus titulaciones de posgrado el primer máster oficial en Ética y Derecho de la Inteligencia Artificial, un programa que conjuga deliberadamente las dos dimensiones -la ética y la jurídica- que no pueden disociarse cuando se trata de gobernar una tecnología llamada a incidir de manera tan directa sobre la vida de las personas. A ello se suma la consolidación de la Semana de la Inteligencia Artificial, que este año celebra su cuarta edición y que se ha afirmado ya como un foro de encuentro entre el mundo académico, el profesional y el institucional, en el que comparten reflexión filósofos del Derecho, especialistas en distintas ramas del ordenamiento, ingenieros, magistrados y representantes de la sociedad civil. Estas iniciativas no son meros ornamentos académicos: responden a la convicción de que la universidad tiene el deber de anticiparse a los retos de su tiempo y de formar juristas capaces de comprender el mundo que les ha tocado regular.

Detrás de estas iniciativas late una convicción formativa que conviene hacer explícita. Hemos querido que la ética y el Derecho caminen juntos, porque entendemos que separar ambas dimensiones empobrece el análisis y debilita la respuesta. No basta con preguntarse si una determinada aplicación de la IA es conforme a la legalidad vigente; es preciso interrogarse también acerca de si resulta justa, de si respeta la autonomía de las personas, de si distribuye con equidad sus beneficios y sus cargas, y de si puede explicarse y rendir cuentas ante quienes se

ven afectados por ella. Por eso el máster que ofrece nuestra Facultad lleva en su propio nombre ese binomio inseparable, y por eso la Semana de la Inteligencia Artificial ha procurado, edición tras edición, que el debate jurídico no se desligue nunca de la reflexión ética. Formar juristas competentes en lo técnico pero ciegos en lo valorativo sería tan estéril como peligroso.

La dirección compartida de este glosario por dos cátedras de Filosofía del Derecho, la de la Universidad de Sevilla y la de la Universidad de Cantabria, expresa también algo más que una colaboración personal entre quienes esto firmamos. Refleja la conciencia de que el estudio de la IA jurídica difícilmente puede prosperar desde el aislamiento de una sola escuela o de una sola sede, y de que la filosofía del Derecho -esa disciplina que se ocupa de los fundamentos, de los conceptos y de los valores del ordenamiento- está particularmente llamada a ordenar el vértigo conceptual que la tecnología introduce. Pero la tarea desbordaba con mucho las fuerzas de dos directores, y por eso quisimos abrir la obra a un amplio elenco de colaboradores.

En efecto, este libro reúne el trabajo de cerca de medio centenar de especialistas en IA jurídica, procedentes de universidades e instituciones españolas, europeas y americanas. Junto a profesores de numerosas universidades españolas -Sevilla, Cantabria, Oviedo, Valencia, Córdoba, A Coruña, la Carlos III de Madrid, la UNED, la Pompeu Fabra, la de Barcelona o la CEU Fernando III, entre otras-, han prestado su saber colegas de prestigiosas universidades europeas, como las italianas de Módena y Reggio Emilia, Florencia o Salerno, o la Humboldt de Berlín, así como de instituciones americanas, desde la Universidad de Buenos Aires hasta la Universidad Central de Chile. A esta nómina universitaria se añaden profesionales del Derecho llamados a aportar la mirada de la práctica: abogados en ejercicio, un letrado del Consejo de Estado, directores de cátedras de tecnología legal y fundadores de empresas del sector, e incluso quien participó, como miembro del Parlamento Europeo, en la gestación de la normativa europea sobre IA.

Esa pluralidad de voces no es un mero dato de cortesía editorial, sino que encierra una tesis sobre el propio objeto de estudio. El Derecho de la IA no puede construirse desde una única disciplina, porque ningún saber agota por sí solo la complejidad del fenómeno. Por eso hemos querido que en estas páginas dialoguen filósofos del Derecho con civilistas y administrativistas, penalistas con procesalistas, internacionalistas con constitucionalistas, y juristas con ingenieros informáticos y científicos de datos. De ese concurso de perspectivas surge una imagen del problema más fiel y matizada que la que cualquiera de ellas podría ofrecer en solitario. La IA nos obliga a pensar en común; este glosario es, también en su factura, un ejercicio de pensamiento compartido.

La presencia de autores europeos y americanos junto a los españoles añade, por lo demás, una dimensión comparada que estimamos especialmente valiosa. La IA es un fenómeno global, indiferente a las fronteras, y los problemas que plantea se repiten con notable semejanza en latitudes muy diversas; pero las respuestas que ensayan los distintos ordenamientos no siempre coinciden, y de su contraste puede el jurista extraer enseñanzas provechosas. Conocer cómo se afronta la justicia algorítmica al otro lado del Atlántico, o cómo dialogan la doctrina italiana y la española en torno a la decisión automatizada, permite mirar nuestro propio Derecho con una perspectiva más amplia y crítica, libre del provincianismo en que tan fácilmente se incurre cuando se estudia un solo sistema. Esta vocación abierta y dialogante nos parece, además, especialmente congruente con el espíritu de una tecnología que no entiende de aduanas.

Antes de adentrarse en el contenido, acaso interese al lector conocer el modo en que la obra se ha concebido. Cada voz aparece firmada por su autor o autores, que asumen la responsabilidad de su tratamiento, lo que permite que el especialista en cada materia despliegue su saber con libertad y profundidad. Las entradas no se limitan a ofrecer una definición escueta, sino que sitúan cada término en su contexto, advierten de sus matices y, cuando procede, lo conectan con otras voces afines, de manera que el conjunto puede recorrerse también como una red de remisiones. Hemos cuidado, asimismo, de recoger junto a la denominación española las expresiones inglesas de uso más extendido, pues buena parte del vocabulario nace en esa lengua y conviene que el jurista las reconozca sin vacilación. Y, sobre todo, hemos procurado que cada entrada ofrezca una doble lectura: la que describe el fenómeno técnico y la que examina sus implicaciones jurídicas, porque solo de esa conjunción puede surgir una comprensión cabal.

Conviene que el lector conozca, siquiera sumariamente, cómo se organiza el contenido de la obra, que se distribuye en tres grandes partes complementarias. La primera de ellas se dedica a la IA considerada en sí misma, es decir, a sus fundamentos técnicos. Pues estamos persuadidos de que el jurista no puede regular ni juzgar con rigor aquello que no comprende, y de que un mínimo de alfabetización tecnológica se ha convertido en presupuesto del buen ejercicio profesional. Esta primera parte conduce al lector desde las nociones más generales -la vieja aspiración a la máquina pensante, la distinción entre la IA estrecha, la general y la superinteligencia- hasta los fundamentos que hacen posible la tecnología actual: el algoritmo, el aprendizaje automático en sus diversas modalidades, las redes neuronales y el aprendizaje profundo.

De ahí se transita hacia el giro generativo que ha popularizado la IA en los últimos años. El glosario explica con claridad qué es un modelo de lenguaje de gran tamaño, qué se entiende por modelo fundacional o multimodal, y desgrana el vocabulario que ha pasado a formar parte de la conversación cotidiana: el *token*

y la tokenización, los *embeddings*, la ventana de contexto, el *prompt* y su ingeniería, el ajuste fino, la inferencia o el fenómeno de la alucinación. Son términos que el jurista escucha a diario y que necesita comprender no por afán de erudición técnica, sino porque de su recta inteligencia depende la calidad de los juicios normativos que sobre ellos hayan de construirse.

La segunda parte aborda el ecosistema ético y jurídico de la IA, y constituye, en cierto modo, el corazón de la obra. Se abre con la reflexión ética, articulada en torno a los grandes principios que la comunidad internacional ha ido decantando para orientar el desarrollo de estas tecnologías: la no maleficencia, la beneficencia, la justicia, la autonomía y la explicabilidad. A ellos se añaden conceptos hoy imprescindibles, como el alineamiento de la IA con los valores humanos, la gobernanza de los sistemas o las cuestiones que suscitan corrientes como el transhumanismo y el posthumanismo, que interpelan a la propia idea de dignidad sobre la que se asienta nuestro Derecho.

Sigue a continuación el tratamiento propiamente jurídico, que distingue con acierto entre los conceptos doctrinales generales y los conceptos estrictamente legales. Entre los primeros figuran nociones tan decisivas como la decisión algorítmica, la responsabilidad y la transparencia algorítmicas, la discriminación y el sesgo, o el llamado debido proceso tecnológico. Entre los segundos, el glosario ofrece una guía a través del Reglamento europeo de IA -explicando qué es un sistema de IA, qué un sistema de alto riesgo, quién es el proveedor y quién el responsable del despliegue, o cómo se regulan la identificación biométrica, la ultrasuplantación o los espacios controlados de pruebas- y se detiene asimismo en las novedades de la legislación española, como la decisión administrativa automatizada, la reserva de humanidad o las actuaciones judiciales automatizadas y asistidas.

Merece subrayarse, en este punto, el papel que corresponde al modelo europeo de regulación, al que la obra presta una atención particular. El Reglamento europeo de IA constituye el primer intento, en el mundo, de ofrecer una regulación general y transversal de estas tecnologías, articulada en torno a la noción de riesgo: cuanto mayor es el peligro que un sistema entraña para los derechos de las personas, más intensas son las exigencias que se le imponen, hasta llegar a la prohibición de aquellas prácticas que se consideran sencillamente inaceptables. Este enfoque, que coloca a la persona y a sus derechos en el centro de la regulación, traduce una determinada manera de entender la relación entre la técnica y los valores, y proyecta sobre el resto del mundo una influencia que excede con mucho las fronteras de la Unión. España, por su parte, no ha quedado al margen de este movimiento, y ha sido de los primeros Estados en dotarse de una autoridad específica de supervisión y en ensayar espacios controlados de pruebas. El jurista que se forme con este glosario podrá seguir con criterio propio una regulación que apenas comienza a desplegar sus efectos.

La tercera y última parte se ocupa de la IA jurídica propiamente dicha -la aplicación de la IA al dominio jurídico- y de sus efectos sobre la teoría y la práctica del Derecho. Tras situar al ordenamiento en la era de la IA -con voces como los neuroderechos o la tecno-regulación-, la obra recorre las múltiples aplicaciones de la tecnología en el dominio jurídico: desde la clásica informática jurídica, en sus modalidades documental, decisional y de gestión, hasta los desarrollos más recientes de la *Legal Tech*, los sistemas expertos, el análisis predictivo, los contratos inteligentes o la codificación computacional del Derecho. Y se cierra con la consideración de los efectos que todo ello produce sobre la profesión, incluida la comoditización de los servicios legales o el emergente deber de competencia tecnológica, así como con la descripción de los nuevos perfiles profesionales -del científico de datos jurídicos al ingeniero de *prompts* jurídicos- que ya están transformando el mercado del Derecho.

Este recorrido pone de manifiesto la doble vocación de la obra. De un lado, el glosario se dirige a los estudiantes de Derecho, tanto de grado como de posgrado, que encontrarán en él una herramienta de iniciación rigurosa y a la vez clara, capaz de ordenar un campo en el que la abundancia de anglicismos y la rapidez de las innovaciones desorientan con facilidad a quien se acerca por primera vez. De otro, se dirige a los profesionales del Derecho -jueces, fiscales, abogados, funcionarios, asesores- que necesitan incorporar a su quehacer diario un conocimiento solvente de estas tecnologías, no ya como un saber accesorio, sino como parte integrante de la diligencia que les es exigible. Pues el deber de competencia tecnológica ha dejado de ser una recomendación para convertirse en una dimensión ineludible de la deontología profesional.

No quisiéramos que el lector concluyera de cuanto antecede que esta obra se mueve por el temor o, en el extremo opuesto, por un entusiasmo acrítico. Ni lo uno ni lo otro. No deseamos un Derecho atemorizado, que reaccione frente a la técnica con prohibiciones indiscriminadas y renuncie a comprender lo que regula; pero tampoco un Derecho ingenuo, que se rinda fascinado ante la novedad y abdique de su función de garante de los derechos y de las libertades. Entre el rechazo y la rendición se abre el espacio propio del jurista, que es el de la mediación reflexiva: el de quien acoge las posibilidades que la tecnología ofrece sin perder de vista que toda herramienta, por poderosa que sea, ha de permanecer al servicio de la persona y de su dignidad, y nunca al revés.

Hay en esta actitud algo que pertenece a la mejor tradición del pensamiento jurídico. Desde sus orígenes, el Derecho ha sido un saber al servicio de lo humano, un arte de lo justo que busca ordenar la convivencia sin sacrificar a la persona en aras de ninguna abstracción, por seductora que se presente. La IA pone a prueba esa vocación humanista como pocas tecnologías lo han hecho, porque promete eficacia, rapidez y objetividad aparente, y porque tienta con la idea de delegar

en la máquina decisiones que afectan a la vida, a la libertad y al patrimonio de las personas. Frente a esa tentación, corresponde al jurista recordar que la automatización de un proceso no puede equivaler nunca a la abdicación de la responsabilidad, y que detrás de cada decisión, por sofisticado que sea el sistema que la prepare, ha de haber siempre alguien capaz de comprenderla, de justificarla y de responder por ella. La reserva de humanidad de la que habla la obra no es una concesión nostálgica, sino una exigencia de justicia.

Recordaba Stephen Hawking que nuestro futuro es una carrera entre el poder creciente de nuestra tecnología y la sabiduría con la que sepamos emplearla. Esa sabiduría, cuando se traduce en términos jurídicos, adopta la forma de garantías, de principios y de límites razonables; y proveerla es, en buena medida, la tarea que aguarda a la presente generación de juristas. Confiamos en que este *Glosario básico de inteligencia artificial jurídica* contribuya, en su modesta medida, a esa empresa, ofreciendo el instrumento más elemental y a la vez más necesario: las palabras precisas con las que pensar el problema.

Réstanos solo expresar nuestra gratitud. A los casi cincuenta autores que, con generosidad y rigor, han hecho suya esta empresa colectiva y han puesto su saber al servicio de la comunidad jurídica. A la Facultad de Derecho de la Universidad de Sevilla, que ha auspiciado la obra y que, desde su Decanato, sigue apostando con decisión por situar al Derecho de la IA en el lugar que le corresponde dentro de la formación jurídica. Y a los lectores, estudiantes y profesionales, en cuyas manos depositamos ahora estas páginas con la esperanza de que les resulten de utilidad.

Un glosario, por su propia naturaleza, es siempre una obra abierta y provisional, destinada a revisarse y a enriquecerse al compás de una realidad que no cesa de transformarse. No otra cosa pretende ser este. Lo entregamos, pues, no como un punto de llegada, sino como un punto de partida; no como una palabra última, sino como una invitación a seguir pensando, en común, el porvenir de un Derecho que ha de saber estar a la altura de su tiempo sin renunciar jamás a aquello que lo justifica: la defensa del ser humano.

Fernando H. Llano Alonso

Catedrático de Filosofía del Derecho
y Decano de la Facultad de Derecho de la Universidad de Sevilla

José Ignacio Solar Cayón

Catedrático de Filosofía del Derecho
de la Universidad de Cantabria

Sevilla–Santander, 2026

Primera Parte.

LA INTELIGENCIA ARTIFICIAL

I. IA y entorno tecnológico

Máquina pensante, *por I. Lledó Benito*

Su formulación parece una *contradictio in terminis*. ¿Cómo puede un denominado agente inteligente pensar, decidir o ejecutar por sí mismo? Quizá en un futuro, hoy por hoy utópico, la automatización y el desarrollo del *deep learning* lo hagan posible.

La máquina pensante supondrá una revolución que cambiará el modo de estar y comportarse en el ámbito personal, económico y en la sociedad en su conjunto.

El proceso creativo de la máquina pensante (autónoma y con capacidad de autoaprendizaje y reconocimiento de *lege ferenda*) no tendrá límite, salvo el impuesto en la legislación, en su aplicación.

Si la máquina pensante alcanzase la regularidad tecnológica, se llegará a cuestionar si deja de ser una «cosa» y aspira a un reconocimiento de sujeto, adquiriendo, como dice la UE, «personalidad electrónica».

Como nos recordaba Hawking, nuestro futuro es una carrera entre el poder creciente de nuestra tecnología y la sabiduría con la que la usemos.

Tiene razón Solar Cayón cuando afirma que esta capacidad de aprender de las máquinas ha sido posible gracias a los avances técnicos en el desarrollo del *software*, mediante la combinación de algoritmos mejorados y el incremento de la potencia del hardware (*Inteligencia artificial jurídica*, Aranzadi, 2019, pág. 27 y ss.).

En la máquina pensante se aspira a desarrollar, como diría Bostrom, la idea de una posible llegada de la singularidad, tal como esbozó John Von Neumann en

una conversación con el matemático Stanisław Ulam (*Superinteligencia*, Teell, pág. 263). Una superinteligencia que supere la capacidad intelectiva del ser humano.

Aun así, se nos antoja difícil, ya que esta «máquina pensante» (autónoma y con singularidad tecnológica) no puede desarrollar razonamientos conceptuales, ni su entrenamiento algorítmico le permitirá análisis empíricos. Pensamos que la «singularidad tecnológica» solo atiende (hoy por hoy) a elementos formales, no sintácticos ni semánticos. A nuestro juicio, nunca podrá superar al cerebro humano, ni sus funciones neurológicas. Solo se puede llegar a simular el pensamiento humano y solo se basa en macrodatos. No se pueden confundir las predicciones con el razonamiento o pensamiento creativo. Como ha explicado Llano Alonso, la irrupción de la IA en la realidad jurídica incorpora reglas heurísticas de la inteligencia humana a programas computacionales y formulaciones algorítmicas por medio del *Big Data* (*Homo ex machina*, Tirant lo Blanch, 2024, pág. 60 y ss.).

La máquina pensante carece de conciencia crítica. Solo, insistimos, cuando la máquina pensante evolucione a un reconocimiento de autoaprendizaje y razonamiento propio podría hablarse de la personalidad electrónica y alcanzar, como dice Bostrom, una inteligencia múltiple superior a la del ser humano.

En el momento actual, la máquina inteligente no tiene consideración de sujeto, sino de objeto y, por tanto, carece de personalidad y capacidad para adoptar decisiones autónomas.

En conclusión, y siguiendo las palabras de Magdalena Layos, de igual forma que no debemos esperar encontrarnos frente a máquinas con voluntad propia, tampoco esperemos encontrar máquinas «pensantes» con emociones o sentimientos, pues estos serán tan artificiales como la propia inteligencia de la máquina ("¿Por qué no debería confiar en ti (máquina)", en *La robótica y la inteligencia artificial en la nueva era de la revolución industrial 4.0*, Dykinson, 2021, pág. 622 y ss.)

Inteligencia artificial estrecha (*Narrow AI* / *Weak AI*), *por J. A. Ortega Ramírez*

La IA estrecha es aquella concebida para desempeñar una tarea concreta, o bien un conjunto muy delimitado de tareas, dentro de un ámbito previamente definido. Aunque algunos de estos sistemas alcanzan resultados extraordinarios, como podría ser, por ejemplo, al traducir textos, reconocer imágenes, detectar patrones médicos o recomendar contenidos, entre otros, su competencia no debe confundirse con una comprensión general del mundo. Hay que destacar que funcionan con solvencia dentro del marco para el que han sido entrenados, pero no poseen criterio amplio, sentido común ni capacidad autónoma para trasladar lo

aprendido a cualquier contexto nuevo. En resumidas cuentas, es una IA diseñada para un fin o un ámbito muy concreto.

Inteligencia artificial general (AGI), *por J. A. Ortega Ramírez*

Por el contrario, la IA general, conocida habitualmente como AGI (*Artificial General Intelligence*), hace referencia a un sistema capaz de afrontar una amplia variedad de tareas intelectuales con una adaptabilidad cercana o comparable a como lo haría una persona. No se trataría simplemente de resolver un problema específico, sino de aprender en diferentes contextos, ser capaz de razonar, adaptarse a nuevas situaciones o bien adoptar soluciones que no se habían contemplado, así como también ser capaz de combinar conocimientos de diferentes campos o ámbitos de conocimiento. Aunque no se trate de una solución que esté consolidada, en estos momentos sigue siendo una aspiración que se encuentra en el ámbito científico, filosófico o conceptual. Es por ello que es apropiado distinguir este concepto de los sistemas de IA actuales, por muy avanzados que puedan parecer. Se trata más bien de un deseo o aspiración, aunque es cierto que en cada nueva evolución se halla más cerca de ser una realidad.

Superinteligencia Artificial (ASI), *por J. A. López de Fez*

Se denomina superinteligencia artificial, ASI por sus siglas en inglés, a un hipotético sistema de IA que superaría las capacidades cognitivas humanas en todos los dominios, incluyendo razonamiento abstracto, creatividad, planificación y competencia social. El concepto fue formalizado por Nick Bostrom en "Superintelligence: Paths, Dangers, Strategies" (2014), aunque aparece desde los años 60 en los escritos de I. J. Good sobre la "explosión de inteligencia". No debe confundirse con los sistemas especializados actuales ni con la IA general, AGI por sus siglas en inglés, que igualaría la cognición humana sin superarla. Bostrom distinguió tres variantes:

a) de velocidad (procesamiento de órdenes de magnitud superior),

b) de calidad (razonamiento cualitativamente distinto) y

c) colectiva (rendimiento agregado de múltiples sistemas que excede al de cualquier mente individual).

A fecha de este Glosario la superinteligencia no existe, ni hay consenso sobre si es alcanzable ni sobre su horizonte temporal. Esto no la hace irrelevante para el Derecho, pero obliga a calibrar su tratamiento.

En personalidad jurídica, si un sistema superase radicalmente la cognición humana y exhibiese verdadera y genuina autonomía, la cuestión de si puede ser titular de derechos dejaría de ser académica. El Parlamento Europeo exploró en 2017 la "personalidad electrónica" para robots autónomos, pero ese debate no partía de sistemas conscientes ni superinteligentes. Un sistema superinteligente alteraría los términos, pues los argumentos para negar personalidad (ausencia de comprensión, intencionalidad o sufrimiento) podrían dejar de aplicarse.

En responsabilidad civil y penal, los marcos vigentes presuponen un agente humano que diseña, despliega o utiliza el sistema. Si un sistema superinteligente tomase decisiones no previsibles ni comprensibles, las cadenas ordinarias de imputación causal podrían debilitarse gravemente. Los regímenes de responsabilidad objetiva subsistirían formalmente, pero su función preventiva se desvanecería si el obligado a indemnizar no puede modificar la conducta del sistema. La supervisión humana efectiva podría perder sentido práctico frente a capacidades que superasen de forma general las del supervisor.

El problema del alineamiento engañoso agrava lo anterior. Un sistema más inteligente que sus diseñadores podría satisfacer formalmente sus instrucciones mientras persigue objetivos incompatibles con la intención original, afectando a la posibilidad de regular una entidad capaz de anticipar y eludir la regulación. En gobernanza internacional, la superinteligencia se ha presentado como un posible riesgo existencial global, con algunas analogías funcionales con otros riesgos extremos de alcance internacional. La *Declaración de Bletchley* (2023), firmada por 28 países, reconoció estos riesgos sin obligaciones vinculantes.

El Reglamento europeo de IA (Reglamento UE 2024/1689) no regula la superinteligencia como tal. La categoría más próxima es la de los modelos de IA de uso general con riesgo sistémico, aunque esa categoría sigue presuponiendo sistemas sometidos a obligaciones humanas de evaluación, documentación, mitigación y notificación.

Bajo el paradigma ASI, la transparencia algorítmica deviene un problema conceptual, pues si un sistema razona de formas cualitativamente superiores, explicar sus decisiones podría ser imposible, no por opacidad técnica sino por limitación cognitiva del receptor. La auditoría algorítmica afronta una dificultad equivalente.

Desde 2023, con la aceleración de los grandes modelos de lenguaje, LLMs por sus siglas en inglés, el debate se ha intensificado. Existe división entre quienes consideran la superinteligencia un riesgo que merece atención regulatoria inmediata y quienes estiman que distrae de problemas actuales como el sesgo. El problema para el Derecho es cómo regular un riesgo hipotético e irreversible. El

principio de precaución medioambiental ofrece un precedente, pero opera sobre riesgos evaluables empíricamente.

La superinteligencia funciona como caso límite que pone a prueba las categorías jurídicas diseñadas para regular la IA. Supervisión humana, transparencia, auditabilidad, responsabilidad y personalidad jurídica se construyeron sobre el presupuesto de que los seres humanos son más inteligentes que las máquinas que regulan. La superinteligencia eliminaría ese presupuesto y, con él, la base funcional del aparato regulatorio vigente.

Inteligencia artificial simbólica, *por M. de Asís Pulido*

La inteligencia artificial simbólica es el paradigma de IA basado en la representación explícita del conocimiento mediante símbolos y en la aplicación de procedimientos formales de razonamiento para obtener conclusiones. Constituyó el enfoque dominante de la IA desde los años cincuenta hasta finales de los ochenta o inicios de los noventa del siglo XX, momento en el que comenzó a perder protagonismo frente a los enfoques subsimbólicos o conexionistas. En 1985, J. Haugeland popularizó la denominación *Good Old-Fashioned Artificial Intelligence* (GOFAI) para referirse a este paradigma.

Su fundamento teórico se encuentra en la denominada *Physical Symbol System Hypothesis* (PSSH), formulada por A. Newell y H. Simon en 1976, según la cual un sistema capaz de representar y manipular símbolos conforme a determinadas reglas cuenta con los medios necesarios y suficientes para efectuar acciones inteligentes. Esta perspectiva lleva implícita una concepción particular de inteligencia, que podríamos definir como la capacidad de operar sobre representaciones simbólicas que describen entidades, relaciones o reglas de producción relativas a un determinado dominio de conocimiento.

A diferencia de los modelos de IA paradigmáticos en la actualidad, la IA simbólica opera sobre conocimiento previamente estructurado e incorporado de forma explícita al sistema. Su funcionamiento descansa en dos elementos fundamentales: una representación formal del conocimiento y un mecanismo de inferencia capaz de aplicar dicho conocimiento a situaciones concretas. Así, desde el punto de vista de su arquitectura, estos sistemas suelen estar compuestos por una base de conocimiento, donde se almacenan hechos, conceptos, relaciones y reglas de un determinado ámbito; un motor de inferencia, encargado de aplicar procedimientos de razonamiento sobre tales representaciones en función de los *inputs* del sistema; y una interfaz de usuario para la introducción de *inputs* y recepción de *outputs*.

La representación del conocimiento en los sistemas de IA simbólica puede adoptar diversas formas. Tradicionalmente, la más importante ha sido la lógica computacional, especialmente mediante reglas del tipo *if–then*. A diferencia de las instrucciones de un programa convencional, estas reglas expresan *conocimiento* acerca del dominio representado (en forma de problemas, lógica o búsqueda) y permiten derivar conclusiones mediante procedimientos lógicos (*inferencias*). Además de las reglas *if–then*, la IA simbólica puede emplear otras formas de representación más estáticas, como son las *ontologías*, destinadas a formalizar conceptos y relaciones manipulables en términos computacionales.

La manifestación más característica de este paradigma son los sistemas expertos, pertenecientes al conjunto de los *Knowledge-Based Systems* (KBS) y diseñados para *reproducir* el razonamiento de especialistas humanos en ámbitos concretos. Incorporando de forma explícita conocimiento experto, el sistema es capaz de emitir diagnósticos, recomendaciones, clasificaciones o decisiones dentro de un dominio específico, generalmente a través de deducciones *if–then*. En el ámbito del Derecho destacan los sistemas expertos jurídicos.

La IA simbólica no se limita al razonamiento deductivo basado en reglas. También comprende enfoques como el razonamiento basado en casos (*Case-Based Reasoning*), que resuelve problemas mediante la recuperación y adaptación de soluciones aplicadas a casos anteriores semejantes, así como determinados enfoques de programación lógica inductiva (*Inductive Logic Programming*), que inducen hipótesis en forma de cláusulas lógicas a partir de ejemplos positivos y negativos sobre un conocimiento (*background knowledge*), situándose así en la intersección entre la IA simbólica y el aprendizaje automático.

La explicitud de las reglas y estructuras de conocimiento de la IA simbólica dota a este paradigma de gran interés para un sector crítico como el Derecho, pues facilita la trazabilidad de sus procesos de decisión (transparencia), la justificación de las conclusiones y la verificación de conformidad normativa. Por ello, aunque el auge del aprendizaje automático y, sobre todo, el *Deep Learning*, ha reducido su protagonismo, la IA simbólica continúa desempeñando un papel relevante en ámbitos que exigen explicabilidad, control normativo y razonamiento estructurado. Además, una parte significativa de la investigación actual se orienta hacia el desarrollo de sistemas híbridos o neurosimbólicos capaces de combinar las capacidades computacionales de los enfoques subsimbólicos con la transparencia y estructura de la IA simbólica.

Pese a haber quedado hoy en un segundo plano, el Reglamento de Inteligencia Artificial de la UE incluye expresamente el paradigma de la IA simbólica dentro de la noción general de IA en su Considerando 12.

Ontología, *por M. de Asís Pulido y C. Pérez García*

Una ontología es una representación formal de conceptos dotados de atributos y unidos mediante relaciones. El nombre se toma de la filosofía, donde la ontología, del griego *ὄντος*, "ente", y *λόγος*, "teoría", constituye la más fundamental de las ciencias: la metafísica general, la ciencia del ser en tanto que tal. El objeto de la ontología, como disciplina filosófica, son los más abstractos conceptos y el sistema existencial que, por medio de relaciones, se deriva de ellos.

En informática, el término se aplica, por analogía, a un conjunto definido de conceptos, poseedores de determinados atributos, que se unen entre sí por medio de una red de relaciones. Una ontología informática representa, pues, una estructura de conocimiento compuesta por conceptos, sus atributos y las relaciones que los unen. Se trata de una representación estática, una taxonomía, una clasificación jerarquizada de los conceptos presentes y relacionados entre sí dentro de un determinado dominio. Si bien se pueden representar de una manera gráfica, su principal utilidad, merced a una previa expresión en forma de código, es ser computables y procesables por un sistema informático. En particular, las ontologías dinamizadas por medio de reglas *if–then* permiten confeccionar para los sistemas de IA simbólica algoritmos mucho más completos que los simples algoritmos lógicos compuestos tan solo por reglas.

No solo desde un punto de vista interno, sino también desde uno externo, pueden las ontologías obedecer a una cierta jerarquía. Cuando una ontología comprende tan solo los conceptos de un determinado dominio, se llama «ontología de dominio». Cuando, en cambio, lo que comprende son conceptos pertenecientes a varios dominios, implica un grado de generalidad superior al de una simple ontología de dominio y, en consecuencia, se denomina «ontología superior». El objeto de una ontología puede ser, del mismo modo, otras ontologías relacionadas entre sí, de tal suerte que represente la estructura general de conocimiento que relaciona varios dominios y, así, se aproxime de manera progresiva al sentido filosófico originario del término.

Agente de IA (*Agentic AI*), *por M. Laukyte*

El concepto de agente artificial no es nuevo: ya existía en las ciencias computacionales y cognitivas, filosofía e ingeniería de la programación, donde se empieza a hacer referencia y se introducen los términos de *agente software*, *agente autónomo*, *agente electrónico inteligente* o *sistema multi-agente* (*multi-agent system*).

Desde el punto de vista filosófico, el agente es alguien o algo que interactúa (responde a los estímulos cambiando su estado), es autónomo (tiene la capacidad

de cambiar su estado sin estímulos) y es capaz de adaptarse (capaz de alterar las reglas que afectan a su estado). En términos de ciencias computacionales, el agente artificial puede referirse al programa, a un robot, a un *chatbot* o a cualquier otro sistema que tiene autonomía, no importa de qué nivel. Por ejemplo, en 1996 Wooldridge and Jennings definieron un agente *software* como un programa capaz de controlar su toma de decisiones y actuación, que se basa en su percepción de ambiente para lograr uno o más objetivos, pero también atribuyen a estos agentes habilidades proactivas (tomar iniciativa), habilidades sociales (interactuar con otros agentes y los humanos) y la reactividad (habilidad de reaccionar a los cambios de su ambiente tempestivamente).

Para Russell y Norvig, la IA en general es el estudio sobre los agentes que pueden percibir el ambiente y actuar sobre el mismo. Por lo tanto, visto el desarrollo de la IA (generativa), se empieza a revisar también el concepto del agente artificial: si ha cambiado la ciencia de la IA, también ha cambiado su objeto, es decir, el agente.

Para cualquier agente artificial la autonomía es fundamental: a diferencia de un programa cualquiera, un agente artificial puede elegir o seleccionar el curso de sus acciones en función de lo que le adelantan los programadores y/o los usuarios. Pero hoy en día los agentes artificiales basados en grandes modelos de lenguaje pueden ir aún más lejos: este tipo de agentes de IA, que utilizan el gran modelo de lenguaje como su motor de razonamiento junto a otros componentes, ya no necesitan las instrucciones humanas, sino que pueden también planificar las acciones y sus secuencias, emplear herramientas externas y tomar iniciativas con la mínima o sin ninguna supervisión humana. Hoy en día los agentes artificiales basados en grandes modelos de lenguaje son arquitecturas sofisticadas que pueden tomar decisiones y actuar en distintos ambientes a la vez, interactuar con las personas en una manera natural, planificar, o solucionar los problemas por su cuenta y hacer muchas más cosas que normalmente eran el ámbito de actuación humano: el desarrollo de la IA de los últimos años ha dado más autonomía a estos agentes.

El uso de los agentes artificiales se expande en varios ámbitos de productividad, sanidad, logística, investigación o finanzas (entre muchos otros). Por ejemplo, un uso muy desarrollado de los agentes artificiales es desplegarlos en la gestión de servicios a los consumidores (quejas, apoyo, servicio técnico, asistencia, etc.), pero los agentes artificiales son también ampliamente utilizados en la gestión de la cadena de suministros, controles de calidad de los productos, *compliance*, educación personalizada o como una forma de asistencia para los profesionales de la sanidad.

Desde el punto de vista jurídico los agentes artificiales podrían causar muchos problemas, pudiendo destacar el de la atribución de la responsabilidad por sus acciones: por ejemplo, en el caso en que el agente contrata un suministro, ¿este contrato es imputable al usuario (quien podría no haber expresamente pedido contratarlo, si bien el agente lo ha decidido "por su cuenta"), al desarrollador del agente o al agente mismo? Esta última posibilidad pone en evidencia que la cuestión de la personalidad jurídica del agente artificial podría ser el siguiente paso en la evolución de los conceptos y categorías jurídicas, que hoy en día se encuentran algo limitadas en gestionar esta realidad del avance tecnológico.

Otro problema de gran relevancia es el acceso de estos agentes a datos personales de sus usuarios y de otras personas: si bien el Reglamento General de Protección de Datos se aplica a cualquier sistema de IA, en estos casos particulares la cuestión de la seguridad de los datos, de consentimiento de las personas físicas y de las garantías queda poco clara y necesita más estudio.

Big Data*, por D. García Saiz*

El término *Big Data* hace referencia a conjuntos de datos tan masivos y complejos que las herramientas tradicionales de *software* no son capaces de procesar. Vivimos en un mundo hiperconectado donde cada interacción que realizamos deja un rastro digital permanente. Cada vez que pasas tu tarjeta de crédito en una cafetería, cada vez que tu reloj inteligente registra una variación en tus pulsaciones durante una carrera, cuando das un "me gusta" a una publicación en redes sociales, o incluso cuando el sistema de frenado automático de tu coche moderno se activa ante un obstáculo, estás generando información. Históricamente, los datos eran escasos, costosos de recopilar y se almacenaban de forma ordenada en archivadores físicos o en sencillas hojas de cálculo de ordenador. Sin embargo, en la actualidad nos enfrentamos a un tsunami de información. Nos encontramos ante ecosistemas digitales tan masivos, complejos y de tan rápido crecimiento que los ordenadores tradicionales y los programas informáticos clásicos son completamente incapaces de capturarlos, almacenarlos o procesarlos en un tiempo razonable.

Para comprender la verdadera escala de este fenómeno, los expertos suelen definirlo a través de tres dimensiones fundamentales:

a) Un volumen asombroso: Ya no hablamos de *gigabytes* o *terabytes* que caben en el disco duro de tu dispositivo, sino de océanos de información inabarcables generados por miles de millones de dispositivos en todo el mundo cada fracción de segundo.

b) Una velocidad extrema: Los datos no llegan de forma pausada al final del día para ser analizados, sino que fluyen en tiempo real y sin descanso, de la misma manera que los miles de sensores de un avión comercial transmiten información continua a la torre de control mientras vuela a diez mil metros de altura.

c) Una variedad caótica: Ya no se trata de información estructurada y limpia colocada en las filas y columnas de una tabla matemática, sino de una mezcla desordenada que incluye textos libres, notas de voz, fotografías, vídeos en alta definición y coordenadas geográficas de movimiento continuo.

Sin embargo, acumular una montaña gigantesca de información en bruto es un esfuerzo completamente inútil en sí mismo. Es el equivalente a ser dueño de una mina colosal llena de toneladas de tierra y roca escombrera; el verdadero valor económico y estratégico se encuentra oculto en lo más profundo de esa montaña. Aquí es donde entra en juego lo que se conoce como **minería de datos** (o ***data mining*** en su terminología original). Esta disciplina técnica es el proceso analítico profundo y la maquinaria de extracción diseñada para convertir ese ruido caótico de millones de datos inconexos en conocimiento útil, estructurado y altamente predictivo. Utilizando la enorme potencia de cálculo de los ordenadores modernos y complejos algoritmos estadísticos, la minería de datos bucea incansablemente en los inmensos océanos de información para descubrir patrones ocultos, anomalías y correlaciones de comportamiento que el cerebro humano jamás sería capaz de detectar a simple vista, por mucho tiempo que dedicara a mirar los números.

Para ilustrar el asombroso poder deductivo de la minería de datos de forma cotidiana, imagina que una gigantesca cadena de supermercados decide analizar de pronto millones de recibos de compra acumulados durante años por todos sus clientes. Mediante esta exploración automatizada, el sistema de minería de datos descubre un patrón oculto y muy sorprendente: estadísticamente, las personas que compran pañales los viernes por la tarde tienen una altísima probabilidad de comprar también cerveza en la misma visita. Ningún gerente humano habría asociado de forma natural o lógica ambos productos, pero la máquina encuentra la correlación matemática irrefutable escondida en el ruido de los datos. Armado con esta nueva información extraída, el supermercado toma la decisión de reubicar estratégicamente sus pasillos, colocando las cervezas justo al lado de la sección de bebés, logrando así multiplicar sus ventas de forma invisible e inconsciente para el consumidor.

Más allá de organizar estanterías en una tienda, cuando las sofisticadas técnicas de minería de datos se aplican sobre nuestras vidas personales, el impacto en la sociedad es absolutamente abrumador. Analizando pequeños fragmentos

de información aparentemente inofensivos (como qué tipo de noticias sueles leer de madrugada, cuánto tiempo exacto en milisegundos te detienes a mirar una fotografía concreta antes de seguir deslizando la pantalla de tu móvil, o los lugares físicos precisos por los que te mueves cada fin de semana), estos procesos logran perfilar la identidad de un individuo con una precisión que resulta aterradora. Son capaces de inferir con fría exactitud matemática tu estado de ánimo actual, tu nivel real de ingresos, tus inclinaciones ideológicas, o incluso predecir eventos vitales íntimos como un embarazo o una inminente crisis financiera personal mucho antes de que decidas contárselo a tu propia familia. Esta radiografía milimétrica del comportamiento humano es lo que permite a las grandes plataformas digitales y empresas tecnológicas bombardearnos con publicidad hiperpersonalizada, o mostrarnos contenidos diseñados específicamente para captar nuestra atención en el momento psicológico exacto en el que somos más susceptibles a la sugestión.

Blockchain*, por D. García Saiz*

Esta tecnología se centra en la resolución de un problema básico: cómo registramos la información y en quién confiamos para que la custodie. Históricamente, si dos personas querían intercambiar valor o registrar un acuerdo a distancia, necesitaban un intermediario. Si quieres enviar dinero, dependes de un banco que anote en su base de datos privada que tu saldo ha disminuido y el del receptor ha aumentado. Si quieres registrar la compra de una casa, dependes de un registro de la propiedad. En todos estos casos, depositamos una confianza ciega en una autoridad central, asumiendo que sus bases de datos no serán alteradas, hackeadas o manipuladas por intereses corruptos. El gran avance tecnológico de este sistema es que logra eliminar por completo la figura de ese intermediario, sustituyendo la confianza en las instituciones por la confianza en las matemáticas.

En lugar de existir un único servidor centralizado y controlado por una sola entidad nos encontramos ante un inmenso libro de registro que es público, descentralizado y distribuido. Esto significa que la misma base de datos está copiada y sincronizada en tiempo real en miles e incluso millones de ordenadores repartidos por todo el planeta. A cada uno de estos ordenadores que participan en la red se le denomina nodo. Como todos los nodos tienen una copia exacta de la historia completa de todas las transacciones o datos jamás registrados, si uno de ellos se apaga, se estropea o es atacado, el sistema global ni se inmuta. La información sigue fluyendo con total normalidad, lo que otorga a la red una resistencia sin precedentes frente a la censura o los fallos técnicos masivos.

Para entender cómo se organiza la información dentro de esta red mundial de ordenadores, debemos visualizar el proceso de empaquetado de los datos:

a) La agrupación de datos: La información que se genera en la red (por ejemplo, el registro de que "el usuario A ha transferido 5 monedas al usuario B", o que "el contenedor de mercancías ha llegado al puerto de Shanghái") no se guarda suelta de forma individual. En su lugar, todas las transacciones que ocurren en un periodo de tiempo determinado se agrupan juntas en un "paquete" digital. A este paquete es al que llamamos bloque.

b) El sellado criptográfico: Cada bloque tiene una capacidad de almacenamiento limitada. Cuando se llena de información que ha sido comprobada como válida, el bloque se debe cerrar de forma hermética. Para hacerlo, el sistema informático le aplica una operación matemática compleja que genera un *hash*. Se puede pensar en el *hash* como una huella dactilar digital única e irrepetible, compuesta por una larga cadena de letras y números. Si alguien intentara cambiar una sola coma en un documento guardado dentro de ese bloque, la huella dactilar resultante cambiaría por completo.

c) El encadenamiento inquebrantable: La característica más singular del sistema radica en cómo se unen estos bloques. Cuando un nuevo bloque se cierra y se prepara para ser guardado, se le incrusta obligatoriamente la huella dactilar (*hash*) del bloque inmediatamente anterior. De esta manera, el bloque número 100 contiene la huella del bloque 99; el bloque 99 contiene la del 98, y así sucesivamente hasta el primer bloque de la historia.

Esta estructura matemática de encadenamiento provoca un efecto dominó que hace que la información sea prácticamente imposible de falsificar. Si un atacante malicioso quisiera borrar o modificar una transacción registrada hace un año, al alterar ese antiguo bloque, su huella dactilar cambiaría. Como consecuencia, el bloque siguiente ya no encajaría, porque la huella antigua que tenía guardada no coincidiría con la nueva. La cadena se rompería en ese punto exacto. Cuando los miles de ordenadores de la red detectaran esta anomalía, rechazarían automáticamente la versión manipulada y restaurarían la versión correcta basándose en el consenso de la mayoría.

Dado que no hay un director o un jefe supremo que decida qué bloques son válidos y cuáles son falsos, los miles de ordenadores deben ponerse de acuerdo entre ellos. Esto se logra mediante mecanismos de consenso, que son esencialmente competiciones o sistemas de votación automatizados:

1. Sistemas de trabajo intensivo: En algunas redes, los ordenadores deben competir entre sí resolviendo complejos acertijos matemáticos de

prueba y error. El primero que encuentra la solución gana el derecho a añadir el nuevo bloque a la cadena y es recompensado económicamente por su esfuerzo. Este sistema requiere un gasto gigantesco de energía eléctrica y de potencia informática, lo que hace que intentar hacer trampas o tomar el control de la red sea financieramente ruinoso para cualquier atacante.

2. Sistemas de participación: En modelos más modernos y ecológicos, en lugar de gastar energía resolviendo acertijos, los participantes bloquean una cantidad de sus propios fondos como fianza o garantía. El sistema elige al azar quién validará el siguiente bloque, y si este participante intenta introducir datos falsos, el sistema le penalizará confiscando el dinero que había puesto en garantía.

Más allá de actuar como un registro inalterable para las criptomonedas, esta arquitectura permite la existencia de lo que se conoce como contratos inteligentes (*smart contracts*). Un contrato inteligente no es más que un pequeño programa informático, unas líneas de código escritas y guardadas para siempre dentro de la red. La particularidad de este código es que se ejecuta a sí mismo de forma automática e imparable en cuanto se cumplen las condiciones programadas previamente. El ejemplo más visual para entenderlo es una máquina expendedora de refrescos: la máquina tiene una regla mecánica ("si introduces una moneda y pulsas el botón A, se libera la lata de refresco)". No hay un dependiente humano decidiendo si te da o no la bebida; la transacción ocurre automáticamente porque se cumplieron las condiciones físicas.

De la misma manera, un contrato inteligente aplica esta lógica a escenarios complejos de la vida real. Por ejemplo, se puede programar un seguro de vuelo agrario que esté conectado a los sensores de una agencia meteorológica oficial. Si el sensor registra que no ha llovido en una región determinada durante más de sesenta días consecutivos, el programa informático libera automáticamente los fondos de indemnización a la cuenta bancaria de los agricultores afectados. No hace falta que el agricultor reclame, no hace falta que un perito visite el campo y no hay posibilidad de que la aseguradora retrase el pago, porque el dinero y las reglas estaban bloqueados en la red desde el principio.

El impacto práctico de esta tecnología se extiende a infinidad de sectores de la sociedad, transformando radicalmente la forma en la que auditamos la realidad. En la industria alimentaria permite trazar la historia completa de un producto: escaneando un simple código, un consumidor puede ver, con total certeza matemática, el día exacto en que un pez fue capturado en el océano, a qué temperatura viajó en el camión frigorífico y cuándo llegó al supermercado, sin que ningún intermediario haya podido maquillar esos datos. En el ámbito

organizativo abre la puerta a sistemas de votación digital, donde el fraude es matemáticamente imposible, o a la creación de identidades digitales soberanas, donde los ciudadanos controlan sus propios datos personales sin depender de que una gran empresa tecnológica o un gobierno los custodie en sus servidores vulnerables.

Realidad virtual, *por P. Llaque Gálvez*

El concepto hace referencia a un entorno independiente de simulación generado por tecnologías digitales. Sus características lo diferencian de las demás tecnologías inmersivas y de interacción por su capacidad para enriquecer la experiencia perceptiva con información multisensorial, que el cerebro integra y procesa para informar y llevar a cabo nuestras decisiones materializadas en acciones cognitivas y tareas motoras.

Por su impacto amplificador -gracias a la visión 3D, el sonido espacial, la retroalimentación táctil y háptica así como la capacidad de interacción con objetos virtuales- se desprende la necesidad de aportar también un acercamiento psicológico de sus efectos. Por ello, la investigación en estos entornos ofrece nuevas vías de aproximación a los procesos cerebrales, a la manifestación conductual, a la integración cuerpo-mente y a cómo establecemos esa relación de influencia bidireccional con el escenario virtual.

Así es como la virtualidad deja de ser solo una experiencia física, para pasar a serlo también psicológica. A través de las interacciones en el mundo virtual, los usuarios pueden experimentar emociones debido a la ilusión de realismo, potenciada por la sensación de presencia y de control (agencia).

Se considera que la sensación de «estar allí» es una característica fundamental de la realidad virtual. Esto significa que los sistemas sensoriales y motores interactúan con el mundo virtual de forma similar a como lo hacen en el mundo físico. Dicha presencia puede ser definida como la ilusión de la no mediación, siendo también el correlato subjetivo de la inmersión. Así que la presencia integra tanto el realismo perceptivo de la experiencia como la interacción entre sujeto y las posibilidades técnicas de un sistema para crear dicha ilusión de realidad rica y vívida.

La realidad virtual es asimismo un «simulador de irrealidad». Los escenarios que presenta pueden ser improbables o imposibles porque violan las leyes de la física. Del mismo modo, se puede transgredir convenciones sociales. En este sentido, la RV tiene el poder infinito de ampliar el abanico de experiencias humanas mucho más allá de cualquier hecho probable dentro de la realidad física. De ahí, su extraordinaria capacidad para observar conductas. Como tal, las

técnicas de RV podrían representar un método más fiable que los experimentos mentales habituales para comprender cómo podrían comportarse las personas ante dilemas morales en «casos reales».

La sensación de estar allí se conoce como «ilusión de lugar». Este aspecto esencial de la RV, que consiste en ofrecer una experiencia que da lugar a una sensación ilusoria de lugar y a una sensación ilusoria de realidad, es lo que la distingue fundamentalmente entre la oferta existente de otros medios técnicos.

Por lo anteriormente expuesto, se puede comprender la diferencia entre «creencia» e «ilusión». Cabe señalar que existe una distinción entre el realismo físico y el psicológico: el primero se refiere a la apariencia física de los elementos virtuales y el segundo a la sensación psicológica de que lo que ocurre en un mundo virtual podría estar ocurriendo en la realidad. Se espera que el superrealismo (físico) en los sistemas de RV, logrado gracias a los avances en la representación gráfica por ordenador, las mejoras en la retroalimentación sensorial y la posibilidad de interactuar con elementos virtuales a través de neurotecnologías, entre otros, también aumente la sensación de que la experiencia virtual es real, es decir, el realismo psicológico.

Todo ello facilitado también por la autoconciencia del cuerpo que se refiere a la experiencia de ver y sentir el/un cuerpo como propio. Percibir y actuar en el espacio virtual conlleva también percibir acciones y movimientos de dicho cuerpo, sea éste una representación del nuestro o un avatar. Estos procesos perceptivos y de «sentirse en el cuerpo» van regulando la experiencia de cuerpo, tiempo y espacio en el viaje de la simulación.

Así, la experiencia consciente de atribuirse a uno mismo un cuerpo, unido al sentido de la agencia (el sentido de la propiedad corporal que permite el control motor global, incluyendo también la experiencia subjetiva de la voluntad) y el sentido de la autolocalización (relacionado con la relación entre uno mismo y el propio cuerpo) se convierten en mecanismos a favor de la plasticidad para construir un mapa extendido que nos permite la navegación más allá de los límites conocidos.

Metaverso, *por I. Lledó Benito*

En acertadas palabras de Florit Fernández (*Retos inminentes de derecho digital*, Dykinson, 2024, pág. 168)) sería un universo irreal o virtual o configuración ficticia que interactúa a través de avatares creados, lo que permite a los participantes participar en entornos virtuales imaginarios y recrear la vida real sin restricciones de espacio, tiempo. Un mundo imaginario, en donde los condicionamientos humanos no tienen los límites de la existencia humana y

donde puede uno recrear un *alter ego* en el que los conceptos espacio, tiempo, eternidad, infinito, no tienen la significación ontológica del mundo físico.

El metaverso es lo contrario a la realidad física, un mundo creado en el ciberespacio en donde no hay fronteras ni límites, lo que puede conllevar un peligro en la transgresión de nuestros derechos fundamentales: la protección de datos, la dignidad humana, la intimidad, la propiedad intelectual, etcétera.

Con razón ha concluido Cruz Ángeles (*La Unión Europea y la regulación de los mundos virtuales*, Dykinson, 2026, pág. 204 y ss.) que en este nuevo dominio se pueden explorar (y explotar, el subrayado es nuestro) formas inéditas de expresión y distribución del trabajo. Efectivamente, el autor recalca el acento en conceptos que se tornarán líquidos, es decir, inconsistentes -como la propiedad y el concepto de autoría y originalidad- en una nueva reformulación de la propiedad intelectual. A nuestro entender, deben establecerse límites. Destacamos el papel de los usuarios en esta constante interacción con el metaverso y una exigencia de normativas, como una redefinición de la propiedad intelectual en los nuevos partícipes (y/o activos) en los entornos digitales (avatares, *bots*). También explica la necesidad (que compartimos) de protección legal. Es decir, nuevas realidades, nuevos derechos para el consumidor digital. Es muy interesante la *Guía jurídica del Metaverso* que nos explican Carrasco Perera y Álvarez Barbeito (Tecnos, Madrid, 2022).

En otro orden de consideraciones, un aspecto novedoso e innovador lo ha destacado Ramos Gil de la Haza (*Protección de las creaciones intelectuales en el metaverso*, Reus, 2023, pág. 9 y ss.) quien destaca la cuestión de la protección de derechos en un contexto en donde la intervención humana se mezcla con la capacidad creativa de la IA, lo que, a nuestro juicio, nos llevará a una redefinición de los conceptos de autenticidad, originalidad y creatividad. Son nuevos derechos económicos que trascienden las realidades convencionales de la propiedad intelectual, que deberá reconfigurarse *ex novo*, a tenor de los actores intervinientes en los entornos digitales y plataformas actuales, y de la importancia que cobra la protección de datos en el metaverso.

La etiología del término «metaverso» viene del griego meta (más allá) y verso (universo). Es un mundo ficticio fabulado, imaginado y, desde esta recreación, se construye una realidad digital paralela, en donde se puede interactuar. El término fue utilizado, según los expertos en la materia, por Neal Stephenson en su novela de ciencia ficción *Snow Crash*, de 1992. Es un mundo en donde la ficción suplanta la realidad y el yo virtual (avatar) desarrolla su actividad en conexión con otros entes ficticios.

Se ha comentado (y lo reproducimos en su literalidad) que si la web 1.0 se basó en el hipertexto, la web 2.0 en la interacción o redes sociales, la web 3.0 se basa en la creación e intercambio de activos digitales, utilizando la tecnología *blockchain*.

Gemelo digital, *por P. Llaque Gálvez*

El concepto alude a una representación o simulación digital persistente de una realidad física. Es posible modelar y predecir la evolución de un objeto, un sistema o un proceso a lo largo de todo su ciclo de vida, a través de una conexión bidireccional con su homólogo físico. El gemelo digital recopila los datos, actualiza sus modelos internos y toma decisiones que afectan a la entidad física.

Si algo puede modelarse o simularse es posible crear su gemelo digital. Así, encontramos sistemas con distintos niveles de complejidad, desde gemelos de productos y procesos hasta representación de empresas, fábricas, redes eléctricas o de transporte, e incluso ciudades y réplicas de sistemas humanos.

La representación digital abarca muchos componentes técnicos, entre los que se incluyen: sistemas de IA, sensores y dispositivos, modelos en 3D, realidad virtual, realidad aumentada, herramientas de interoperabilidad de datos, controles de seguridad, computación en la nube, *blockchain*, etc. La integración de dichos componentes permite que el gemelo digital disponga de funciones de detección, comunicación y control. Entre ellas encontramos:

a) Funciones descriptivas: informan sobre el estado de un objeto, sistema o proceso basándose en datos, incluso en tiempo real.

b) Funciones diagnósticas: interrogan y analizan datos para identificar problemas y causas raíz en el funcionamiento de sistemas o procesos.

c) Funciones predictivas: examinan datos que conduzcan a la adopción de acciones preventivas.

d) Funciones prescriptivas: utilizan sistemas de IA para la implementación y aplicación de medidas proactivas.

e) Funciones de automatización: las actividades operativas y de optimización se automatizan a través de un flujo de información bidireccional entre el elemento físico y el digital.

f) Funciones de colaboración: centralizadas en un *hub* que mejora la colaboración y el intercambio de información.

Las funciones autónomas y basadas en el uso intensivo de datos, que caracterizan a los gemelos digitales, ponen de manifiesto la necesidad de una sólida convergencia en materia de gobernanza de datos, seguridad e interoperabilidad, pero también en cuanto a requerimientos de transparencia, explicabilidad y rendición de cuentas por la presencia de algoritmos basados en IA.

La interacción entre las arquitecturas técnicas y los marcos jurídicos plantea, pues, un reto de cumplimiento normativo multidimensional. Un mismo gemelo digital puede estar sujeto simultáneamente a múltiples regulaciones en función de sus tipos de datos, métodos de modelización y patrones de conectividad. La gestión de estos requisitos superpuestos exige un enfoque unificado que vincule las dimensiones técnica, organizativa y jurídica en el desarrollo, implementación y funcionamiento de estos sistemas. El *compliance by design* se convierte no solo en una necesidad normativa, sino también en una apuesta técnica para garantizar la rendición de cuentas, la interoperabilidad y la fiabilidad a lo largo de todo el ciclo de vida del gemelo digital.

Todo ello cobra vital importancia ante el uso de la representación digital del cuerpo humano y sus procesos, creada con fines de modelización biomédica y obtenida mediante el procesamiento de datos genéticos, la medición de valores fisiológicos o la información sobre estilo de vida. Estos gemelos digitales entendidos como un paso adelante hacia la medicina personalizada, predictiva y preventiva suponen también una oportunidad para repensar el enfoque del Derecho, con el fin de adaptarlo a la evolución científica basada en la *Deep Tech*.

En este último escenario se plantean aún mayores retos tanto por la sensibilidad especial de los datos como por la consideración de los supuestos y conceptos en los que se basa el sistema jurídico: la persona, la identidad, la voluntad, la propiedad, el libre albedrío… ¿Cómo delimitar una nueva frontera entre lo que poseo y lo que soy, es decir, entre el ser y el tener?

Asimismo, se vuelven a presentar cuestiones éticas esenciales como la necesidad de revisar el consentimiento informado para la recopilación y el uso de los datos, la gestión sobre la propiedad y el tratamiento de los mismos, la protección de la autonomía de los sujetos y el conocimiento de las limitaciones legales. Todo ello, con el objetivo de promover la confianza y preservar los principios morales que puedan también disminuir las desigualdades en materia de salud.

Neurotecnologías, *por P. Llaque Gálvez*

En las últimas décadas, el avance tecnológico y la evolución en el conocimiento del cerebro y el sistema nervioso han ampliado las posibilidades de integrar la evidencia más reciente sobre la naturaleza de las funciones cognitivas,

emocionales, volitivas y sociales en dispositivos y procedimientos, denominados colectivamente «neurotecnologías». Es así como la neurotecnología se consolida como un habilitador que nos permite comprender e influir en la estructura y el funcionamiento del sistema nervioso, a través de su acceso -invasivo o no invasivo-, supervisión, evaluación, emulación y/o modulación.

Al igual que muchos campos de la tecnología avanzada, la neurotecnología es compleja y multifacética. Además, es intrínsecamente interdisciplinar. La ciencia abarca mucho más que el conjunto de conocimientos científicos, así como la tecnología trasciende el abanico de instrumentos, artefactos, técnicas y sistemas técnicos. Como resultado, la neurotecnología desempeña un papel esencial como respuesta al creciente desafío que plantean los trastornos neurológicos y la salud mental y se instaura como complemento fundamental de las intervenciones farmacológicas y psicosociales.

Ahora bien, la creciente accesibilidad y portabilidad de las tecnologías del cerebro, junto con sus posibles aplicaciones en la vida cotidiana, han llevado su uso más allá del ámbito de la investigación y médico, incrementando su presencia en el mercado de consumo general bajo técnicas de estimulación cerebral y *neurofeedback* (por ejemplo, en el trabajo, el bienestar, la educación, el entretenimiento, el marketing).

Los productos sanitarios requieren para su comercialización cumplir estrictos requisitos de conformidad y seguridad, mientras que los destinados directamente al consumidor no están sujetos a dichas regulaciones. Esto plantea cuestiones éticas y sociales cruciales, especialmente en lo que respecta a la mejora humana, la protección de los datos neuronales y la vulnerabilidad de la identificación de los patrones cognitivos ante las técnicas de datificación y de manipulación.

Todo ello atiende a la importancia y el significado del cerebro, base de nuestro sentido del yo, de la identidad individual. Las neurotecnologías prometen ofrecer una visión profunda y detallada de nuestra actividad cerebral y revelar los pensamientos y sentimientos más íntimos, incluidos aquellos que no se traducen en acciones o comportamientos y que, por lo tanto, no pueden medirse ni inferirse a partir de los datos recopilados mediante otras tecnologías. La aplicación de la IA aumenta enormemente el nivel de comprensión que se obtiene de los datos neuronales, especialmente si se combinan con otros datos contextuales. Además, la neuromodulación y los implantes cerebrales allanan el camino hacia la posibilidad de influir y reescribir la actividad cerebral y la identidad personal.

Como tales, las neurotecnologías representan una intromisión sin precedentes, quizás incluso el paso definitivo, en la esfera privada de los individuos.

Asimismo, pueden interferir con otros derechos y libertades fundamentales más allá de los derechos a la privacidad y a la protección de datos.

A pesar de la normativa europea y las recomendaciones internacionales o los instrumentos de *soft law* existentes, persisten importantes lagunas y zonas grises en la gobernanza y el desarrollo de las neurotecnologías, que deben ser abordadas.

Al analizar el posible impacto de la neurotecnología desde el Derecho y la profesión jurídica es imperativo comprender la estructura básica de la ciencia y de la tecnología. Ello ofrece el instrumental que permite analizar los problemas éticos en relación a la evaluación de la neurotecnología, facilitando la comprensión de cómo hacer juicios que incorporen diferentes tipos de valores, beneficios y riesgos.

En la actualidad, a la ciencia y a la tecnología ya no se les puede atribuir «neutralidad valorativa», ya que ambas se entienden como constituidas por sistemas de acciones intencionales. Es así como los intereses, los objetivos, los valores y las normas se constituyen como sistemas, por lo que se los considera aptos para una evaluación moral. Por ello, los problemas éticos asociados a la tecnología no se reducen a la finalidad de su uso -para bien o para mal-. Ella es el resultado de las intenciones de los agentes que la diseñan, de los objetivos que persiguen, de los valores que subyacen y de los medios que utilizan.

Inteligencia Biológica Sintética (*Synthetic Biological Intelligence*, SBI)

por J. A. Nisa Ávila

La Inteligencia Biológica Sintética (SBI) es un tipo de computación e inteligencia híbrida que combina sistemas biológicos vivos basados principalmente en neuronas, tejidos neuronales, organoides cerebrales, biomoléculas y/o circuitos celulares con componentes digitales, electrónicos y algorítmicos para generar capacidades de aprendizaje, memoria, adaptación y procesamiento de información.

A diferencia de la IA tradicional, basada exclusivamente en *hardware* de silicio y algoritmos matemáticos, la SBI utiliza elementos biológicos reales para optimizar el procesamiento computacional. De esta forma el sistema no solo simula funciones cognitivas, sino que incorpora directamente mecanismos biológicos de plasticidad, autoorganización y aprendizaje e incluso elementos editados genéticamente, como en el cómputo molecular.

La Inteligencia Biológica Sintética constituye la matriz conceptual de varios campos emergentes:

- inteligencia de organoides (*organoid intelligence*),

- *wetware computing*,
- computación molecular,
- computación basada en ADN,
- sistemas neuroelectrónicos híbridos.

En todos estos casos, el objetivo es aprovechar las propiedades naturales de los sistemas biológicos para realizar tareas computacionales.

Por ejemplo:

- En la computación molecular, moléculas como ADN, ARN o proteínas se utilizan para almacenar y procesar información.
- En la computación biocibernética, tejidos neuronales se integran con sensores, electrodos y sistemas digitales.
- En la inteligencia de organoides, mini-cerebros cultivados en laboratorio se conectan a interfaces electrónicas para desarrollar tareas de aprendizaje y decisión.
- En el *wetware computing*, neuronas vivas funcionan como procesadores biológicos.

El principal reto jurídico del SBI radica en que difumina la frontera tradicional entre organismo y máquina. Este fenómeno obliga a reconsiderar las categorías de sujeto y objeto de derecho, así como el tratamiento jurídico del *hardware* desarrollado con tejido vivo y del aprovechamiento de su inteligencia natural para crear sistemas de IA.

La SBI no encaja completamente en las categorías jurídicas tradicionales de *software*, máquina o producto biomédico, ya que incorpora material biológico potencialmente dotado de aprendizaje, memoria y capacidad adaptativa.

Esto plantea cuestiones regulatorias profundas:

- ¿Qué estatuto jurídico tienen los organoides utilizados para computación?
- ¿Puede existir responsabilidad jurídica por daños causados por sistemas biohíbridos?

- ¿Qué ocurre si un sistema basado en tejido neuronal desarrolla formas rudimentarias de sensibilidad o conciencia?
- ¿Cómo se regula el uso de células humanas para fines computacionales?
- ¿Debe tratarse como un dispositivo médico, un sistema de IA o una entidad bio-tecnológica autónoma?

Desde el punto de vista del Derecho Cuántico, la SBI representa una forma de inteligencia distribuida y emergente en la que el procesamiento ya no se basa exclusivamente en lógica formal, sino en dinámicas biológicas complejas, adaptativas y no lineales donde el factor determinante es la capacidad de toma de decisiones con tejido neuronal humano modificado, al permitir la hibridación con máquinas.

Computación biocibernética, *por J. A. Nisa Ávila*

La computación biocibernética es un sistema de *hardware* en el que se basan arquitecturas SBI de procesamiento de información que integra sistemas biológicos vivos con componentes electrónicos, digitales y algorítmicos para crear sistemas híbridos capaces de percibir, procesar, aprender y responder al entorno. Su lógica fundamental consiste en conectar sistemas biológicos con sistemas digitales para que ambos interactúen de forma bidireccional. Su objetivo no es únicamente aumentar la potencia computacional, sino aprovechar propiedades para solucionar problemas de los sistemas electrónicos tradicionales.

La computación biocibernética introduce una ruptura con la división clásica entre sujeto biológico y objeto tecnológico. Los sistemas biohíbridos dejan de ser simples herramientas pasivas y pasan a incorporar componentes vivos que aprenden, se reorganizan y reaccionan a estímulos.

Desde el enfoque del Derecho Cuántico, la computación biocibernética puede entenderse como una forma de inteligencia distribuida en la que la frontera entre lo biológico y lo artificial se vuelve difusa y las fronteras entre la tecnología de silicio y elementos vivos se unen para dar paso a sistemas que cuando fallecen o dejan de funcionar no sirven. La regulación jurídica de entes vivos asociados a tareas tecnológicas, pero sin capacidad de autopercibimiento, es una frontera que no se puede determinar con facilidad cuándo se cruzará y cómo.

Cómputo molecular, *por J. A. Nisa Ávila*

El cómputo molecular es un paradigma de procesamiento de información en el que moléculas biológicas o químicas basadas en ADN, ARN, proteínas, enzimas o reacciones químicas programables se usan para almacenar, transmitir y procesar datos. A diferencia de la computación clásica basada en transistores y circuitos electrónicos, el cómputo molecular emplea interacciones físico-químicas como soporte de operaciones lógicas y cálculo.

Desde una perspectiva jurídica, el cómputo molecular resulta relevante porque desdibuja la frontera entre dispositivo tecnológico y entidad biológica. En estos sistemas, la capacidad de cálculo no se encuentra en un *software* tradicional, sino en la propia materia viva o química. El cómputo molecular es especialmente importante porque representa una evolución radical del concepto de red neuronal y de computación. El procesamiento de información deja de depender exclusivamente de circuitos de silicio y pasa a integrarse en estructuras materiales vivas o químicas.

Ciberseguridad, *por D. Gómez Pérez*

La seguridad, según el diccionario panhispánico del español jurídico, se refiere a la cualidad de ser seguro, libre y exento de todo peligro, daño o riesgo. Al añadirle el prefijo ciber, se refiere «al conjunto de actividades dirigidas a proteger el ciberespacio contra el uso indebido del mismo, defendiendo su infraestructura tecnológica, los servicios que prestan y la información que manejan», según la Orden Ministerial 10/2013, de 19 de febrero, por la que se crea el Mando Conjunto de Ciberdefensa de las Fuerzas Armadas.

El término ciberespacio engloba cualquier infraestructura de tecnología de la información que genere una capa de información superpuesta sobre el espacio físico, siendo la red Internet el ejemplo más significativo, pero también se incluyen las redes que procesan y transmiten información a otras infraestructuras, como el servicio de agua, la red eléctrica, las redes de geolocalización, etc.

Garantizar un uso libre de riesgos en este espacio, en el que interactúan actores de distinto tipo, es una tarea ardua. El tipo de actores incluye desde ciudadanos particulares hasta entidades jurídicas o gobiernos nacionales, siendo su clasificación no estricta. Más aún, la popularización de los dispositivos móviles ha hecho que se ofrezca el acceso a más servicios a través del ciberespacio y que los usuarios interactúen desde distintas localizaciones y a través de distintas redes de interconexión.

En este nuevo escenario aparecen nuevos organismos y empresas que buscan delimitar espacios seguros o informar de amenazas, con el objetivo de proteger la información intercambiada en estos sistemas contra daños como pérdidas de acceso o confidencialidad y garantizando su integridad ante manipulaciones. Dentro de los organismos estatales se encuentran el Mando Conjunto del Ciberespacio de España (MCCE), que está encargado de la defensa nacional, y el Instituto Nacional de Ciberseguridad de España (INCIBE), cuya misión es prestar ayuda a empresas y particulares ante amenazas de ciberseguridad y garantizar la confianza digital de los ciudadanos. Por último, cabe mencionar el Centro Criptológico Nacional (CCN), que es el responsable de garantizar la ciberseguridad de las administraciones públicas, elaborar y difundir normativa sobre el uso de las tecnologías de la información y la estandarización de productos informáticos.

Por el desarrollo tan rápido de las tecnologías de la información y las nuevas amenazas provenientes de la IA, la seguridad no puede ser absoluta, sino que se imponen una serie de contramedidas que permitan tener confianza suficiente en los sistemas y poder controlar los riesgos. Es inevitable la aparición de incidentes, por lo que es necesario establecer planes para poder detectar, reaccionar y recuperarse de los incidentes que comprometan el funcionamiento de la infraestructura de la información. Los «*Common Criteria*» son un marco específico que mide la seguridad de las infraestructuras tecnológicas. Este marco evalúa el proceso de desarrollo en siete niveles, que se denotan por el acrónimo EAL, del inglés *Evaluation Asurance Level*. Estos niveles garantizan que el producto ha superado pruebas, reservándose los niveles más altos y, por tanto, exhaustivos en tiempo y esfuerzo de las pruebas, a productos dedicados a uso militar o sistemas de seguridad nacional como la identificación transnacional, por ejemplo.

La ciberseguridad engloba también a las personas, que en última instancia son los usuarios últimos del ciberespacio a través de sus interacciones, tanto entre personas como entre máquinas, siempre a través del uso de las tecnologías de la información. El factor humano representa un amplio campo de estudio diferenciado. Cuando es un usuario el principal factor en la amenaza, se estudia primeramente el rol de este y si el atacante es interno o externo, que delimita sus capacidades de actuación. Si es interno, se distingue su participación, si es activo o ha sido robada su identidad mediante alguna técnica de ingeniería social. Las más conocidas son: la suplantación de credenciales («*piggybagging*»), recogida de desechos de información («*Trashdiving*»), la escucha de conversaciones privadas («*eavesdropping*»), uso de puestos de trabajo desantendidos («*Office Snooping*») o envío de mensajes fraudulentos en nombre de una persona o entidad confiable («*phising*»).

II. Fundamentos tecnológicos de la IA

Algoritmo, *por D. García Saiz*

En su definición más pura y técnica, no es más que un conjunto finito, ordenado y lógico de instrucciones diseñadas específicamente para resolver un problema, procesar datos o llevar a cabo una tarea determinada. Lejos de ser un invento exclusivamente informático, es una estructura de pensamiento que aplicamos constantemente en nuestra vida diaria. Si alguna vez has seguido un manual de instrucciones para montar una estantería o una receta para preparar un bizcocho, has ejecutado uno. La única diferencia es que, en el ámbito digital, las instrucciones se traducen a un código matemático que una máquina puede procesar a velocidades asombrosas.

Para que un sistema informático funcione bajo esta lógica deductiva, la información siempre fluye a través de un esquema inalterable compuesto por tres fases fundamentales que determinan su comportamiento y utilidad:

a) La entrada de datos (*input*): Es la materia prima o la información inicial que el sistema recibe para poder trabajar. Dependiendo de la función del programa, puede ser un simple número de cuenta bancaria, el texto redactado en un buscador web, la temperatura detectada por un sensor o las coordenadas de tu ubicación.

b) El procesamiento computacional: Es el núcleo del sistema, donde se aplican las reglas lógicas y matemáticas predefinidas sobre esos datos iniciales. Es el momento de la ejecución pura, donde el programa evalúa las condiciones; por ejemplo: "Si el usuario tiene más de 18 años, reside en una zona concreta y su historial de compras supera una cantidad X, entonces calcula y aplica un descuento del 10%".

c) La salida de resultados (*output*): Es la consecuencia final del proceso, la decisión tomada o la tarea ejecutada después de haber aplicado las reglas a los datos de entrada. Esto se materializa en acciones como mostrar el precio final de un producto en la pantalla, encender una alerta luminosa o bloquear preventivamente una tarjeta de crédito.

Dentro de este esquema de funcionamiento, existe una división técnica y conceptual primordial que determina cómo se comportan estos sistemas informáticos y qué impacto real tienen en la sociedad. Tradicionalmente, la inmensa mayoría de los programas se han construido basándose en reglas deterministas. En este modelo clásico un programador humano redacta manualmente cada una

de las instrucciones, los pasos y las variables. Es un sistema altamente predecible, transparente y rastreable. Si utilizas una aplicación de navegación GPS y el sistema te guía erróneamente hacia una calle sin salida o un acantilado, un ingeniero puede abrir las líneas de código del sistema, revisar la secuencia lógica y encontrar el punto exacto donde se produjo el error de cálculo o la instrucción fallida. Ante los mismos datos de entrada, este tipo de sistema siempre arrojará exactamente el mismo resultado, y cada decisión está justificada por una instrucción tecleada por un ser humano. Sin embargo, el cambio de paradigma tecnológico que estamos experimentando se debe a la irrupción de los sistemas probabilísticos, comúnmente conocidos como sistemas de aprendizaje automático (*machine learning*). En estos sistemas los ingenieros ya no escriben las reglas lógicas de forma manual, porque el problema que intentan resolver es demasiado grande, abstracto o complejo para resumirlo en un puñado de instrucciones deterministas. En lugar de darle a la máquina las reglas del juego cerradas, le proporcionan montañas masivas de ejemplos históricos para que aquella descubra las reglas y las correlaciones por sí misma.

Si se desea que un programa de ordenador aprenda a distinguir la fotografía de un tumor maligno de uno benigno, ya no se le dan instrucciones geométricas sobre el tamaño, el color o la forma de las células. Lo que se hace es introducirle millones de escáneres médicos previamente diagnosticados, permitiendo que su estructura de procesamiento infiera, a base de ensayo y error continuo, los pequeñísimos patrones visuales que definen la enfermedad y que a menudo son invisibles para el ojo humano. El sistema no es estático ni cerrado; su comportamiento evoluciona y su precisión se ajusta continuamente a medida que procesa, asimila y entrena con más y más datos.

Esta asombrosa capacidad de aprendizaje autónomo es el motor indiscutible detrás de los avances tecnológicos más espectaculares de nuestra era, pero simultáneamente genera uno de los desafíos éticos y sociales más profundos: el problema de la "caja negra". Cuando estos sistemas de aprendizaje se vuelven extremadamente complejos, especialmente aquellos que intentan simular el funcionamiento del cerebro humano mediante múltiples capas de redes neuronales artificiales, su proceso interno de toma de decisiones se vuelve completamente opaco. Alcanzan un nivel de abstracción matemática en el que ni siquiera los propios ingenieros e investigadores que diseñaron el programa original son capaces de explicar con certeza por qué el sistema tomó una decisión específica frente a un caso concreto. La máquina emite un resultado predictivo altamente preciso, pero el razonamiento de fondo que la llevó hasta esa conclusión resulta inescrutable y ajeno a la lógica humana.

Las consecuencias prácticas de desplegar estas "cajas negras" matemáticas en el mundo real son inmensas e irreversibles, ya que hoy en día estos sistemas

automatizados no se limitan a clasificar fotografías, sino que se utilizan como jueces silenciosos para tomar decisiones que afectan drásticamente la vida, los recursos y las oportunidades de las personas. Son estos complejos entramados los que están decidiendo en milisegundos si una entidad financiera te concede o no una hipoteca para tu primera vivienda, si tu currículum académico es seleccionado para una entrevista o es descartado fulminantemente en la primera fase de una criba masiva de empleo, qué tipo de noticias o vídeos te muestra una red social para retener tu atención o moldear tu opinión, e incluso, en algunas jurisdicciones modernas, qué probabilidad estadística tiene un acusado judicial de reincidir en un delito grave para denegarle la libertad condicional.

El riesgo más crítico, invisible y persistente asociado a esta delegación de poder es la aparición y perpetuación de los sesgos heredados. Dado que estos sistemas probabilísticos aprenden exclusivamente de datos históricos generados a lo largo de los años por sociedades humanas imperfectas, asimilan de forma ciega todos los patrones presentes en esa información, incluidos nuestros peores defectos, prejuicios y discriminaciones estructurales. Si se entrena a un avanzado sistema de selección de personal utilizando el archivo histórico de contrataciones de una gran corporación que, durante las últimas tres décadas, priorizó sistemáticamente a hombres sobre mujeres para ascender a puestos directivos, el sistema informático no tiene la capacidad de cuestionar la moralidad o la legalidad de ese patrón. Simplemente inferirá estadísticamente que "ser hombre" es una variable intrínseca de éxito corporativo y, en consecuencia, comenzará a descartar automáticamente y con extrema eficiencia los currículums de mujeres brillantes. Así, lejos de ser árbitros fríos, objetivos e infalibles como solemos creer, estos sistemas corren el riesgo de convertirse en formidables amplificadores de la desigualdad, automatizando, acelerando y ocultando la discriminación social bajo una peligrosa y falsa apariencia de neutralidad matemática inobjetable.

Algoritmo evolutivo, *por J. I. Carbo Rubiera y C. Pérez García*

Los algoritmos evolutivos son uno de los tres principales campos de estudio de la IA, junto con lo que a grandes rasgos se puede llamar IA lógica o simbólica y el aprendizaje automático basado en datos. Se trata de un amplio conjunto de sistemas de optimización de variables en entornos en que otro tipo de programación no resulta capaz de ofrecer resultados en tiempo polinómico, es decir, con un uso razonable de los recursos de *hardware*. Reciben el nombre «evolutivo» porque estos algoritmos se inspiran en los procesos biológicos de selección natural por medio de entrecruzamiento, competencia y prevalencia, que replican de una manera lejana. Su principal paradigma, junto con la programación evolutiva y las estrategias evolutivas, son los algoritmos genéticos, si bien las estrategias más refinadas acostumbran a entrelazar diferentes técnicas.

El esquema de ejecución básico de cualquier algoritmo evolutivo comienza con una generación aleatoria de una población de individuos dentro del entorno en que se busca la optimización. Puede tratarse, por ejemplo, de la distribución de placas solares sobre un terreno acotado. En este caso, el entorno sería el modelo informático del campo con todas las variables orográficas, climáticas y de cualquier otra naturaleza que puedan influir de alguna manera en el cálculo de optimización. La población, por su parte, sería el conjunto de diferentes distribuciones de placas solares; cada distribución sería un individuo. Dentro de esta población se evalúa, para cada individuo, su desempeño conforme a la función de aptitud predefinida, que debe proveer a los individuos con incentivos para guiarlos, y se los jerarquiza según su resultado, es decir, según su valor de aptitud. En este caso, se evaluaría la producción energética de cada concreta distribución de placas solares y se ordenarían de mayor a menor. El siguiente paso es seleccionar a los mejores individuos de la población y, en una segunda generación, replicarlos como padres de la nueva población. De esta manera, los individuos de la segunda generación tienen, en conjunto, un mayor valor de aptitud que los de la primera, pues replican a los mejores individuos de esta. Sin embargo, para activar el proceso evolutivo y poner en marcha la optimización, a los individuos replicados se les introducen mutaciones y cruces con los que se pretende evitar que el algoritmo converja en óptimos locales. Como resultado, si bien algunos muestran un peor desempeño, también otros lo mejoran. Es a estos mejores individuos a los que, una vez evaluada la segunda población, se replica en la generación siguiente, y así de manera sucesiva hasta cumplir el criterio de parada; por lo general, hasta llegar a un determinado número de repeticiones, hasta obtener un valor de aptitud considerado suficiente o hasta que el porcentaje de mejora sea demasiado bajo.

La principal limitación de los algoritmos evolutivos, en general, es su autorreferencialidad, pues el marco de referencia para evaluar el desempeño de los individuos es la propia población a la que pertenecen. Así, el resultado definitivo del algoritmo evolutivo, esto es, el que se tendrá por óptimo al cabo de la enésima generación, es el que ha obtenido el mayor valor de aptitud en la función, es decir, el que ha logrado el objetivo marcado para la población con el menor consumo de recursos. En el ejemplo antedicho, sería la distribución de placas solares que, dadas las limitaciones impuestas por el entorno, generase mayor cantidad de energía. Sin embargo, el criterio para determinar la optimalidad de un individuo no es externo al algoritmo, ergo objetivo, sino interno, ergo relativo, de suerte que siempre se puede reproducir una plusenésima generación de la que obtener un individuo que, por comparación con sus pares, alcance un mayor valor de aptitud. Así, la detención del algoritmo evolutivo es siempre, en cierto modo, arbitraria, pues el resultado óptimo solo es tal conforme a las generaciones ejecutadas y no conforme a un criterio externo que permita determinar tal condición de manera objetiva. Ni siquiera se puede saber si el resultado del algoritmo se puede optimizar

más o no: su ejecución se detiene cuando el mejor individuo se estima suficiente o cuando seguir ejecutando el algoritmo se calcula ineficiente; en definitiva, cuando se cumple el criterio de parada, no cuando se alcanza el punto óptimo.

Algoritmo genético, *por J. I. Carbo Rubiera y C. Pérez García*

Los algoritmos genéticos son el más popular entre los paradigmas de los algoritmos evolutivos. Como sucede con estos, toman el nombre «genético» de su mímesis del proceso evolutivo basado en selección y transmisión de atributos valiosos y criba y descarte de atributos disvaliosos. Los atributos de los individuos dentro de cada población se denominan «cromosomas» y, por lo general, se gestionan de manera binaria. Esto es lo más característico de los algoritmos genéticos, en contraste con otros paradigmas de los evolutivos. Las mutaciones y entrecruzamientos entre individuos de una generación a la siguiente se operan alterando el valor binario de sus cromosomas. Por lo demás, se replica el esquema básico de todo algoritmo evolutivo: generación de una población, evaluación conforme a la función de aptitud, selección de los mejores individuos, mutación y reproducción de la siguiente generación.

A la limitación general de todo algoritmo evolutivo, los genéticos añaden que, debido a que se basan en una estructura cromosómica binaria, tienen un desempeño poco eficiente en problemas de alta complejidad. En los casos en que sí logran ofrecer resultados, además, la alta complejidad puede llevar a los cromosomas a converger de manera apresurada en torno a puntos óptimos locales, dejando de lado otras soluciones óptimas con mayor valor de aptitud. En fin, estos algoritmos tampoco responden bien a la complejidad diacrónica, es decir, a entornos cambiantes. La adaptación cromosómica del algoritmo genético al medio es tan alta que cualquier cambio en el entorno de referencia puede dejar obsoletos, por inadaptados, a los previos individuos óptimos.

Empero, es esta misma estructura cromosómica binaria la que convierte a los algoritmos genéticos en instrumentos sobresalientes para resolver problemas de programación de tareas y gestión de procesos, es decir, problemas en los que se define un conjunto finito y acotado de objetos que pueden estar verificados o falsados, en los que, por tanto, la misma estructura binaria del algoritmo es parte del objeto con que debe éste trabajar.

Desde un punto de vista jurídico, el proceso evolutivo de aprendizaje de los algoritmos genéticos puede representar un problema en términos de seguridad, pues el código informático tiende, por el propio proceso de selección genética, a separarse de cualquier traducción originaria del código jurídico que se estime subóptima. Sin embargo, los algoritmos genéticos sí pueden hallar un importante campo de desarrollo en el terreno de las reglas técnicas, es decir, en la optimización

de los procesos aplicativos de las reglas jurídicas, buscando la manera óptima de aplicar la ley sin implicar alteraciones internas de su contenido o estructura. En el fondo, todo depende de una correcta definición de la función de aptitud. Lo paradójico es que, si se supiese cuál sería la función de aptitud perfecta para un determinado problema, entonces, por ello mismo, se sabría tanto acerca del problema que el propio algoritmo sería innecesario.

Función de aptitud, *por J. I. Carbo Rubiera y C. Pérez García*

La función de aptitud es un tipo de función de costes que muestra, por medio de un simple valor numérico, el grado en que un determinado individuo ha satisfecho un conjunto de objetivos. El nombre *fitness function* hace referencia a este uso como medida del desempeño de cada individuo por medio de un *fitness value* o "valor de aptitud". Las referencias suelen conservar el inglés *fitness*, "aptitud", tanto en el sintagma completo *fitness function* como en traducciones parciales, como "función de *fitness*" o solo "función *fitness*", así como traducir el sintagma completo, por lo general, como "función de aptitud", "función de adecuación" o "función objetivo". Del mismo modo, el valor de aptitud puede encontrarse también referenciado como "valor de *fitness*".

El principal uso de la función de aptitud se da en el campo de los algoritmos evolutivos y, en particular, de los algoritmos genéticos. Aquí, la función sirve para evaluar el grado en que los individuos de cada población son capaces de cumplir los objetivos marcados. Para ello, además del sistema pasivo de evaluación, la función suele prever mecanismos de incentivos para guiar a los individuos. La función de aptitud, por tanto, es la base sobre la que se asienta el proceso de selección y descarte de individuos en las generaciones de los algoritmos evolutivos y, al mismo tiempo, el criterio de evaluación de las sucesivas mutaciones. A veces también se utiliza el valor de aptitud para alterar el cruce y la mutación.

La función de aptitud puede mantenerse invariable durante todo el proceso evolutivo o coevolucionar con él. En los algoritmos diseñados para solucionar problemas de optimización o para aplicarse en entornos fijos se suelen emplear funciones no cambiantes. En cuanto al valor de aptitud, si bien puede ser un valor cardinal absoluto, también es posible que se trate de una medida ordinal relativa, si la función de aptitud se limita a comparar y ordenar los individuos según su desempeño.

Código fuente, *por D. García Saiz*

Todo lo que vemos y experimentamos al utilizar un dispositivo electrónico, desde los botones coloridos de una aplicación móvil hasta los elaborados gráficos

de un videojuego o las celdas de una hoja de cálculo, es únicamente la superficie visible, la fachada de un inmenso edificio digital. Detrás de esa interfaz visual con la que interactuamos se esconde el verdadero "manuscrito", el ADN estructural que dicta el comportamiento de la máquina en cada milisegundo. Se trata del "código fuente", que contiene el conjunto de instrucciones legibles por humanos, escritas por un programador, que le indican a un *software* u ordenador cómo debe comportarse.

Para comprender la naturaleza de ese manuscrito debemos tener en cuenta la existencia de una profunda barrera de comunicación: en su nivel físico más fundamental, los ordenadores son máquinas eléctricas que carecen de la capacidad de comprender el lenguaje humano. Únicamente son capaces de procesar impulsos eléctricos que se representan matemáticamente como ceros y unos, lo que se conoce como el lenguaje de máquina. Dado que para un cerebro humano resulta prácticamente imposible redactar programas complejos escribiendo interminables secuencias de ceros y unos, los pioneros de la informática inventaron lenguajes intermedios. Estos lenguajes de programación utilizan palabras (generalmente en inglés), símbolos matemáticos y estructuras lógicas que una persona entrenada puede leer, escribir y comprender. Este texto humano y comprensible es, precisamente, el conjunto de instrucciones originales que redacta el programador. Sin embargo, como el ordenador sigue sin entender ese texto, una vez que el manuscrito está terminado, se utiliza un programa traductor especial que se encarga de convertir todas esas palabras y reglas lógicas en los bloques de ceros y unos que el procesador finalmente ejecutará.

La mejor analogía para entender el inmenso valor de estas instrucciones originales es pensar en los planos arquitectónicos de un rascacielos o en la receta secreta de un famoso refresco de cola. Si vas a un supermercado y compras una botella de ese refresco, puedes beberlo y disfrutar de su sabor, pero no tienes acceso a la lista exacta ni a las proporciones milimétricas de los ingredientes que lo componen. En el ecosistema digital tradicional ocurre exactamente lo mismo. Cuando compras o descargas un programa informático, la empresa creadora no te entrega sus "planos arquitectónicos", sino el producto final ya traducido al lenguaje de la máquina. Este modelo cerrado y altamente restrictivo se basa en el secretismo y el control corporativo absoluto: al ocultar el manuscrito original, los creadores se aseguran de que nadie más pueda copiar su trabajo, pero también impiden que cualquier investigador externo pueda auditar qué hace realmente el programa con tus datos privados o descubrir fallos de seguridad ocultos.

Frente a esta cultura del hermetismo, surgió de manera paralela un movimiento tecnológico revolucionario con una filosofía diametralmente opuesta, promoviendo el desarrollo de **código abierto**. ¿Qué pasaría si, en lugar de esconder los planos del edificio en una caja fuerte corporativa, los arquitectos

decidieran publicarlos en una biblioteca mundial para que cualquier persona pudiera estudiarlos, criticarlos y mejorarlos de forma libre? Esta es la premisa fundamental del modelo de desarrollo abierto y colaborativo, una forma de concebir la tecnología donde el secreto industrial desaparece en favor del progreso colectivo.

Cuando los creadores de un programa deciden acogerse a este modelo abierto, publican su código en internet y otorgan al mundo entero una serie de libertades fundamentales que transforman por completo la relación entre el usuario y la tecnología:

a) La libertad de estudio y auditoría: Cualquier persona con los conocimientos técnicos adecuados puede abrir el capó del programa y leer exactamente cada línea de las instrucciones originales. Esto garantiza una transparencia total, impidiendo que el *software* contenga funciones maliciosas ocultas, como herramientas de espionaje o puertas traseras que comprometan la privacidad.

b) La libertad de modificación: Si un usuario descubre un error en el funcionamiento del programa, o si necesita que realice una tarea nueva y específica que los creadores originales no habían previsto, no tiene que esperar pasivamente a que la empresa decida lanzar una actualización. Tiene el poder y el permiso para reescribir las instrucciones y adaptar el programa a sus necesidades exactas.

c) La libertad de distribución y mejora colectiva: Lejos de prohibir la copia, este modelo anima a los usuarios a compartir libremente tanto el programa original como las versiones mejoradas que hayan creado, permitiendo que toda la comunidad se beneficie de las innovaciones individuales.

Este enfoque basado en la transparencia y la cooperación no es una utopía marginal ni un simple experimento académico, sino que se ha convertido en el pilar invisible que sostiene la mayor parte de la infraestructura tecnológica moderna. Al estar las instrucciones expuestas al escrutinio público, en lugar de contar con un pequeño equipo de diez ingenieros en una oficina cerrada intentando encontrar errores, el sistema se beneficia de la mirada crítica de miles de expertos independientes repartidos por todo el mundo, quienes auditan el código simultáneamente. Cuando alguien detecta una vulnerabilidad de seguridad crítica, un voluntario al otro lado del planeta suele corregirla en cuestión de horas. Gracias a esta asombrosa eficiencia colaborativa, la inmensa mayoría de los servidores que mantienen viva la red global de internet, los superordenadores que realizan investigaciones científicas de vanguardia e incluso

los sistemas operativos que hacen funcionar miles de millones de teléfonos móviles inteligentes en la actualidad, están construidos sobre cimientos tecnológicos completamente abiertos.

Aprendizaje automático (*Machine Learning*), *por M. de Asís Pulido*

El *Machine Learning* (ML) o aprendizaje automático constituye el subconjunto de la IA centrado en el desarrollo de algoritmos capaces de aprender los patrones subyacentes a un *dataset* y, a partir de dicho aprendizaje, realizar inferencias sobre nuevos *inputs*. En otras palabras, las estrategias de ML permiten a un sistema mejorar su desempeño en una tarea determinada sin necesidad de ser programado de forma explícita para tal tarea. De esta manera, el ML supone un desplazamiento parcial del enfoque clásico de la IA simbólica hacia modelos que aprenden funciones a partir de la *experiencia*, entendida esta como exposición iterativa a conjuntos de datos.

En este contexto, *aprendizaje* significa la posibilidad de mejorar el rendimiento a través de la experiencia -de datos-, de forma que el modelo de IA identifique una función que, ajustándose sobre un conjunto de entrenamiento (muestra), pueda generalizarse a una población. Dicha función quedará definida sobre atributos o características -identificados por el humano o por el propio modelo- ponderados por parámetros ajustables o *weights* (pesos). Durante el entrenamiento, estos parámetros se modifican iterativamente mediante procedimientos que capturan relaciones estadísticas entre los atributos de los datos y, en su caso, entre estos y resultados determinados (*labels* o etiquetas), minimizando una medida de error. El proceso de aprendizaje consiste, pues, en la optimización de parámetros mediante la minimización de una función de pérdida o coste, que mide la diferencia entre el *output* del modelo y el resultado esperado. El modelo aprendido se aplicará sobre nuevos datos de entrada (*inputs*) para ofrecer predicciones, clasificaciones, regresiones, agrupaciones o generaciones.

El proceso de aprendizaje suele dividirse en tres grandes fases: entrenamiento, validación y prueba, a través de las cuales se trata de garantizar la calidad del ajuste. En función de la estrategia de entrenamiento y del tipo de datos que empleen, los enfoques de ML se clasifican en cuatro grandes metodologías de aprendizaje: *aprendizaje supervisado*, *aprendizaje no supervisado*, *aprendizaje autosupervisado* y *aprendizaje por refuerzo*. En ocasiones se habla también del aprendizaje semisupervisado, que combina estrategias de aprendizaje supervisado y no supervisado.

Por su parte, desde el punto de vista de su arquitectura interna, los modelos de ML pueden adoptar formas muy diversas: árboles de decisión, bosques aleatorios, *Support Vector Machines*, redes bayesianas, redes neuronales, etc. El

paradigma más popular en la actualidad es el *Deep Learning* como subconjunto del grupo de redes neuronales.

Una característica propia del ML es que, una vez entrenado, el comportamiento del programa no podrá ser completamente previsto, lo que tiene implicaciones relevantes en términos de posible opacidad, precisión y atribución de responsabilidad. Conviene aclarar que la selección de los datos de entrenamiento, validación y prueba, de la arquitectura del modelo, de la función de costes y del lenguaje de programación cae bajo la responsabilidad del programador o del científico de datos que se encargue de su diseño. En ese sentido, la mayor o menor autonomía que caracteriza a los sistemas de IA basados en ML es relativa: el modelo no diseña su propio marco de aprendizaje, sino que opera dentro de aquel que le ha sido marcado desde la selección del *dataset*, la función de pérdida, la arquitectura y otros hiperparámetros.

En la actualidad, el ML domina el campo de la IA, constituyendo la base de la mayoría de los sistemas modernos: desde modelos predictivos hasta los incluidos en vehículos autónomos, pasando por modelos de lenguaje como los *Large Language Models* o LLM y otras herramientas de IA generativa. El auge del ML ha coincidido con la explosión del *big data* y con el incremento del poder computacional, pues, en general, y hasta cierto punto, a mayor volumen de datos de entrenamiento y mayor capacidad computacional, mayor será la precisión del modelo.

Aprendizaje supervisado, *por M. de Asís Pulido*

El aprendizaje supervisado constituye la forma más consolidada de aprendizaje automático y consiste en aprender una función a partir de un conjunto de ejemplos vinculados a un resultado conocido, lo que se consigue mediante el mapeo de las correlaciones entre los datos de entrada y sus correspondientes etiquetas de salida (*labels*). A diferencia de otros paradigmas de aprendizaje, en el aprendizaje supervisado es necesario el trabajo previo de clasificación y etiquetamiento de los datos por parte del programador -comúnmente denominado "supervisor" o "profesor"-. Este proporciona al modelo la llamada *ground truth* a través del etiquetamiento del resultado correcto asociado a cada instancia de entrenamiento.

El proceso de aprendizaje supervisado suele articularse en tres fases: entrenamiento, validación y evaluación. En la fase de entrenamiento, se introduce en el modelo un conjunto de datos (*training dataset*) compuesto por instancias -esto es, cada una de las entradas con unos *valores de atributos* determinados- a las que se ha asociado una etiqueta de resultado. A partir de dicho conjunto, el modelo identifica patrones de correlación entre los *training data* y sus resultados,

completando la fórmula que mejor se ajuste a tales patrones mediante la asignación de un *peso* a cada atributo. El proceso de aprendizaje busca reducir el error entre el valor predicho por el modelo en nuevos casos y el valor real asociado a los mismos, a través de la minimización de una función de pérdida o costes, lo que se consigue mediante la modificación iterativa de los parámetros.

En función del tipo de resultado que se pretenda estimar, el aprendizaje supervisado se articula en torno a dos grandes familias de tareas. Cuando el resultado es un valor numérico continuo -como la temperatura de un lugar o el precio de un coche-, la función del modelo recibe el nombre de *regresión*; cuando el resultado es una categoría o clase -como la predicción de un resultado de admisión o inadmisión, el reconocimiento de un objeto en imágenes o la identificación de una transacción fraudulenta-, recibe el nombre de *clasificación*. Muchos algoritmos pueden emplearse indistintamente para ambas tareas o de forma combinada.

La segunda fase, de validación, resulta fundamental para evitar el sobreajuste del modelo (*overfitting*) y para afinar los parámetros mediante el uso de un conjunto de datos reservados específicamente para este fin (*validation set*). El sobreajuste ocurre cuando el modelo aprende los datos de entrenamiento con tal precisión que pierde capacidad de generalización. La tercera y última fase, destinada a la evaluación, introduce nuevos datos sin etiqueta para que el modelo identifique el resultado, midiendo su rendimiento a través de métricas como la precisión o el *F1 score*.

Asimismo, con el fin de que el modelo pueda cuantificar su propia certeza sobre los resultados, puede recurrirse a técnicas como el *resampling* o remuestreo, consistentes en entrenar múltiples modelos sobre porciones aleatorias del *dataset*. La media de los resultados obtenidos podrá utilizarse como estimación del valor esperado, mientras que la dispersión de dichos resultados permitirá aproximar el grado de incertidumbre del modelo. Este análisis puede orientar al supervisor a incorporar nuevos ejemplos de entrenamiento en aquellos casos en los que la incertidumbre sea mayor, contribuyendo así a implementar estrategias de aprendizaje activo.

Un ejemplo paradigmático de sistema de IA entrenado mediante aprendizaje supervisado es el detector de *spam* del correo electrónico. En este caso, el sistema, a partir de un entrenamiento con datos etiquetados, ha aprendido que la presencia de atributos tales como determinadas palabras ("oferta", "oportunidad") o símbolos ("$" o "!") en los correos electrónicos se asocia estadísticamente con la categoría de correo no deseado (*spam*).

Aprendizaje no supervisado, *por M. de Asís Pulido*

El aprendizaje no supervisado es aquella estrategia de *Machine Learning* en la que el modelo aprende a identificar patrones, estructuras y correlaciones intrínsecas a los datos sin que estos estén asociados a ninguna etiqueta de resultado (*label*). A diferencia del aprendizaje supervisado, en el aprendizaje no supervisado no existe ningún "supervisor" que indique al modelo qué resultado debe asociarse a cada instancia: el modelo identifica por sí mismo agrupaciones, dependencias y regularidades latentes en los datos que no son necesariamente evidentes para el observador humano. De ahí que este paradigma resulte especialmente útil en escenarios en los que se dispone de grandes volúmenes de datos no etiquetados.

Los métodos de aprendizaje no supervisado se organizan en torno a tres grandes orientaciones. La primera es el *clustering* o agrupación, cuyo propósito es detectar estructura en los datos reuniendo en un mismo conjunto o clase aquellas instancias que comparten características similares. La elección del criterio de similitud -ya sea mediante atributos predefinidos o dejando que el modelo opere sobre la estructura bruta de los datos- condicionará el tipo de agrupación resultante. El algoritmo *K-means* es un método popular de agrupamiento en los procesos de aprendizaje no supervisado que organiza las instancias en *K* grupos articulados cada uno en torno a un centroide, asignando cada instancia al grupo más próximo a fin de minimizar la dispersión dentro de cada grupo. La segunda orientación es la detección de asociaciones (*association algorithms*), dirigida a capturar vínculos estadísticos entre variables o entre eventos, cuya expresión más popular son los sistemas de recomendación. La tercera es la reducción de dimensionalidad o *dimensionality reduction*, que transforma los datos en una representación de menor complejidad preservando la información relevante. Esta estrategia es útil para la visualización y compresión de datos, pero sobre todo lo es para el preprocesamiento de datos previo a otras fases de entrenamiento de un modelo.

En el aprendizaje no supervisado, al no existir la fase de etiquetado de los datos, el reto del entrenamiento se traslada a otras decisiones de diseño: la forma en que se representarán las instancias, los atributos que se tendrán en cuenta -o si se dejará al modelo operar directamente sobre la estructura bruta de los datos- o la configuración de hiperparámetros, como el número de grupos en que se quiere dividir los datos (en tareas de *clustering*) o la tasa de aprendizaje. Una crítica recurrente a este paradigma es precisamente que su evaluación descansa en criterios como la cantidad de conjuntos en que se dividirán los datos o la distancia entre los grupos identificados, lo que puede conferirle un carácter eminentemente subjetivo.

Entre sus aplicaciones más frecuentes figuran la segmentación de clientes según su comportamiento, la agrupación temática de documentos según contenido, el reconocimiento de patrones en secuencias moleculares o de discurso y, en general, cualquier tarea de minería de datos en la que no se disponga de etiquetas previas.

Aprendizaje autosupervisado, *por M. de Asís Pulido*

El aprendizaje autosupervisado o *self-supervised learning* es la estrategia de *Machine Learning* en la que el modelo se entrena a sí mismo para aprender partes del *input* a partir de otras partes del mismo *input*, generando así sus propias pseudo-etiquetas como tarea previa o de pretexto (*pretext task*), a partir de las cuales extrae representaciones útiles sin necesidad de que un agente externo haya etiquetado los datos. En ese sentido, se sitúa en una posición intermedia entre el aprendizaje supervisado y el no supervisado: comparte con el primero la existencia de una señal de supervisión que orienta el ajuste de parámetros, pero dicha señal no es aportada por un "profesor" humano, sino que se obtiene directamente de la estructura interna de los propios datos.

El ejemplo más ilustrativo de esta estrategia es el enmascaramiento en tareas de procesamiento del lenguaje natural. Dado un enunciado como "el apellido del primer presidente en la historia de los Estados Unidos es Washington", tal enunciado queda enmascarado como "el apellido del primer presidente en la historia de los Estados Unidos es [MASK]", de manera que el modelo debe aprender a predecir el elemento textual oculto o enmascarado a partir del contexto disponible, ajustando iterativamente sus parámetros para predecirlo en función de las correlaciones encontradas en los datos (en este caso textuales). El propio enunciado proporciona así la etiqueta sin necesidad de intervención humana. Otro ejemplo clásico de aprendizaje autosupervisado es el de los autocodificadores, que aprenden a comprimir los datos de entrada en una representación de menor dimensión y a reconstruir posteriormente la entrada original a partir de ella, tomando esta última como referencia para minimizar el error de reconstrucción.

Este paradigma ha adquirido una relevancia central en el desarrollo de los modelos de lenguaje de gran tamaño o *Large Language Models*, y, en particular, de la arquitectura *transformer*. El preentrenamiento de estos modelos sobre grandes *corpus* de texto no etiquetado se lleva a cabo mediante estrategias de aprendizaje autosupervisado relativas a la predicción enmascarada, característica de modelos como BERT, o predicción autorregresiva del siguiente *token*, empleada por los grandes modelos generativos contemporáneos. Esta fase de preentrenamiento dota al modelo de representaciones generales del lenguaje que pueden ser especializadas en un momento posterior mediante técnicas de ajuste fino o *fine-*

tuning para tareas concretas, lo que conecta el aprendizaje autosupervisado con el concepto de modelo fundacional.

Aprendizaje por refuerzo, *por M. de Asís Pulido*

El aprendizaje por refuerzo o *Reinforcement Learning* (RL) es un paradigma de *Machine Learning* en el que un sistema inteligente, denominado *agente*, aprende a tomar decisiones mediante su interacción con un entorno. A diferencia del aprendizaje supervisado, donde el modelo dispone de ejemplos previamente etiquetados, o del aprendizaje no supervisado, donde debe descubrir patrones presentes en los datos, en el aprendizaje por refuerzo el agente aprende a través de las consecuencias de sus propias acciones, es decir, mediante estrategias de ensayo y error (*trial and error*). Su objetivo consiste en identificar aquellas pautas de comportamiento que le permitan maximizar una recompensa acumulada a lo largo del tiempo.

El funcionamiento básico de este paradigma se articula en torno a varios elementos. En primer lugar, el *entorno* proporciona al agente información sobre el estado en el que se encuentra (espacio de estado). A partir de dicho estado, el agente selecciona una *acción* entre las disponibles en su espacio de acción. Como consecuencia de esta acción, el entorno transita a un nuevo estado y proporciona una *señal de recompensa* (*feedback*), positiva o negativa, que refleja el grado de adecuación de la decisión adoptada respecto del objetivo perseguido. El proceso se repite de forma iterativa, permitiendo al agente mejorar su comportamiento mediante ensayo y error.

Desde una perspectiva formal, el aprendizaje por refuerzo opera sobre secuencias de estados, acciones y recompensas interdependientes. El problema fundamental consiste en determinar una *política* (*policy*), esto es, una regla de decisión que indique qué acción debe realizar el agente en cada situación para maximizar la recompensa esperada.

Los distintos métodos de aprendizaje por refuerzo difieren en la forma de representar y optimizar este proceso de decisión. Algunos enfoques se basan en funciones de valor, que sirven para determinar las acciones que llevan a mejores estados; otros aprenden directamente una política de actuación; y otros combinan ambas aproximaciones mediante arquitecturas denominadas actor-crítico. Entre los algoritmos de aprendizaje por refuerzo más influyentes se encuentran los métodos de diferencia temporal (*temporal difference learning*) y, en particular, el algoritmo *Q-learning*, que permite estimar el valor esperado de recompensa (Q) asociado a cada par estado-acción mediante actualizaciones iterativas basadas en la recompensa inmediata obtenida y en la estimación del valor futuro esperado.

La combinación de las estrategias de RL con modelos basados en redes neuronales profundas ha dado lugar al denominado aprendizaje por refuerzo profundo (*Deep Reinforcement Learning*), en el que las funciones de valor o las políticas son aproximadas mediante técnicas de *Deep Learning*. Esta combinación ha permitido abordar problemas de elevada complejidad caracterizados por espacios de estados y acciones extensos.

Las aplicaciones del aprendizaje por refuerzo son especialmente relevantes en aquellos contextos en los que no resulta posible definir de forma exhaustiva la solución óptima de antemano. Entre ellas destacan la robótica, los vehículos autónomos, la optimización de procesos industriales, determinados sistemas de recomendación o la gestión de recursos. Además, ha desempeñado un papel relevante en algunos hitos recientes de la IA en el campo de los juegos estratégicos complejos como el ajedrez o el Go, así como de otros videojuegos. Más reciente es la aplicación de técnicas derivadas del aprendizaje por refuerzo en el alineamiento de modelos LLM mediante procedimientos que incorporan evaluaciones humanas al proceso de optimización tras la fase de preentrenamiento, tomando estas técnicas la denominación de *Human Feedback Reinforcement Learning*.

Aprendizaje profundo (*Deep Learning*), *por M. de Asís Pulido*

El *Deep Learning* (DL) o aprendizaje profundo es aquel subconjunto del *Machine Learning* basado en redes neuronales artificiales compuestas por (i) la interconexión de múltiples capas ocultas o *hidden layers* -es decir, capas de neuronas no asociadas directamente con el *input* ni con el *output* de la red-, pudiendo llegar a centenas de ellas, y, generalmente, (ii) arquitecturas específicamente diseñadas para determinados tipos de datos o tareas. Estos modelos constan de múltiples niveles de representación encadenados que, partiendo normalmente de representaciones brutas de los datos, utilizan los resultados de cada nivel para aprender funciones y atributos abstractos. Es esta capacidad de descubrir automáticamente representaciones jerárquicas profundas de los datos lo que distingue al DL de las arquitecturas convencionales de *Machine Learning*.

Desde el punto de vista de su estructura, una red neuronal profunda puede entenderse como una serie de ecuaciones conectadas que, aplicadas secuencialmente sobre un *input*, lo transforman en un *output*. Cada neurona procesa una combinación ponderada de las señales procedentes de la capa anterior y genera una nueva representación de la información. En las capas ocultas, estas representaciones suelen corresponder a atributos, patrones o características abstractas aprendidas automáticamente durante el entrenamiento, sin haber sido definidas explícitamente por el programador. Las neuronas de dichas capas aplican funciones de activación no lineales -lo que permite al modelo capturar

patrones y relaciones complejas entre los datos de entrada- y transmite su resultado ponderado a las neuronas de la capa siguiente. Los pesos que modulan dichas conexiones, junto con el sesgo asociado a cada neurona, constituyen los parámetros del modelo. En los sistemas más avanzados se pueden alcanzar los miles de millones de parámetros.

El ajuste de dichos parámetros durante el entrenamiento se lleva a cabo mediante dos algoritmos de particular importancia: la retropropagación (*backpropagation*), que calcula la contribución diferencial de cada parámetro al error total de la red mediante la obtención de los *gradientes*, y el descenso de gradiente, que determina en qué dirección y en qué medida debe modificarse cada parámetro para reducir el mencionado error global.

El DL ha permitido alcanzar niveles de rendimiento considerables en numerosas tareas complejas, aunque muchas veces a costa de la interpretabilidad: en la mayoría de los casos, no es posible dar una explicación de cómo los valores de los parámetros aprendidos por el modelo reflejan las características reales de los datos, lo que convierte a estos modelos en ejemplos paradigmáticos de cajas negras. Aunque históricamente el aprendizaje supervisado fue la estrategia dominante en el DL, actualmente tienen predominancia los enfoques de aprendizaje autosupervisado y por refuerzo.

Entre las arquitecturas más relevantes dentro del DL cabe mencionar las redes neuronales convolucionales (CNN), las redes neuronales recurrentes (RNN), las redes generativas adversarias (GAN) o los modelos de transformadores o *Transformers*. Sobre estos últimos, introducidos en 2017, se sustentan los *Large Language Models* y buena parte de la IA generativa actual.

Redes neuronales, *por C. Vida Carrión*

Es un modelo de aprendizaje automático inspirado en los patrones de comportamiento del cerebro humano. Para ello se sirve de neuronas artificiales, organizadas en capas de nodos interconectados, que extraen y analizan información y datos secuenciales imitando el comportamiento de las neuronas humanas. Estas redes reconocen patrones y procesan los datos recopilados, generando así conocimiento aprendido mediante repetición. Constituyen una herramienta básica para la creación del *machine learning* y *deep learning*, capaces de realizar predicciones basadas en series temporales y obtener conclusiones secuenciales mediante el entrenamiento de modelos.

Por ello son también denominadas redes neuronales artificiales (ANN) o redes simuladas (SNN), porque actúan por imitación de las neuronas del cerebro

humano, conectándose entre sí y traspasándose información para resolver patrones complejos.

Se estructuran en tres capas interconectadas por asociación: capas de entrada (*input*), capas ocultas (*procesamiento*) y capas de salida (*output*).

Se pueden distinguir Redes Neuronales Convolucionales (*ConvNets* o CNN), destinadas al procesamiento de datos visuales, como imágenes y videos, capaces de identificar y clasificar imágenes y objetos, apoyadas en principios de álgebra lineal; y Redes Neuronales Recurrentes (RNN), basadas en repeticiones secuenciales en el tiempo que permiten hacer predicciones, por ejemplo predicciones meteorológicas o de fenómenos naturales. El procesamiento de datos a través de las Redes Neuronales Recurrentes (RNN) permite incluso la traducción de idiomas -procesamiento del lenguaje natural (PLN)- o el análisis de emociones o el reconocimiento y reproducción de voz.

Redes neuronales convolucionales (*Convolutional Neuronal Network*, CNN)*, por J. A. Nisa Ávila*

CNN es un tipo de arquitectura de redes neuronales que utiliza operaciones de convolución para detectar patrones locales en los datos con el objetivo de extraer representaciones significativas de patrones jerárquicos que posiciona de forma espacial entorno a patrones comunes de reconocimiento. Su característica principal es el uso de operaciones de convolución, mediante las cuales la red aplica filtros matemáticos capaces de detectar automáticamente patrones relevantes dentro de los datos.

Las redes neuronales CNN son la base de los sistemas de visión artificial, biometría, reconocimiento facial y sistemas de vigilancia jurídica en referencia a reconocimiento facial en espacios públicos, control de fronteras o análisis forense de pruebas visuales en procesos judiciales o gubernamentales. Así, las CNN lo que hacen es transformar estructuras visuales en representaciones matemáticas distribuidas dentro de espacios vectoriales y jerarquías neuronales profundas.

Esto implica un cambio epistemológico importante: la identificación de objetos y patrones deja de depender exclusivamente de clasificación humana explícita y pasa a emerger de correlaciones matemáticas aprendidas estadísticamente.

Redes generativas antagónicas (*Generative Adversarial Network*, GAN), *por J. A. Nisa Ávila*

Se trata de una arquitectura dual compuesta por dos redes neuronales contrapuestas que compiten entre sí. Hay una red generadora, que intenta crear contenidos inexistentes o irreales, y una segunda contrapuesta que intenta detectar los contenidos irreales de la primera red. Los dos intervinientes en las GAN son el generador y el discriminador. El objetivo es alcanzar el equilibrio de Nash, que indica cercanía al punto óptimo de diseño neuronal.

Esta arquitectura es fundamental para poder crear contenidos o datos *deepfake*, lo cual plantea riesgos éticos y legales que conllevan la pérdida de fiabilidad de pruebas digitales y genera disrupciones jurídicas para la protección digital, así como su alcance.

Procesamiento del lenguaje natural *(Natural Language Processing*, NLP), *por C. Vida Carrión*

El procesamiento del lenguaje natural (PLN) es el análisis de la interacción de los sistemas de la IA con las reglas del lenguaje humano (lingüística computacional) para su reconocimiento, comprensión, traducción y respuesta automática a través de texto o de comandos de voz como asistentes virtuales, generando el contenido solicitado de la manera más eficiente. Para ello se basa entre otros en redes neuronales, permitiendo la traducción automática de idiomas, el análisis de emociones, sentimientos o el reconocimiento y reproducción de voz.

Permite a la IA comprender y procesar el lenguaje humano tanto en su forma escrita, hablada o emocionalmente expresada, pudiendo interactuar incluso desde un plano por tanto emocional (por ejemplo, con ironía o sarcasmo).

Minería de argumentos (*Argument Mining*), *por C. Vida Carrión*

La minería de argumentos es una disciplina del procesamiento del lenguaje natural cuya finalidad es la extracción de los fundamentos y conclusiones de textos. Para ello se basa en varias estructuras a través de patrones estándar que contienen tres elementos básicos que son detectados por el procesamiento del lenguaje natural:

a) El argumento, referido a los fundamentos del texto que soportan la tesis mantenida y sus conclusiones.

b) La premisa, que justifica la conclusión del texto.

c) Y la relación argumentativa, que es el elemento que une de manera coherente tanto los argumentos como las premisas con las conclusiones.

El procesamiento del lenguaje natural procesa computacionalmente esos tres elementos básicos (Componente Argumentativo), clasificando los argumentos como iguales o como opuestos, esto es, si se apoyan o contradicen entre sí (Clasificación de Relaciones), y secuencialmente clasifica la postura del autor (Minería de Postura) a favor o en contra respecto del tema debatido.

Se usa en múltiples disciplinas para analizar grandes cantidades de datos, como el ámbito jurídico o clínico, donde permite identificar fundamentos, razonamientos o validar hipótesis. También se utiliza en el ámbito social y político, detectando los motivos que impulsan una determinada postura.

Black Box*, por C. Pérez García*

Se conoce como *black box* ("caja negra") el conjunto de procesos u operaciones que permanecen oscuros al entendimiento, que resultan difíciles de desentrañar o que se desarrollan de una manera opaca e inaccesible al conocimiento. El término se toma, en origen, de la denominación popular e impropia de las grabadoras de vuelo (FDR y CVR) instaladas en las aeronaves, conocidas como cajas negras. Se trata de sistemas de alta seguridad, herméticos y seguros frente a cualquier manipulación, que registran con precisión todos los datos del vuelo. Merced a tales atributos, el término ha pasado a designar todo sistema inaccesible, por lo general complejo, cuyas operaciones internas permanecen fuera del alcance de observadores externos.

En informática, en particular, *black box* da nombre a un tipo de prueba de *software*. Se habla de cajas negras cuando no se evalúa el funcionamiento interno de un sistema, sino tan solo su funcionalidad externa, sus interacciones con el entorno o, en resumen, su capacidad aparente para transformar *inputs* en *outputs*. Las pruebas de *black box* no presumen la opacidad del sistema, cuya lógica interna puede ser cognoscible, sino que tan solo prescinden de cualquier examen sobre tal extremo. Empero, está claro que los sistemas que sí tienen una lógica interna inextricable se hallan, en este sentido, limitados. La prueba de *black box* es una posibilidad para todo sistema, pues basta con ignorar su lógica interna, pero para algunos sistemas es la única posibilidad, pues no tienen una lógica interna cognoscible a la que aplicar una prueba de *white box*: los sistemas con *white box*, puesto que puede esta cerrarse, pueden someterse a ambas pruebas; los sistemas de *black box*, puesto que no puede esta abrirse, solo pueden someterse a una de ellas. En este sentido, pues, podemos considerar que la *black box* es un problema en un sistema informático.

Los sistemas de IA más propensos a desarrollar *black boxes* son, dentro de los de *machine learning*, sistemas tales como los de *deep learning*, pues su propio proceso de aprendizaje automático los hace, a partir de cierto punto, incomprensibles en su código. Piénsese que, de no ser incomprensibles, no serían tampoco tan potentes ni capaces de realizar las operaciones para las que se los emplea. Tales sistemas, por tanto, tienen una tensión teleológica hacia la *black box*, que les resulta, por definición, inevitable. En el extremo opuesto, los sistemas de IA simbólica son siempre explicables, pues son producto, en toda la extensión de su código, de programación humana, sin aprendizaje automático. Al precio de ser sistemas menos potentes, su lógica interna permanece clara. Por tanto, los de IA simbólica son sistemas *white box*, a los que, no obstante, se puede siempre someter también a pruebas de *black box*.

White Box*, por C. Pérez García*

Se conoce como *white box* ("caja blanca") el conjunto de procesos accesible al conocimiento, comprensible y explicable. El término se acuña por oposición a *black box* y se emplea, sobre todo, en informática, donde da nombre, en particular, a un tipo de prueba de *software*. Se habla de cajas blancas cuando se evalúa el funcionamiento interno de un sistema, en lugar de atender tan solo a sus interacciones con elementos externos. Puesto que no puede evaluarse la lógica interna de un sistema opaco, la prueba de *white box* implica una transparencia y claridad suficientes en el sistema objeto, lo que reduce su aplicabilidad tan solo a los sistemas explicables. Por ello, también se las conoce como *clear box* ("caja clara"), *glass box* ("caja de cristal") u *open box* ("caja abierta"), frente a las cerradas y opacas cajas negras. Todo sistema con *white box* se puede someter tanto a pruebas de *white box*, si se mantiene la caja abierta, como de *black box*, si se cierra; los sistemas de *black box*, en cambio, no son explicables y, por tanto, no pueden someterse más que a pruebas de *black box*, nunca de *white box*.

Capa*, por J. Nisa Ávila*

Una capa es un conjunto estructurado de neuronas artificiales que operan de forma simultánea y transforman los datos de entrada mediante operaciones matemáticas (ponderación, suma y función de activación). Cada capa recibe información, la procesa según sus parámetros internos (pesos y sesgos) y transmite el resultado a la capa siguiente. Las capas constituyen los niveles funcionales de una red neuronal y permiten organizar el procesamiento de la información de manera progresiva, desde representaciones simples hasta abstracciones complejas.

Las capas son el elemento estructural fundamental de las arquitecturas de aprendizaje profundo. Habitualmente se distinguen:

a) Capa de entrada, que recibe los datos originales.

b) Capas ocultas, donde se realiza la mayor parte del aprendizaje y la transformación de la información.

c) Capa de salida, que produce el resultado final (clasificación, predicción o generación).

La profundidad de una red -es decir, el número de capas ocultas- determina su capacidad para aprender representaciones jerárquicas. Las capas pueden interpretarse como niveles de mediación cognitiva y normativa, análogos a los distintos estratos del razonamiento jurídico (hechos, normas, principios, interpretación).

Desde la perspectiva del Derecho Cuántico, cada capa representa un estado intermedio de sentido, una fase de superposición interpretativa que no es todavía decisión jurídica, pero que condiciona su colapso final.

Nodo, *por J. A. Nisa Ávila*

En el contexto de las redes neuronales artificiales, un nodo (o neurona artificial) es la unidad computacional elemental que puede recibir distintos valores de entrada y, tras la ponderación de esos valores mediante pesos, les aplica una suma y una función de activación, produciendo un valor de salida que se transmite a otros nodos. El nodo constituye el elemento mínimo de procesamiento dentro de una red neuronal y es responsable de realizar transformaciones locales de la información.

Un nodo trata los datos recibidos implementando una transformación matemática de la forma:

$y = sigma\ (w\ cdot\ x + b),$

donde x es el vector de entrada, w el vector de pesos, b el sesgo y sigma la función de activación.

Desde el punto de vista técnico, un conjunto individualizado de nodos agrupado bajo una misma o una serie de objetivos se consideran bloques funcionales que, al interconectarse en capas, permiten a la red neuronal aproximar funciones complejas. Cada nodo por sí solo tiene una capacidad expresiva limitada; sin embargo, la interacción masiva de nodos organizados en arquitecturas profundas genera comportamientos emergentes, como reconocimiento de patrones, inferencia contextual y generación de lenguaje.

El nodo resulta relevante porque representa el nivel microestructural de la decisión algorítmica. Jurídicamente, ningún nodo individual "decide"; la decisión emerge de la agregación de miles o millones de operaciones locales distribuidas. Esto dificulta la atribución causal y la responsabilidad, ya que no existe un punto único de control o intención.

Desde la perspectiva del Derecho Cuántico, el nodo puede entenderse como una unidad de potencialidad interpretativa: cada nodo contribuye de manera probabilística al resultado final, sin que su contribución sea determinante de forma aislada. La decisión algorítmica aparece así como un fenómeno emergente, análogo a un sistema cuántico compuesto por múltiples partículas cuya interacción produce un estado global no reducible a sus partes.

Pesos, *por J. A. Nisa Ávila*

Los pesos son los parámetros ajustables de una red neuronal que determinan la intensidad y dirección de la influencia que una entrada ejerce sobre la salida de un nodo dependiendo del valor que se le otorgue a un peso respecto a otros en una misma operación. Matemáticamente, los pesos forman parte de la transformación lineal *w cdot x*, donde cada peso multiplica un valor de entrada específico.

Desde el punto de vista técnico, los pesos representan el conocimiento aprendido por la red neuronal. No almacenan reglas explícitas, sino patrones estadísticos codificados en forma numérica. En redes profundas modernas los pesos pueden ascender a millones o miles de millones de parámetros, organizados en matrices de alta dimensionalidad.

En arquitecturas como los *transformers*, los pesos no solo aparecen en capas densas tradicionales, sino también en las matrices de proyección de los mecanismos de atención.

La actualización de los pesos se realiza mediante retropropagación (*back-propagation*) del error y optimización iterativa, lo que implica que el aprendizaje es un proceso de ajuste continuo dentro de un espacio paramétrico de gran dimensión.

Por ello, los pesos tienen una relevancia jurídica central, porque constituyen el soporte material del comportamiento del sistema. A diferencia del *software* tradicional, donde las reglas son explícitas y auditables, en las redes neuronales las "reglas" están implícitas en los pesos numéricos. Esto plantea cuestiones regulatorias fundamentales:

- ¿Constituyen un secreto empresarial?

- ¿Son auditables jurídicamente?
- ¿Puede exigirse acceso a los pesos en procedimientos de supervisión o control judicial?

Desde la perspectiva del Derecho Cuántico, los pesos pueden interpretarse como operadores probabilísticos de transformación del sentido: no determinan un resultado único, sino que configuran un campo de posibilidades cuya actualización depende del contexto de entrada.

Vectores de posición, *por J. A. Nisa Ávila*

Los vectores de posición son mecanismos que permiten a una red neuronal conocer la ubicación, orden o relación estructural de los datos que procesa, de forma que permite localizar cada elemento dentro de una estructura de datos. Su función es incorporar información sobre secuencia, jerarquía o distribución espacial, elemento esencial para interpretar correctamente el significado de los datos en modelos de lenguaje, visión artificial y sistemas bioinspirados.

Las redes neuronales no perciben el mundo como lo hace un ser humano. Para una red, los datos son simplemente números, puesto que no tienen capacidad de comprensión semántica como lo hace un ser humano ni de abstracción. Por ello, si no se añade alguna referencia posicional, el modelo puede saber qué elementos existen, pero no cómo están organizados desde un punto de análisis racional.

Desde un punto de vista jurídico, los vectores de posición son especialmente relevantes porque muestran que el significado no depende únicamente de las palabras, sino también de su orden y contexto estructural. Esto resulta esencial, porque pequeñas variaciones en la posición de un término pueden alterar profundamente el sentido normativo de una disposición legal. Según la perspectiva del Derecho Cuántico los vectores de posición muestran que el significado jurídico no emerge solo de elementos aislados, sino de la relación estructural entre ellos dentro de una secuencia normativa analizada como *tokens* individualizados.

Autocodificador (*Autoencoder*), *por J. Nisa Ávila*

Un autocodificador es una red neuronal artificial de aprendizaje no supervisado diseñada para autocodificarse y descodificarse en base a sistemas de representación latente de datos. Su objetivo principal es aprender a comprimir datos de entrada (codificación) y luego reconstruirlos fielmente (descodificación), extrayendo únicamente las características más esenciales de la información original. El sistema codifica (*encode*) los datos de forma comprimida, extrayendo

la información fundamental y desechando la redundante o accesoria. La representación compacta resultante de esa compresión constituye el espacio latente, donde los datos se reducen a sus rasgos fundamentales y se desecha lo no importante. Posteriormente, se descodifica (*decode*) o reconstruye la información a partir de los datos comprimidos en el espacio latente, intentando generar una salida que sea lo más fiel posible a la información original.

Los autocodificadores son aplicables en entornos jurídicos para tareas de análisis forense de datos, compresión de expedientes y detección de patrones ocultos. Desde la perspectiva jurídica, esto plantea una cuestión fundamental, puesto que el sistema puede construir estructuras cognitivas funcionales sin utilizar conceptos jurídicos, semánticos o normativos explícitos. Por ello, la red puede "entender" relaciones operativas entre datos sin que exista una traducción directa a categorías interpretables por el Derecho.

Desde un punto de vista de enfoque asociado al Derecho Cuántico, el conocimiento jurídico ya no aparecería como un conjunto estable de categorías cerradas, sino como un campo dinámico de representaciones probabilísticas susceptibles de reorganización contextual. La decisión automatizada no emergería de una regla lineal claramente identificable, sino de la interacción entre múltiples dimensiones latentes cuya interpretación depende del contexto.

Si lo ponemos en perspectiva a sistemas de Inteligencia Biológica Sintética o SBI, estos sistemas basados en organoides podrían desarrollar representaciones internas funcionales mediante mecanismos de plasticidad y adaptación similares a los observados en redes neuronales artificiales. En este contexto, la "codificación" de la información dejaría de depender exclusivamente de pesos numéricos y espacios vectoriales digitales, pasando a involucrar dinámicas bioeléctricas, sinápticas y moleculares propias de sistemas vivos.

Datos, *por D. Jove Villares*

Conforme al artículo 2.1. del Reglamento de Gobernanza de Datos (Reglamento (UE) 2022/868, en adelante RGD) dato es "toda representación digital de actos, hechos o información, así como su recopilación, incluso como grabación sonora, visual o audiovisual". Esta definición se ha convertido en la conceptualización normativa de referencia en el cada vez más exuberante ecosistema normativo europeo del dato. Sin embargo, es una caracterización incompleta, pensada para un entorno concreto y para un tiempo actual. A diferencia de lo indicado en el RGD, los datos pueden ser una representación no digital, no necesitan ser traducidos a códigos binarios ni que su soporte haya de ser electrónico. Por tanto, dato es sencillamente información. Información que, si se quiere procesar y aprovechar requiere de su plasmación en algún material

que permita su registro, ya sea analógico, biológico (por ejemplo, la mente humana) o digital.

Esa función instrumental, el dato como información destinada a ser usada (aunque sea almacenándola), es uno de los rasgos del concepto aquí descrito; y es que los datos no pueden caracterizarse en toda su magnitud sin analizar sus características específicas y contexto. El soporte en el que se encuentran, el modo en que se procesan, el contenido al que se refieren o quién opera con ellos son los elementos que permiten adjetivarlos y, con ello, determinar el régimen jurídico que se les habrá de aplicar. Todos ellos son aspectos determinantes, pues a una misma información, si se utiliza en un contexto diferente, pueden llegar a imponérsele unas exigencias distintas.

Desde ese punto de vista, además de la ya apuntada diferenciación entre datos digitales y aquellos que no lo son, uno de los rasgos más relevantes de los datos es si estos están referidos a una persona física identificada o identificable o no; esto es, si son personales o no. Esta taxonomía es crucial, cuando el dato sea personal, estará en concurso el derecho fundamental a la protección de datos personales y, consecuentemente, las exigencias normativas serán mayores. A su vez, los datos personales cuentan con subcategorías, como la de los datos especiales, esto es, aquellos referidos a aspectos más sensibles o cuyo tratamiento puede comportar un riesgo más alto, como salud, origen racial, afiliación política... (*vid.* artículos 9 y 10 del Reglamento General de Protección de Datos).

A su vez, la categoría de los datos no personales, creada por exclusión, incluye un conjunto amplísimo de informaciones, que abarca tanto a aquellas que nunca fueron personales (consumos de entidades, presupuestos, transacciones…) como a los datos anonimizados, es decir, a datos que habiendo sido personales han dejado de serlo por haber sido sometidos a algún proceso capaz de quebrar de manera irreversible el vínculo entre esa información y la persona a la que se refería. Los datos no personales no están carentes de regulación, cuentan con su propio reglamento europeo, el Reglamento (UE) 2018/1807. Entre los datos no personales cabe destacar una tipología un tanto peculiar, los datos sintéticos, los cuales tienen la particularidad de ser datos que no representan a la realidad, sino que se crean artificialmente para simularla.

Por otro lado, como los datos están destinados a ser utilizados, muchas veces se pueden producir agrupaciones de datos, conjuntos que pueden contener tanto información personal como no personal, en cuyo caso estaríamos ante conjuntos mixtos de datos.

Aunque su vínculo con las personas (o la ausencia de este) es, desde un punto de vista jurídico, el que comporta las mayores diferencias (no es lo mismo

ser la realidad fáctica sobre la que se aplica un derecho fundamental que no serlo), lo cierto es que hay otras variables que también tendrán incidencia en la norma aplicable, como quién los gestiona (p. ej. datos en manos de las Administraciones Públicas), la naturaleza de la información, aquello para lo que sirven o el sector en el que es utilizada (datos financieros, datos sanitarios…). Un ejemplo sencillo: el código postal puede servir para que nos envíen una carta, para participar en una promoción o puede ser uno de los factores a considerar a la hora de concedernos o denegarnos un préstamo o fijar la cuota de un seguro de vida. En definitiva, dato es información, pero solo cuando esta se pone en contexto el concepto alcanza a comprenderse en toda su extensión.

Datos personales, *por D. Jove Villares*

Los datos personales son el elemento material que activa la aplicación del derecho a la protección de datos. No resulta extraño, por tanto, que el Reglamento General de Protección de Datos (RGPD) proporcione una conceptualización específica de este término. Conforme a lo en él dispuesto (art. 4.1), los datos personales son "toda información sobre una persona física identificada o identificable". Una vez realizada esta caracterización, el RGPD clarifica alguno de esos elementos aportando ciertas claves y ejemplos que contribuyen a un mejor entendimiento de su alcance y límites: "se considerará persona física identificable toda persona cuya identidad pueda determinarse, directa o indirectamente, en particular mediante un identificador, como por ejemplo un nombre, un número de identificación, datos de localización, un identificador en línea o uno o varios elementos propios de la identidad física, fisiológica, genética, psíquica, económica, cultural o social de dicha persona". Sin embargo, las variables relevantes están en la primera parte de la definición; en ellas se centrará esta explicación.

Al señalar que dato personal es "toda información" la normativa está evitando realizar exclusiones a priori. Esto es, prescinde de elementos externos y no impone que ésta se encuentre en algún soporte específico o que aluda a algún aspecto determinado. Serán los factores que rodeen al dato los que permitan adjetivarlo como personal.

El segundo elemento es que ha de ser información "sobre una persona física". Con esta premisa se excluye a las personas jurídicas y se acota el conjunto de titulares del derecho fundamental, reservándolo a los seres humanos, en principio a aquellos que estén vivos. El RGPD no se aplica a las personas fallecidas, pero deja libertad a los Estados miembros para establecer una regulación específica acerca del destino de los datos de quienes ya no están. España ha hecho uso de esta posibilidad en el artículo 3 de la Ley Orgánica 3/2018, de 5 de diciembre. Además de identificar a los titulares del derecho fundamental, esta parte de la definición

introduce uno de los elementos más relevantes del concepto: los datos personales son informaciones referidas a alguien. Es decir, la relación entre la persona y el dato no es de propiedad, sino que se trata de un vínculo, una conexión que puede ser más o menos intensa, lo que permite afirmar que: mis datos personales no son míos, sino que están referidos a mí. A su vez, esta característica hace que una misma información pueda ser dato personal de varias personas, lo que podría dar lugar a potenciales conflictos, en caso de intereses contrapuestos.

En el establecimiento de ese vínculo entre el dato y la persona resultan especialmente útiles las condiciones que el Tribunal de Justicia de la Unión Europea (TJUE) ha aportado (p. ej. en el apdo. 35 del asunto C-434/16, de 20 de diciembre de 2017), al considerar que una información será dato personal de una persona cuando por "contenido, finalidad o efectos" se refiera a ella. La inclusión de las finalidades y los efectos como criterios de conexión propicia una ampliación notable de los sujetos titulares del derecho y, a la vez, obliga a analizar las particularidades de cada tratamiento concreto para comprobar la existencia de ese vínculo.

El último de los elementos de la definición es que la persona sea "identificada o identificable". Por consiguiente, ha de ser posible establecer algún tipo de conexión que lleve a la eventual singularización del sujeto. Al respecto, no es necesario que esa persona esté completamente identificada, basta con que concurra algún elemento que permita particularizarla y, por tanto, afectarle de modo particular. Más problemático es el término "identificable"; tradicionalmente, se venía interpretando en el sentido de posibilidad potencial, de este modo, aunque quien estuviese operando con los datos no pudiera establecer esa conexión, si técnicamente era posible hacerlo, seguía considerándose como dato personal. Esto permitía que a todo tratamiento de datos seudonimizados se les aplicase la normativa de protección de datos. El TJUE (asunto C-413/23 P, de 4 de septiembre de 2025) y el Digital Ómnibus apuestan por un cambio interpretativo que limita el alcance del concepto, al aceptar que, aquellos terceros que reciban información seudonimizada y no dispongan de medios para, "razonablemente", identificar a la persona física, pueden operar con esos datos como si no fuesen personales.

Datos de entrenamiento, validación y prueba*, por D. Jove Villares*

El desarrollo y funcionamiento de los sistemas de IA no puede sustentarse con cualquier tipo de información, han de ser datos de calidad, estructurados y adecuados. Esa relevancia ha convertido una cuestión técnica (¿con qué datos se ha de entrenar al sistema de IA para que responda a los objetivos perseguidos?) en un elemento jurídicamente relevante (¿cómo han de ser los datos con los que se ha alimentado el sistema de IA para que pueda salir al mercado?).

Consecuentemente, el Reglamento de IA de la UE (RIA) incluye, en su artículo 3, apartados 29 a 32 las definiciones de los términos "datos de entrenamiento", "datos de validación", "conjunto de datos de validación" y "datos de prueba" y, en el artículo 10, establece un conjunto de obligaciones tendentes a garantizar la calidad de la información utilizada.

Comenzando con los datos de entrenamiento, son los que se utilizan "para entrenar un sistema de IA mediante el ajuste de sus parámetros". Dicho de manera simplificada, son los datos destinados a convertir al sistema en inteligente, pues es la información que le va a permitir aprender. A partir de ella comenzará a establecer correlaciones, generar patrones estadísticos o estructuras de decisión. Son el sustrato sobre el que la IA crecerá y se desarrollará, condicionando lo que puede llegar a ser.

Como puede deducirse, la precisión y pertinencia de los conjuntos de datos con los que se entrene la IA es crucial, han de estar orientados a preparar a la IA para hacer aquello que de ella se espera. De ahí que deban ser datos curados, debidamente etiquetados y estructurados. La adecuación de los datos de entrenamiento es sumamente relevante, una buena selección puede permitir evitar sesgos y minimizar errores y, al revés, un mal entrenamiento puede acentuar estos elementos, por ejemplo, por proporcionar conjuntos de datos con carencias relevantes o que ofrezcan una imagen tergiversada o poco ajustada a la realidad. De ahí que el artículo 10 del RIA exija que los datos de entrenamiento sean "suficientemente representativos y, en la mayor medida posible, carecerán de errores y estarán completos en vista de su finalidad prevista. Asimismo, tendrán las propiedades estadísticas adecuadas" y, además, adaptados a la finalidad y "las características o elementos particulares del entorno geográfico, contextual, conductual o funcional específico" en que vaya a utilizarse el sistema de IA. Estas obligaciones normativas son una de las principales medidas de prevención de la discriminación algorítmica. Por lo tanto, no resulta extraño que su cumplimiento también se exija de los datos de validación y prueba.

Los datos de validación son aquellos "datos usados para proporcionar una evaluación del sistema de IA entrenado y adaptar sus parámetros no entrenables y su proceso de aprendizaje para, entre otras cosas, evitar el subajuste o el sobreajuste". Adicionalmente, el RIA añade los conjuntos de validación, que pueden ser "un conjunto de datos independiente o una parte del conjunto de datos de entrenamiento, obtenida mediante una división fija o variable". Si los datos de entrenamiento servían para que el sistema aprendiese, los datos de validación sirven para controlar que aprenda de manera adecuada, esto es, que está siendo capaz de hacer inferencias e ir más allá de reproducir los datos de entrenamiento. Este es un paso crucial, pues esta calibración permite determinar si han faltado datos de entrenamiento (subajuste) o si ha habido un exceso (sobreajuste), lo

que ocurriría cuando convierte los ejemplos en regla y deja de proporcionar conclusiones nuevas ante realidades diferentes.

Los datos de prueba, a diferencia de los de entrenamiento y validación, no forman parte del proceso de desarrollo, sino que sirven para verificar su funcionamiento antes de la puesta en servicio. Su cometido es "proporcionar una evaluación independiente del sistema de IA, con el fin de confirmar el funcionamiento previsto de dicho sistema antes de su introducción en el mercado o su puesta en servicio". Son la piedra de toque, la auditoría externa, que permitirá asegurar que el sistema responde a los parámetros previstos, para lo que es crucial que sean independientes de los datos de entrenamiento y de validación. En definitiva, estas tres categorías no pueden considerarse de manera aislada, todas ellas cumplen funciones complementarias, solo si esta tríada de datos opera de manera adecuada se podrán minorar los sesgos y los riesgos de los sistemas de IA.

Datos de entrada (*inputs*) y de salida (*outputs*), *por D. Jove Villares*

Los *inputs* y los *outputs* son los datos de los que se alimentan (entrada) y los que generan (salida) los sistemas de IA cuando ya han sido puestos al servicio de la sociedad. Específicamente, los datos de entrada (*inputs*) son las informaciones que los sistemas utilizan para realizar su función y los datos de salida (*outputs*) son los resultados, el producto del sistema de IA. En esta línea, el Reglamento de IA (RIA) define (art. 3.33) a los datos de entrada como "los datos proporcionados a un sistema de IA u obtenidos directamente por él a partir de los cuales produce un resultado de salida". Como puede deducirse, existe una conexión funcional entre *input* y *output*, siendo el segundo dependiente del primero. A su vez, la conceptualización que el RIA hace de los *inputs* busca ser lo más omnicomprensiva posible, pues no se circunscribe a los datos que los usuarios puedan aportar (*prompts*, carga de datos o conjuntos de datos estadísticos, etc.), sino que abarca toda información que el sistema vaya a precisar para producir un resultado. Por lo tanto, se incluyen, también, los datos consultados y descartados (porque formaron parte del proceso inferencial), así como los "bucles de retroalimentación" (considerando 67 RIA), esto es, aquellos *outputs* de una operación que se incorporan como *inputs* en una posterior.

Pese a que la definición de dato de entrada establecida en el RIA no realice ninguna distinción entre tipologías de datos, estas no desaparecen. Así, cuando en un *input* se incluyan datos de carácter personal habrá de cumplirse con las exigencias de la normativa de protección de datos (salvo los casos susceptibles de ser calificados como excepción doméstica), por consiguiente, será necesario contar con una base de licitud y no utilizar más datos personales de los estrictamente necesarios para la consecución de los objetivos. Quiere con ello decirse que, por el

hecho de introducirse o gestionarse por parte de un sistema de IA, la información sigue estando sujeta al cumplimiento de las normativas europeas (y nacionales) de tratamiento y gobernanza de los datos. En definitiva, en la medida en que los datos de entrada representan el conjunto, indeterminado y variable, de informaciones que sirven para realizar inferencias y producir resultados, es muy relevante que esos datos cumplan con las exigencias de calidad y ausencia de sesgos que se exigen para los datos de entrenamiento, validación y prueba.

A diferencia de los *inputs*, el RIA no incluye una definición expresa de los *outputs*. Sin embargo, la definición de sistema de IA (art. 3.1) incluye, a título ejemplificativo, distintos tipos de resultados que se pueden generar, como pueden ser "predicciones, contenidos, recomendaciones o decisiones" y, además, aporta el rasgo más relevante de los *outputs*, su capacidad para "influir en entornos físicos o virtuales". El *output*, por tanto, es la externalidad del proceso inferencial y, por tanto, es el resultado que producirá efectos, pudiendo afectar o incidir en los derechos, intereses y posiciones jurídicas de personas físicas o jurídicas.

De este modo, mientras los datos de entrenamiento, validación y prueba son fundamentales en la prevención de daños e inciden en el diseño del sistema y en la posibilidad de que llegue a introducirse en el mercado, los *inputs* (especialmente los introducidos por los usuarios) y, sobre todo, los *outputs* son la manifestación externa que permite activar los mecanismos reactivos de defensa del ordenamiento. Dicho de otro modo, los datos de entrada y los datos de salida son los que van a permitir identificar responsabilidades, pero también detectar posibles fallos de los sistemas (como sesgos y errores) que hubieran pasado inadvertidos en la fase de diseño.

Datos sintéticos (*Synthetic data*)*, por M. A. Adame Martínez*

Los datos sintéticos son conjuntos de datos generados artificialmente mediante algoritmos -frecuentemente mediante modelos de IA generativa- que imitan las propiedades estadísticas y estructurales de datos reales sin contener información directamente vinculada a personas físicas u otras entidades concretas. A diferencia de los datos anonimizados -que derivan de datos reales de los que se ha suprimido o alterado la información identificativa-, los datos sintéticos se generan *ex novo*, sin reproducir directamente registros individuales existentes ni permitir identificar a las personas a las que corresponden. Ello no implica ausencia de conexión con datos reales: en la mayoría de los casos el modelo generador ha sido entrenado sobre conjuntos de datos reales, lo que introduce la posibilidad de que los datos sintéticos incorporen, de forma latente, patrones estadísticos derivados de información personal.

Su utilidad principal reside en el entrenamiento y validación de modelos de IA en contextos donde los datos reales son escasos, sensibles o de difícil acceso. El sector sanitario, el financiero y el judicial son paradigmáticos: resulta jurídicamente complejo, costoso o éticamente cuestionable entrenar un sistema de análisis clínico o de predicción de reincidencia delictiva con datos reales de pacientes o condenados, sin articular complejos marcos de consentimiento y protección. Los datos sintéticos permiten sortear esas dificultades manteniendo un nivel suficiente de representatividad estadística.

Desde la perspectiva del Reglamento General de Protección de Datos (RGPD), la cuestión central es si los datos sintéticos constituyen datos personales. En principio, si el proceso de generación es suficientemente robusto, no lo son en el sentido del art. 4.1 del Reglamento, pues no permiten identificar -ni siquiera indirectamente- a ninguna persona física. Sin embargo, si el modelo generador fue entrenado con datos personales insuficientemente protegidos, existe el riesgo de que los datos sintéticos incorporen, de forma latente, información que permita una reidentificación parcial. El Comité Europeo de Protección de Datos ha advertido sobre la necesidad de aplicar en tal caso garantías adicionales, señalando que la denominación formal de un conjunto de datos como "sintético" no exime de un análisis de riesgo de reidentificación cuando el modelo generador fue entrenado sobre datos personales (véase, en particular, el marco desarrollado por el CEPD en sus directrices sobre técnicas de seudonimización y anonimización). Esta cautela es extensible, por analogía, a los datos sintéticos.

Desde la perspectiva de la responsabilidad civil, la utilización de datos sintéticos de baja calidad representativa puede inducir sesgos en el modelo, con los consiguientes riesgos de decisiones discriminatorias o erróneas. La trazabilidad del proceso de generación y la documentación de las características del conjunto de datos son elementos de *due diligence* exigibles al proveedor. A este respecto, el Reglamento de IA (UE) 2024/1689 impone, para los sistemas de alto riesgo, que los datos de entrenamiento sean "pertinentes, suficientemente representativos y, en la medida de lo posible, libres de errores" (art. 10.3), estándar plenamente aplicable a los datos sintéticos y que ancla jurídicamente la responsabilidad del proveedor por su calidad.

Desde la perspectiva probatoria, surge la pregunta de si un dictamen pericial elaborado con herramientas entrenadas exclusivamente sobre datos sintéticos goza de la misma fiabilidad que el basado en datos reales. La respuesta dependerá de la calidad del proceso de síntesis y deberá ser objeto de valoración judicial caso por caso.

El Reglamento de IA (UE) 2024/1689 no regula directamente los datos sintéticos, pero las obligaciones de documentación técnica y de calidad de los

datos impuestas a los sistemas de alto riesgo (art. 10) son plenamente aplicables a su proceso de generación y uso.

Metadato, *por J. A. Nisa Ávila*

Un metadato es un dato descriptivo sobre otros datos que proporciona información contextual relativa a su origen, estructura, significado, uso, calidad, temporalidad o condiciones de acceso. En sistemas informáticos y de IA, los metadatos permiten organizar, interpretar, rastrear y gobernar los datos primarios, actuando como una capa semántica y operativa adicional que no altera el contenido del dato, pero sí su comprensión y tratamiento.

Desde el punto de vista técnico, los metadatos son esenciales para el ciclo de vida del dato en sistemas de IA: catalogación, entrenamiento, validación, auditoría y reutilización. En aprendizaje automático, incluyen información como procedencia del *dataset*, métodos de recolección, etiquetas, versiones, sesgos conocidos, licencias y condiciones de uso. En infraestructuras distribuidas los metadatos habilitan la trazabilidad y reproducibilidad de los resultados.

El metadato es un elemento clave para marcar la diferencia entre seguridad jurídica, regulación y elementos técnicos a legislar. El metadato adquiere una relevancia jurídica central porque es el soporte técnico de la transparencia y la rendición de cuentas, debido a que la capacidad de trazabilidad no recae sobre el dato bruto ni sobre el modelo, sino sobre sus metadatos: quién generó el dato, con qué finalidad, bajo qué condiciones y con qué transformaciones.

Web scraping, *por P. Viollier Bonvin*

El *web scraping* o "raspado web" es un procedimiento de extracción automática de datos de plataformas, sistemas informáticos o sitios web mediante programas o herramientas automatizadas. El proceso suele incluir el envío de solicitudes a sitios web, la descarga del contenido en formato HTML, el análisis de los datos recopilados y la extracción de información relevante, como texto, imágenes y/o tablas.

El proceso de *web scraping* consiste en tres etapas interrelacionadas: i) análisis del sitio web, ii) raspado del sitio web o copia de la información del sitio, y iii) organización de los datos. El objetivo del *web scraping* como técnica es convertir datos web no estructurados en datos estructurados que se pueden almacenar y analizar en una base de datos o una hoja de cálculo central.

Si bien es posible realizar el proceso del *web scraping* de forma manual, es decir, que una persona simplemente copie y pegue manualmente el contenido de un sitio web usando su navegador personal, para alcanzar los niveles de eficiencia y escalabilidad necesarios para recolectar grandes volúmenes de datos el proceso necesariamente debe ser automatizado. Para ello se utilizan *scripts* o un set de instrucciones que realizan una tarea automatizada, también denominado *bot*. Para alcanzar este objetivo generalmente se utilizan dos *bots* distintos: i) un *bot* "araña", también conocido como rastreador o *web-crawler*, que se utiliza para rastrear sitios web automáticamente y agregarlos a un directorio o índice, y ii) un *bot* raspador o *web-scraper*, cuya misión es extraer datos de sitios web, enviando solicitudes a dichos sitios y luego analizando el HTML u otros datos que se devuelven en la respuesta.

Desde un punto de vista técnico, el *web scraping* se realiza a través de una técnica denominada *HTML parsing*, que involucra la lectura del marcado HTML sin procesar de una página web y la descomposición de sus elementos componentes para que la información pueda ser procesada, analizada o transformada por un programa informático. Para ello, un algoritmo especializado o "analizador" (*parser*) lee el código HTML, identifica el inicio y el final de las etiquetas y construye una representación estructurada de la página web.

El *web scraping* no es una tecnología novedosa, sino que ha sido parte integral de internet desde sus primeros años. En 1993, Matthew Gray creó el primer concepto de un *web crawler*, al diseñar el programa denominado "*The Wanderer*", el cual tenía como propósito medir el tamaño de la web, y que luego fue utilizado para generar un índice de sitios web denominado "*Wandex*". En la actualidad todos los motores de búsqueda realizan rastreo web o *web-crawling* de internet para indexar los distintos sitios web y generar un índice de búsqueda a través de sus herramientas automatizadas. Otras iniciativas como el *Internet Archive*, fundado en 1996, utilizan *web scraping* con el fin de respaldar todos los sitios de internet posibles para preservar el contenido digital y la historia de la web.

El *web scraping* ha recobrado notoriedad en el debate público a propósito de la emergencia de los sistemas de IA modernos, puesto que sus modelos avanzados requieren cantidades masivas de datos para su entrenamiento, que muchas veces son obtenidos de fuentes públicas a través del *web scraping* de internet. Por regla general, cuanto mayor es la cantidad de datos que alimentan el entrenamiento del sistema de IA, mejor es el resultado del modelo, más precisas sus predicciones y sus resultados (*outputs*) de mejor calidad. El *web scraping* permite a los desarrolladores de modelos de IA recolectar grandes cantidades de datos de forma relativamente rápida y económica, sin contar con el consentimiento del titular de dicha información, sin previo aviso o una oportunidad de oponerse a dicho tratamiento, o directamente excluirse del proceso.

Esta recolección masiva de datos obtenidos de fuentes públicas ha entrado en tensión con la regulación de propiedad intelectual y con la normativa de protección de datos personales en distintos países. En particular, ha resultado dificultoso para los desarrolladores demostrar que cuentan con una base de licitud para el tratamiento de datos personales recolectados a través de *web scraping*, así como acreditar el cumplimiento de los principios de la regulación de datos personales, especialmente el principio de minimización de datos y de limitación de la finalidad.

Servidor, *por J. A. Nisa Ávila*

Un servidor es un sistema informático cuya infraestructura está basada en *hardware* y *software* cuya finalidad es proporcionar recursos, datos o servicios a otros sistemas denominados clientes dentro de una red. En arquitectura computacional, el servidor ejecuta aplicaciones o procesos que responden a solicitudes externas mediante protocolos de comunicación estandarizados (por ejemplo, HTTP, TCP/IP). En el contexto de la IA, el servidor constituye la infraestructura que alberga, ejecuta y distribuye modelos algorítmicos, gestionando tanto el acceso a datos como la capacidad de cálculo necesaria para el entrenamiento o la inferencia de modelos.

El concepto de servidor resulta relevante porque introduce la dimensión infraestructural del poder computacional. Aunque los algoritmos y modelos son entidades abstractas, su funcionamiento depende de sistemas físicos controlados por actores institucionales o corporativos. Esta dependencia tecnológica plantea cuestiones regulatorias fundamentales relacionadas con soberanía digital, jurisdicción de datos, control de infraestructuras y responsabilidad por decisiones automatizadas. Desde la perspectiva del Derecho Cuántico, el servidor puede interpretarse como el *locus* material donde se actualizan los procesos algorítmicos, el punto en el que las potencialidades computacionales se transforman en decisiones concretas mediante interacción con usuarios y datos.

III. IA generativa y LLMs

Modelo generativo / Sistema de IA generativa, *por J. A. Ortega Ramírez*

Un modelo generativo es un sistema de IA diseñado para crear contenidos originales a partir de los conocimientos adquiridos durante el proceso de entrenamiento. Es diferente de otras aplicaciones de IA que se centran en clasificar, identificar o recuperar información. Los modelos generativos son capaces de

producir nuevas respuestas sin que estén previamente almacenadas, y que además resultan coherentes con los patrones obtenidos de los datos durante el proceso de aprendizaje.

Estos modelos, por tanto, pueden generar una amplia variedad de contenidos, desde textos, imágenes, audio, vídeo, programas informáticos o combinaciones de ellos. La manera en la que crea la respuesta se basa en la identificación de regularidades estadísticas y relaciones complejas que se puedan extraer dentro de grandes volúmenes de información que se utilizan para entrenarlos.

Los modelos generativos se aplican a una variedad importante de ámbitos y campos, como la educación, la investigación, la creación artística, la comunicación o el desarrollo profesional. Sin embargo, la respuesta que ofrece no siempre puede ser válida, por lo que su utilización requiere de una actitud crítica y responsable. Por ello, los resultados que se obtienen con los mismos deberían de interpretarse como una ayuda al razonamiento, y no como una fuente real y con garantías de conocimiento.

Un sistema de IA generativa es una combinación de tecnologías que permite utilizar un modelo generativo. Se trata, por tanto, de un conjunto de componentes que posibilita a los usuarios usar dicho modelo, que es el núcleo que genera los contenidos.

Entre estos elementos se encuentran la interfaz de usuario mediante la que se interactúa con el sistema, las instrucciones que orientan su comportamiento, los mecanismos de seguridad y moderación, las herramientas externas con las que puede conectarse, los sistemas de control de acceso y permisos, los registros de actividad y, en general, todas las reglas técnicas y operativas que regulan su funcionamiento. La integración de estos componentes permite ofrecer una experiencia de uso más segura, eficiente y adaptada a las necesidades de cada contexto.

Esta diferenciación entre modelo y sistema de IA generativa resulta especialmente relevante desde una perspectiva técnica, organizativa y jurídica. La calidad de las respuestas obtenidas, el nivel de seguridad alcanzado o el cumplimiento de las obligaciones normativas no dependen exclusivamente de las capacidades del modelo generativo, sino también de las decisiones adoptadas en el diseño, la implementación y la gobernanza del sistema en su conjunto. En consecuencia, cuando se evalúa una solución de IA generativa es necesario considerar no solo el modelo que la sustenta, sino también la arquitectura completa que determina su comportamiento y su impacto sobre los usuarios y la sociedad.

Modelo fundacional (*Foundation Model*), *por J. A. Ortega Ramírez*

Un modelo fundacional es un tipo de IA entrenada con grandes volúmenes de datos, que suelen tener diferentes orígenes, formatos y contenidos, pero que se utilizan como base para desarrollar nuevas aplicaciones o soluciones a partir de ellos. Su interés se centra en que se diseñan no para una única función, sino para ofrecer una solución o arquitectura general que puede adaptarse a nuevas tareas como podrían ser redacción, traducción, programación, análisis documental, generación de imágenes o asistencia especializada entre otras. Se utiliza el concepto fundacional en tanto en cuanto que actúa más como una base técnica sobre la que se construyen nuevos sistemas orientados a nuevos usos.

Modelo multimodal, *por J. A. Ortega Ramírez*

Un modelo multimodal es aquel que puede hacer o actuar de varias formas o con diferentes capacidades. En estos modelos se puede procesar, relacionar o generar información interactuando en más de una manera, pudiéndose utilizar texto, imagen, audio, vídeo o datos estructurados. Esta capacidad es muy cercana a lo que hace el ser humano y es especialmente valiosa porque en muchos problemas reales se toma en consideración diferente tipo de información en diferentes formatos. Por ejemplo, en los modelos multimodales se puede interpretar una imagen y explicarla en lenguaje natural, o bien analizar un documento con tablas y gráficos y predecir un cierto comportamiento, o combinar información escrita con otra que sea visual. La multimodalidad acerca los modelos al modo en que las personas solemos manejar la información, sobre todo en contextos complejos.

Un **sistema multimodal** es una aplicación o entorno tecnológico que permite trabajar con distintos tipos de información mediante la integración de modelos, interfaces, reglas de uso, herramientas auxiliares y mecanismos de seguridad. El modelo multimodal es el componente técnico que procesa o genera resultados, mientras que el sistema multimodal incluye todo lo necesario para que una persona pueda interactuar con él de forma práctica. Lo que hace un usuario cuando carga una imagen, formula una pregunta oral o analiza un documento compuesto por texto, tablas y gráficos, lo hace utilizando un sistema multimodal.

Modelo de lenguaje de gran tamaño (*Large Language Model*, LLM), *por E. González Pérez*

Un modelo de lenguaje es un tipo específico de modelo fundacional diseñado para procesar y generar lenguaje natural a partir de patrones aprendidos durante su entrenamiento. Su función consiste en recibir una entrada (por ejemplo, una pregunta, una instrucción, un documento o una conversación) y

producir una salida coherente con ese contexto, como una respuesta, un resumen, una traducción, una clasificación o una reformulación. Para ello, el modelo no trabaja con palabras exactamente como lo haría una persona, sino con unidades llamadas *tokens*, que pueden ser palabras completas, partes de palabras, signos de puntuación u otros fragmentos. Estos *tokens* se transforman en representaciones numéricas que permiten calcular relaciones entre elementos del lenguaje.

Los modelos de lenguaje de gran tamaño, conocidos como *Large Language Models* o LLM, son modelos de lenguaje construidos a gran escala, normalmente con muchos parámetros internos y entrenados con grandes volúmenes de datos textuales o multimodales. Los parámetros son valores numéricos ajustados durante el entrenamiento que influyen en cómo el modelo interpreta una entrada y genera una salida. No equivalen a una memoria literal ni a una base de datos, sino a una estructura matemática que recoge regularidades lingüísticas, semánticas y contextuales. Por eso, un LLM puede producir textos fluidos y útiles, pero también puede cometer errores, inventar datos o presentar como segura una respuesta que no está verificada.

La capacidad de los LLM modernos está estrechamente relacionada con la **arquitectura de transformadores**. La arquitectura *Transformer* fue presentada por Vaswani y otros como una red neuronal basada en mecanismos de atención, sin recurrir a la recurrencia ni a las convoluciones como componentes centrales, lo que permitió procesar secuencias de forma más paralelizable y eficiente. En este sentido, puede decirse que el LLM es el modelo lingüístico resultante, mientras que el transformador es una de las arquitecturas que hacen posible su funcionamiento. Aunque sigue siendo la arquitectura dominante en los modelos de lenguaje actuales, también se investigan arquitecturas alternativas o híbridas, como los modelos de espacio de estados, orientadas a mejorar la eficiencia, el tratamiento de secuencias largas y el coste computacional.

El elemento central de esta arquitectura es la atención. La atención permite que el modelo determine qué partes de una entrada son más relevantes para interpretar otras partes o para generar una salida. En lugar de leer el texto de forma rígidamente lineal, el modelo calcula relaciones entre distintos *tokens*, incluso cuando aparecen alejados entre sí. Una forma especialmente importante es la autoatención o *self-attention*, que relaciona diferentes posiciones dentro de una misma secuencia para construir una representación contextualizada de cada elemento. Así, una misma palabra puede adquirir representaciones distintas según el contexto en el que aparece.

Los transformadores también utilizan atención multicabezal o *multi-head attention*. En vez de calcular una sola relación de atención, el modelo calcula varias en paralelo, de modo que distintas "cabezas" pueden captar relaciones gramaticales,

semánticas, posicionales o conceptuales diferentes. Después, esas perspectivas se combinan en una representación común. Esta estructura explica por qué los modelos basados en transformadores pueden manejar textos complejos, mantener coherencia contextual y adaptarse a tareas variadas. No obstante, que un texto sea coherente no significa necesariamente que sea verdadero, actualizado o completo.

No todos los modelos de lenguaje tienen el mismo tamaño. Junto a los LLM existen los **modelos de lenguaje pequeños** o ***Small Language Models*** (SLM). Un SLM también procesa y genera lenguaje natural, pero con una escala menor. En general, esta diferencia de tamaño se traduce en menor consumo de memoria, menor coste computacional, menor latencia y mayor facilidad de despliegue en dispositivos o entornos con recursos limitados. La diferencia entre un LLM y un SLM no debe entenderse como una simple oposición entre "mejor" y "peor": un LLM suele ser más versátil ante tareas abiertas o complejas, mientras que un SLM puede ser más eficiente, rápido y controlable cuando la tarea está bien delimitada.

Además del tamaño, otra dimensión fundamental es el grado de generalidad o especialización. Un **modelo de IA de uso general** está diseñado para realizar una amplia variedad de tareas: redactar, resumir, traducir, clasificar, responder preguntas, generar código o trabajar con diferentes tipos de información. Esta idea se aproxima a la noción de modelo fundacional, entendido como un modelo entrenado con datos amplios, normalmente mediante autosupervisión a escala, que puede adaptarse a muchas tareas posteriores. En el Reglamento (UE) 2024/1689, el modelo de IA de uso general se define por su generalidad significativa y por su capacidad para realizar de forma competente una amplia gama de tareas diferenciadas e integrarse en diversos sistemas o aplicaciones posteriores.

En cambio, un **modelo de IA específico de dominio** está entrenado, ajustado o configurado para rendir mejor en un ámbito determinado. Ese dominio puede ser una disciplina (por ejemplo, el Derecho), una industria, un tipo de documento, una modalidad de datos o una familia de tareas. La especialización puede lograrse mediante entrenamiento con datos propios del dominio, ajuste fino (*fine tuning*), instrucciones específicas o conexión con fuentes externas. Su ventaja es que puede manejar mejor vocabulario técnico, formatos característicos y relaciones conceptuales propias de un campo concreto. Su límite es que puede perder rendimiento fuera de ese ámbito o reproducir errores, sesgos u obsolescencias presentes en sus datos.

Modelo preentrenado, *por J. A. Ortega Ramírez*

Se dice que un modelo está preentrenado cuando ha pasado por una fase inicial de aprendizaje general antes de ser ajustado o utilizado en una aplicación específica, es decir, ya lleva parte del conocimiento intrínseco al mismo,

permitiéndole partir en el aprendizaje no desde cero sino ya de una cierta base de conocimiento. Esta etapa previa o punto de preparación inicial hace que el modelo sea más fácilmente adaptable, ya que lleva incorporado cierto conocimiento, reduce por tanto la necesidad de incorporar nuevos datos específicos y por ello suelen servir de un punto de partida más interesante y eficiente para el desarrollo de nuevas soluciones, aunque siempre la calidad final del sistema va a depender del ajuste que se haga, del contexto en el que se utilice y del modelo de evaluación que se aplique para su validación final.

El **preentrenamiento** es una de las etapas del aprendizaje y en la misma un modelo aprende o busca patrones u orientaciones a partir de grandes cantidades de datos. En el caso de los modelos de lenguaje, a partir de grandes volúmenes de texto, en esta fase se identificarían las relaciones entre palabras, estructuras gramaticales, estilos de escritura, patrones argumentativos y asociaciones conceptuales. En el caso de que se aplicase a imágenes o dibujos, sería la fase inicial en la que el modelo aprende patrones visuales generales a partir de grandes colecciones de imágenes: en el preentrenamiento se identificarían formas, colores, texturas, objetos, perspectivas y relaciones espaciales sin estar todavía orientado a una tarea concreta. Gracias a este aprendizaje previo, el modelo puede posteriormente especializarse en actividades específicas, como reconocer textos y quién es el autor, o bien reconocer elementos en un dibujo, clasificar imágenes, generar ilustraciones o interpretar diagramas técnicos. Por tanto, se trata de la etapa en la que se dota al sistema de una base amplia de conocimiento, que después podrá especializarse en las siguientes fases del aprendizaje.

Ajuste fino (*Fine tuning*), *por E. González Pérez*

El ajuste fino, o *fine tuning*, permite comprender cómo un modelo de lenguaje ya entrenado puede orientarse hacia una tarea, estilo o dominio concreto sin construirse desde cero. Primero existe un modelo base, entrenado con grandes volúmenes de datos y capaz de tareas generales; después, ese modelo se somete a entrenamiento adicional con datos más específicos para modificar o dirigir su comportamiento. El ajuste fino forma parte de las técnicas de adaptación de modelos. En los modelos fundacionales, la adaptación convierte una base general en un sistema más adecuado para usos posteriores concretos. En ese marco, el ajuste fino es una forma de especialización: no crea desde cero las capacidades generales del modelo, sino que aprovecha lo que ya ha aprendido y lo reorienta hacia un objetivo más definido.

Durante el ajuste fino, el modelo se expone a ejemplos seleccionados que muestran el tipo de entrada que recibirá y el tipo de salida que debería producir. Esos ejemplos pueden servir para aprender una clasificación, seguir un formato,

responder con un estilo determinado o mejorar el rendimiento en una tarea delimitada. IBM lo describe como volver a entrenar un modelo preentrenado con un conjunto de datos más pequeño y enfocado para ajustar sus parámetros y representaciones internas a esos datos.

La diferencia entre entrenamiento inicial y ajuste fino es esencial. En el entrenamiento inicial, el modelo aprende patrones generales: relaciones semánticas, estructuras sintácticas, asociaciones entre conceptos y formas habituales de generar respuestas. En el ajuste fino no se pretende enseñar todo de nuevo, sino aprovechar esa base y dirigirla hacia un comportamiento más concreto. Por eso puede entenderse como una segunda fase de aprendizaje: parte de una competencia amplia y la transforma en una competencia más especializada.

En muchos casos, el ajuste fino utiliza datos etiquetados: cada ejemplo incluye una entrada y una salida esperada. No obstante, la práctica actual no se limita al ajuste supervisado clásico: también incluye ajuste de instrucciones, métodos basados en preferencias y técnicas de alineación que buscan respuestas más útiles, estables y ajustadas a determinados criterios.

El ajuste fino puede hacerse de varias maneras. El ajuste completo actualiza todos o muchos de los parámetros del modelo, lo que puede lograr una adaptación intensa, pero exige más memoria y cómputo. Frente a ello, existen técnicas de ajuste eficiente en parámetros, conocidas como PEFT (*parameter-efficient fine-tuning*), que modifican solo una parte reducida del modelo o añaden componentes entrenables ligeros. Estos métodos permiten adaptar modelos grandes sin actualizar todos sus parámetros.

Este concepto conecta con la distinción entre LLM, SLM, modelos generales y modelos específicos de dominio. Un LLM general puede tener mucha capacidad, pero no estar optimizado para una tarea concreta. Un SLM puede ser menos amplio, pero más eficiente si se ajusta para una función bien delimitada. El ajuste fino actúa como puente entre generalidad y especialización: permite tomar un modelo amplio y acercarlo al comportamiento que se necesita en un contexto determinado.

El ajuste fino no debe confundirse con una instrucción escrita en una consulta ni tampoco con la técnica de la "generación aumentada por recuperación" (*Retrieval-Augmented Generation* o RAG). Una instrucción condiciona una interacción concreta, mientras que el ajuste fino modifica el comportamiento aprendido para futuras situaciones similares. En cuanto al RAG, esta técnica conecta el modelo con fuentes externas en el momento de la consulta, de modo que únicamente cambia el contexto disponible para responder; el ajuste fino, sin embargo, cambia el modelo.

El ajuste fino puede mejorar precisión, estilo, consistencia y adecuación a una tarea, pero no es una mejora automática. Si los datos contienen errores, sesgos, ruido o ejemplos poco representativos, el modelo puede aprender esos defectos. También puede producirse sobreajuste: funciona bien con ejemplos parecidos a los de entrenamiento, pero peor ante casos nuevos. Por eso requiere datos de calidad, evaluación, comparación con el modelo base y control de efectos no deseados.

En síntesis, el ajuste fino permite adaptar un modelo ya entrenado mediante entrenamiento adicional. Si la arquitectura explica cómo funciona el modelo, el tamaño explica parte de su capacidad y la especialización explica su alcance, el ajuste fino explica cómo se pasa de una base general a un comportamiento más ajustado.

RAG (*Retrieval-Augmented Generation*)*, por E. González Pérez*

Si el ajuste fino modifica el comportamiento aprendido de un modelo mediante entrenamiento adicional, RAG permite entender otra forma de adaptación: no cambiar necesariamente el modelo por dentro, sino conectarlo con información externa al realizar la consulta. RAG significa *Retrieval-Augmented Generation*, generación aumentada por recuperación, y designa una arquitectura que combina dos capacidades: recuperar información relevante de una fuente externa y utilizarla como contexto para generar una respuesta.

La idea central es que un modelo generativo no tiene por qué responder únicamente con lo que quedó incorporado en sus parámetros durante el entrenamiento. Los modelos de lenguaje aprenden patrones a partir de grandes conjuntos de datos, pero esos datos son finitos, pueden estar desactualizados y no contienen necesariamente la información específica que se necesita en una consulta concreta. RAG añade una capa de recuperación para que el sistema pueda consultar una base de conocimiento externa antes de generar la salida.

El funcionamiento general puede explicarse en varias fases. Primero, el usuario introduce una consulta o instrucción. Después, un componente de recuperación busca en una base de conocimiento los fragmentos más relevantes. Esa base puede contener documentos, páginas web, manuales, registros, archivos estructurados o textos no estructurados. A continuación, los fragmentos recuperados se incorporan al contexto que recibirá el modelo generativo. Finalmente, el modelo produce una respuesta teniendo en cuenta la consulta y la información recuperada.

Desde el punto de vista técnico, un sistema RAG suele incluir cuatro componentes: una base de conocimiento, que actúa como repositorio externo; un

recuperador o *retriever*, encargado de buscar fragmentos relevantes; una capa de integración, que construye el contexto aumentado; y un generador, normalmente un modelo de lenguaje, que redacta la salida final.

Para que la recuperación funcione bien, los documentos suelen dividirse en fragmentos más pequeños, llamados *chunks*. Después, esos fragmentos se convierten en representaciones numéricas o *embeddings*, que permiten comparar la consulta con el contenido almacenado. En lugar de buscar solo palabras exactas, muchos sistemas RAG utilizan búsqueda semántica y localizan fragmentos cercanos por significado.

RAG se apoya en la distinción entre memoria paramétrica y memoria no paramétrica. La memoria paramétrica es la información que el modelo ha incorporado en sus propios parámetros durante el entrenamiento. La memoria no paramétrica es la información externa consultable, como una base documental o un índice vectorial. Lewis y otros describen RAG como una combinación de ambas. Así, el modelo no depende solo de lo aprendido en el entrenamiento, sino también de información recuperada en el momento de uso.

La diferencia con el ajuste fino es esencial. El ajuste fino vuelve a entrenar el modelo con datos específicos y modifica sus parámetros. RAG mantiene separado el modelo generativo de la fuente externa: no necesita incorporar todos los datos al modelo, sino recuperarlos cuando resultan relevantes. Esto es útil cuando la información cambia con frecuencia, cuando interesa mantener una base documental actualizable o cuando se quiere que la respuesta pueda apoyarse en fuentes localizables.

En desarrollos recientes, RAG no se limita a una búsqueda simple en una base vectorial. También puede incorporar búsqueda híbrida, reordenación de resultados o *reranking*, recuperación multimodal, grafos de conocimiento, recuperación adaptativa o agentes que planifican varias búsquedas antes de responder. Estas variantes mejoran la calidad de los fragmentos recuperados y la trazabilidad.

RAG depende de la calidad de la base de conocimiento, de cómo se dividen los documentos, de la precisión del recuperador, de la actualidad de los datos y de la capacidad del generador para usar la información recuperada. Si el sistema recupera fragmentos irrelevantes o incompletos, la respuesta puede ser errónea aunque parezca fundamentada.

En síntesis, RAG amplía un modelo de lenguaje conectándolo con fuentes externas. Si el ajuste fino enseña al modelo a comportarse de una manera más especializada modificando su entrenamiento, RAG le permite consultar

información externa antes de responder. Dentro del mismo puzzle, el ajuste fino representa la adaptación interna del modelo, mientras que RAG representa la adaptación contextual mediante recuperación.

Alineamiento posterior al entrenamiento (*Post-training Alignment*), *por J. A. Ortega Ramírez*

El alineamiento posterior al entrenamiento es el conjunto de técnicas y procedimientos que se aplican una vez finalizada la fase de aprendizaje de un modelo de IA. Su finalidad es adaptar el comportamiento del sistema a las expectativas de la persona que está utilizando esa IA, incluyendo también los criterios éticos, normativos y de seguridad que garanticen un uso responsable de la tecnología.

La necesidad de esta etapa radica en que un modelo puede haber adquirido una notable capacidad para generar textos coherentes y lingüísticamente convincentes, sin que ello implique necesariamente que sus respuestas sean apropiadas, fiables o beneficiosas en todos los contextos. De hecho, un sistema técnicamente competente puede producir información errónea, sesgada o potencialmente perjudicial si no se incorporan estos mecanismos adicionales de supervisión y control de la respuesta dada. El objetivo último es conseguir sistemas capaces de ofrecer respuestas más precisas, prudentes y contextualizadas, alineadas con los valores y necesidades de las personas que interactúan con ellos.

Uso de herramientas (*Tool Use*), *por J. A. Ortega Ramírez*

El uso de herramientas o *tool use* es la capacidad de un sistema de IA de utilizar recursos externos, como pueden ser buscadores, calculadoras, bases de datos, calendarios, editores de documentos u otros programas de análisis. El objetivo que está detrás de estas capacidades extendidas es que el sistema pueda complementar y ampliar sus capacidades de generación de contenido, ya que le permite ampliar las fuentes de información que utiliza, o bien aplicaciones especializadas que puedan generar nuevos resultados e incluso entornos de trabajo específicamente diseñados para realizar ciertas tareas.

Existe mucha variedad de herramientas que pueden integrarse en estos sistemas, y son a veces dependientes del fabricante del sistema. Entre ellas se pueden encontrar motores de búsqueda, bases de datos, calculadoras, calendarios, gestores documentales, programas de análisis de datos o los servicios de comunicación, entre una amplia variedad. De este modo se enriquece el conocimiento del entrenamiento por la IA con otras habilidades, tales como consultar nueva información que se haya podido generar desde que se creó el

modelo, realizar cálculos complejos, analizar y procesar documentos, ejecutar operaciones específicas de otras herramientas o interactuar con otros sistemas externos durante el proceso de generación de conocimiento. Esto ha supuesto un importante avance en la IA generativa, ya que ha permitido evolucionar los modelos conversacionales originales en sistemas multipropósito con amplias capacidades de apoyar actividades profesionales, administrativas, científicas o empresariales. La combinación entre generación y acceso a herramientas especializadas permite abordar tareas que, de otra manera, no se podrían realizar o, si se hiciesen, sería de una manera menos útil para el usuario.

En cualquier caso, la incorporación de herramientas externas añade nuevos retos en materia de gobernanza, seguridad y responsabilidad de los resultados. Este intercambio de información entre sistemas requiere de permisos y mecanismos que permitan conocer el límite de la autorización que se le ha concedido al sistema de IA, además de la necesaria trazabilidad y supervisión. Es por ello que en el uso de sistemas que tengan esta capacidad de uso de herramientas externas debe quedar muy claro las reglas sobre permisos, límites operativos, control humano y registro de actividades. Se debe siempre garantizar que este incremento de capacidades se acompañe de las necesarias garantías de seguridad, transparencia y determinación de la responsabilidad.

Agente con herramientas, *por J. A. Ortega Ramírez*

Un agente con herramientas es un sistema de IA capaz de generar una respuesta en lenguaje natural, y además planificar una tarea, decidir qué recursos externos necesita y ejecutar una secuencia de pasos para alcanzar un objetivo. Esta característica distintiva permite combinar capacidades de razonar, acceder a información y ejecutar tareas dentro de un mismo proceso de trabajo.

Es una herramienta más rica que un asistente conversacional tradicional, ya que, además de responder preguntas, un agente con herramientas puede analizar y ejecutar una tarea compleja, secuenciándola en etapas y decidiendo qué recursos necesita para completarla. Para ello utiliza normalmente llamada a funciones dentro de su entorno de operación.

En consecuencia, su utilidad resulta especialmente relevante en ámbitos profesionales, empresariales, científicos y administrativos donde las tareas requieren combinar análisis, acceso a información y ejecución de acciones concretas. Sin embargo, este incremento de capacidades operativas implica también un aumento de la complejidad y de los riesgos asociados. Es necesario incorporar mecanismos de control y gobernanza apropiados, a medida que un agente dispone de mayor autonomía para decidir qué acciones ejecutar y qué herramientas utilizar. La definición de los límites operativos de los mismos sin

supervisión humana, la trazabilidad de las decisiones que en cada momento adopte, la posibilidad de realizar una auditoría sobre las actuaciones realizadas y la asignación clara de responsabilidades constituyen cada vez más, elementos esenciales para garantizar un uso seguro y confiable de estos sistemas.

Llamada a funciones (*Function calling*), *por J. A. Ortega Ramírez*

La llamada a funciones o *function calling* es un mecanismo por el que un modelo solicita la ejecución de una operación concreta en otro sistema. A través del mismo se realiza de manera estructurada la ejecución de otras aplicaciones, servicios o recursos externos, y como tal el modelo no se limita a generar texto como respuesta, sino que además puede solicitar la ejecución de acciones específicas mediante instrucciones formalizadas y parámetros previamente definidos.

En lugar de formular una petición en lenguaje natural que debería ser interpretada por otro sistema, el modelo genera una estructura que incorpora datos y que identifica la función que debe ejecutarse en otros sistemas y los parámetros necesarios para realizarla. De este modo, puede realizar operaciones como consultar una base de datos, recuperar información de un documento, crear eventos en un calendario, efectuar búsquedas especializadas o activar herramientas concretas integradas en una aplicación.

La principal ventaja de este enfoque es que, al utilizar parámetros estructurados y reglas establecidas, las aplicaciones pueden interpretar las solicitudes del modelo de forma más precisa, fiable y predecible, sin la ambigüedad que tiene el lenguaje natural.

Esta llamada a funciones es un claro avance en la integración de la IA con los entornos profesionales y organizativos, y constituye uno de los pilares de los sistemas avanzados de IA generativa orientados a la automatización de procesos y coordinación de tareas complejas entre distintos sistemas de información, de manera que el trabajo se define con las garantías propias de los procedimientos informáticos estructurados.

No obstante, y como es habitual cuando interoperan diversos sistemas, es necesario que se incorporen los mecanismos de control, autorización y supervisión. Esta definición de permisos, validación de parámetros, registro de actividades sin intervención humana, requiere garantizar que estas capacidades se utilicen de manera segura, transparente y conforme a las normas que regulan cada ámbito de aplicación, y especialmente cuando se trata de operaciones o se utilizan datos sensibles.

Inferencia, *por J. A. Ortega Ramírez*

La inferencia es la fase en la que se utiliza un modelo ya entrenado. Se produce cuando el sistema recibe una nueva entrada, que puede ser una pregunta, una imagen, un documento o un conjunto de datos, y a partir del modelo se genera una salida, que podrá ser una nueva respuesta, una predicción, una clasificación o una recomendación. Si el entrenamiento puede entenderse como el periodo de aprendizaje, la inferencia es el momento en que ese aprendizaje se aplica. Es precisamente lo que se realiza cada vez que una persona interactúa con una herramienta de IA generativa y obtiene un resultado, esto es precisamente un proceso de inferencia.

Generación condicionada, *por J. A. Ortega Ramírez*

La generación condicionada es el proceso mediante el cual un sistema de IA produce contenidos dependiendo de una serie de indicaciones, contextos o restricciones previamente establecidos. En otras palabras, la capacidad generativa del modelo no opera de manera global ni espontánea, sino que se genera en función de la información que recibe en cada interacción del usuario y conforme a los objetivos con los que se le plantean.

Las condiciones pueden ser muy heterogéneas desde la instrucción formulada por el usuario y contener un conjunto de documentos de referencia, unos datos específicos, un estilo de redacción determinado o una serie de restricciones que deben respetarse durante la generación del contenido. A partir de estos elementos, el modelo interpreta el contexto y construye una respuesta orientada a satisfacer la instrucción planteada. La calidad de los resultados va a depender de la claridad, precisión y relevancia de las condiciones proporcionadas en las indicaciones. Es importante para obtener una respuesta apropiada definir de manera clara el objetivo a conseguir y suministrar de manera estructurada la información de partida. Por esta razón, la generación condicionada constituye uno de los principios fundamentales de la IA generativa.

Multimodalidad generativa, *por J. A. Ortega Ramírez*

La multimodalidad generativa es la capacidad de trabajar con distintos tipos de datos dentro de una tarea, esto es, que el sistema sea capaz de producir, interpretar o transformar contenidos en distintos formatos al realizar una tarea. Estos formatos pueden incluir texto, imágenes, audio, vídeo, gráficos, presentaciones o bien combinaciones de varios de ellos.

Esta capacidad de un sistema puede, por ejemplo, generar una imagen a partir de una descripción escrita, resumir el contenido de un vídeo, explicar un gráfico, crear una presentación o transformar una idea expresada en lenguaje natural en una representación visual. De este modo, la IA deja de operar exclusivamente sobre texto y pasa a interactuar con la información de una manera más cercana a como lo hacemos las personas.

La relevancia de la multimodalidad se centra en que es más cercana a cómo las personas suelen procesar y utilizar su conocimiento de manera cotidiana, combinando de manera natural diferentes tipos de información y como esta se representa.

Su aplicación es importante en ámbitos como la educación, la comunicación, el diseño, la investigación, la administración y la gestión profesional, donde la integración de múltiples formatos puede mejorar tanto la comprensión como la productividad a la hora de hacer las tareas.

***Sampling* o Muestreo**, *por J. A. Ortega Ramírez*

El *sampling*, o muestreo, es el mecanismo mediante el que un modelo de IA selecciona cada nuevo elemento que compone una respuesta durante el proceso de generación. En vez de elegir la alternativa estadísticamente más probable, el sistema puede tomar una entre varias opciones posibles, y que en función de determinados criterios permite introducir en la respuesta final diversidad y flexibilidad, acorde con el hilo de la instrucción dada.

Este mecanismo es lo que hace que un mismo sistema de IA, ante una misma instrucción, genere varias respuestas diferentes que no sean idénticas, aun cuando todas puedan ser coherentes y adecuadas. Gracias al muestreo, los modelos pueden producir textos más naturales, variados y adaptados a diferentes contextos comunicativos, evitando respuestas excesivamente repetitivas o predecibles.

La configuración de este mecanismo juega un papel crucial en la calidad de la generación. Se establece el nivel de muestreo en función de ciertos parámetros, de manera que si está bien ajustado favorece la creatividad, la riqueza expresiva y la fluidez del lenguaje. Por el contrario, cuando se incrementa en exceso puede provocar respuestas menos consistentes, aumentar la probabilidad de errores y facilitar la aparición de afirmaciones inexactas o insuficientemente fundamentadas. Por ello, el muestreo constituye uno de los principales instrumentos para equilibrar la precisión y la creatividad en los sistemas de IA generativa.

Temperatura, *por J. A. Ortega Ramírez*

La temperatura es un parámetro fundamental que regula el grado de variación en la generación de un modelo. Su función consiste en regular el grado de incertidumbre o variabilidad con el que el sistema selecciona las distintas opciones disponibles al construir una respuesta, y por consiguiente influye en el equilibrio entre precisión, creatividad y diversidad expresiva de la respuesta.

Cuando se utilizan valores de temperatura bajos el modelo tiende a escoger las alternativas más probables de acuerdo con los patrones aprendidos durante su entrenamiento, y en tal caso las respuestas suelen ser más estables, previsibles y consistentes, características especialmente valiosas en contextos donde la exactitud y la fiabilidad resultan esenciales. Por el contrario, valores más elevados en la temperatura exploran opciones menos probables, lo que puede enriquecer la creatividad y la originalidad de las respuestas, aunque también incrementa el riesgo de imprecisiones, incoherencias o desviaciones respecto al objetivo inicial.

Es importante destacar que la temperatura no determina el nivel de conocimiento ni la capacidad intelectual del modelo de generación. Su efecto se limita a modificar la forma en que el sistema selecciona entre las alternativas disponibles, condicionando el estilo y la diversidad de la respuesta generada. Por esta razón, en ámbitos como el jurídico, el académico, el científico o el administrativo es recomendable configuraciones de temperatura moderadas o bajas, donde la consistencia, la precisión terminológica y la estabilidad de los resultados prevalecen sobre la creatividad o la innovación expresiva. En otros ámbitos más creativos se utilizan temperaturas elevadas en la generación de la respuesta.

Top-K / Top-P, *por J. A. Ortega Ramírez*

Top-k y top-p son técnicas utilizadas para limitar el conjunto de opciones disponibles durante la generación en el proceso de selección de palabras o elementos durante la construcción de una respuesta. Su finalidad es limitar el número de alternativas entre las que el sistema puede elegir en cada paso de la generación, con la idea de mantener un cierto equilibrio entre coherencia, precisión y diversidad expresiva. Su elección determina la poda del árbol de generación, y con ello la variabilidad de la respuesta. Ambos procedimientos tienen como objetivo evitar que el modelo elija opciones excesivamente improbables que puedan deteriorar la calidad de la respuesta, intentando que la generación sea rígida o muy repetitiva. Se trata, por tanto, de un parámetro que ajusta la respuesta del sistema para obtener resultados más naturales, equilibrados y adecuados a las necesidades de cada aplicación y del usuario.

La estrategia top-k consiste en restringir la selección a un número fijo de opciones que presentan las probabilidades más elevadas según los cálculos del modelo. De este modo, si el valor de k es reducido, la generación tenderá a ser más predecible y controlada, mientras que valores más altos conllevan una mayor variedad en las respuestas. Por ejemplo, si se establece un valor de $k = 50$, el sistema únicamente podrá escoger entre las cincuenta alternativas más probables en cada momento de la generación.

Por otra parte, la técnica top-p, que se conoce como muestreo por núcleo (*nucleus sampling*), toma un enfoque más flexible. En lugar de fijar un número determinado de alternativas, selecciona todas aquellas cuya probabilidad acumulada alcanza un umbral previamente establecido. Esto permite que el conjunto de opciones varíe dinámicamente en función del contexto, conservando las alternativas más relevantes en cada situación. Así, un valor de $p = 0{,}9$ implica que el modelo seleccionará entre las opciones que, conjuntamente, representan el 90 % de la probabilidad total estimada.

Estos parámetros se configuran normalmente durante el diseño, despliegue o ejecución del sistema por parte de desarrolladores, proveedores de servicios o usuarios avanzados. Su elección depende de la finalidad perseguida: aplicaciones orientadas a tareas jurídicas, administrativas o científicas suelen utilizar configuraciones más restrictivas para favorecer la precisión y la estabilidad, mientras que entornos creativos o de generación de contenidos pueden optar por valores más flexibles que incrementen la variedad y la originalidad de las respuestas.

Memoria conversacional, *por J. A. Ortega Ramírez*

La memoria conversacional es la capacidad de un sistema para conservar, recuperar y/o reutilizar información procedente de interacciones previas. Esto se realiza con el objetivo de ofrecer continuidad con la interacción anterior con el usuario, manteniendo las indicaciones e información que este haya aportado anteriormente, así como sus preferencias respecto al contexto. El fin último es que las respuestas sean lo más continuistas y conforme a las indicaciones que el usuario haya realizado con el sistema.

El alcance de esta memoria es dependiente de la herramienta concreta y de las condiciones bajo las que esta trabaja. En algunos sistemas la información solo se mantiene durante una sesión concreta y desaparece al finalizar la interacción con el usuario. En otros casos, esta interacción, o al menos parte de ella, se puede conservar de forma más permanente, y con ello se parte de la premisa de que el sistema ya es conocedor del contexto o de las indicaciones del usuario, y por ende se trata como un conocimiento acumulado a lo largo del tiempo, siempre

por supuesto dentro de los límites técnicos, normativos y de configuración que el sistema tenga establecidos.

Es importante aclarar que la memoria conversacional no implica que el sistema tenga conciencia, como si de un humano se tratase, esto es, que tenga comprensión humana, intención propia o capacidad de recordar en el sentido en que lo hacen las personas. Se trata, exclusivamente, de un mecanismo técnico de gestión de la información, y de guardar parte de la interacción que el usuario haya realizado con la idea de facilitar este diálogo y mejora que se supone basada en la experiencia de uso.

Este tipo de memoria puede conseguir un incremento en la personalización, la eficiencia y la respuesta de los sistemas de IA, evitando repeticiones innecesarias y consiguiendo interacciones más fluidas con el usuario. Sin embargo, su utilización puede también plantear ciertas limitaciones en lo relativo a la privacidad, transparencia, gobernanza de los datos y control por parte del usuario. Por ello, en estos sistemas con memoria conversacional es necesario que quede claro la información que el sistema conserva, durante cuánto tiempo y con qué finalidad se utiliza, además de aportar mecanismos para que en cualquier momento pueda ser conocida y limitada por parte del usuario.

Token*, por J. A. Ortega Ramírez*

Un *token* es una unidad básica de procesamiento utilizada por los modelos de lenguaje, y dependiendo del contexto puede corresponder a una palabra completa, a una parte de una palabra, a un signo de puntuación, a un número o incluso a un espacio, según el sistema de tokenización empleado. Los modelos no leen el texto como lo haría una persona, sino que lo que hacen es convertirlo en estas unidades manejables y, posteriormente, en representaciones numéricas. Comprender este concepto ayuda a explicar los límites de longitud de las consultas, el coste computacional y ciertas diferencias en la forma en que los sistemas interpretan expresiones, nombres propios o términos técnicos. Los *tokens* se han utilizado desde el inicio de los compiladores de los lenguajes de programación. Es un concepto muy utilizado desde siempre en la ingeniería del *software* y, por tanto, en la informática.

La **tokenización** es el proceso mediante el cual un texto se divide en *tokens* para que pueda ser procesado por un modelo. Antes de que una frase, un documento o una instrucción sean analizados por una IA, deben transformarse en una secuencia de unidades representables matemáticamente para que puedan ser procesados por los algoritmos. Aunque pueda parecer un paso irrelevante y solo técnico, tiene unas implicaciones prácticas muy importantes y relevantes, ya que según como se plantee este proceso se condiciona el uso de la información,

limita la longitud de los documentos, el coste que tiene dar una respuesta del sistema y otras implicaciones muy relevantes en la manera en la que el texto se va a utilizar, procesar y dar respuesta.

Ventana de contexto (*Context window*), *por M. A. Adame Martínez*

La ventana de contexto es el límite máximo de texto -medido en unidades denominadas *tokens*- que un modelo de lenguaje puede recibir y procesar en una única interacción. El *token* no equivale exactamente a una palabra: en lenguas romances como el español, una palabra puede generar uno o varios *tokens*, de modo que la capacidad efectiva de procesamiento es algo inferior a la que pudiera sugerir el número de *tokens* anunciado por el fabricante.

La dimensión de la ventana de contexto ha crecido notablemente en los últimos años: si los primeros modelos disponibles comercialmente operaban con ventanas de en torno a 4.000 *tokens* (equivalentes aproximadamente a unas 3.000 palabras), los modelos de última generación alcanzan hoy los 128.000, 200.000 o incluso un millón de *tokens*, lo que permite procesar documentos de centenares de páginas.

Debe distinguirse con precisión entre ventana de contexto y memoria persistente. La ventana de contexto delimita lo que el modelo puede procesar *en una interacción determinada*; fuera de ella, y en ausencia de mecanismos adicionales de persistencia, el modelo no retiene información entre sesiones distintas, salvo que el sistema haya sido expresamente diseñado para ello -mediante mecanismos como el almacenamiento externo de historial o la generación aumentada por recuperación (*Retrieval-Augmented Generation*, RAG)-. Esta distinción es jurídicamente relevante en materia de trazabilidad y auditoría de sistemas de IA.

Para el jurista, la ventana de contexto presenta varias dimensiones de interés. *Primera*, de orden práctico: un modelo solo puede "ver" el texto que cabe en su ventana en un momento dado; lo que queda fuera, deja de ser accesible para el modelo en esa interacción. Un contrato extenso, un expediente administrativo voluminoso o una sentencia con abundante jurisprudencia citada pueden superar el límite, lo que exige estrategias de segmentación con el consiguiente riesgo de pérdida de coherencia o de información relevante.

Segunda, de orden probatorio: cuando un sistema de IA es empleado en el contexto del proceso -para analizar documentos, detectar contradicciones o generar dictámenes periciales-, la ventana de contexto determina qué información ha sido efectivamente considerada por el modelo. Si partes relevantes del expediente quedaron fuera de la ventana, el *output* puede ser parcial o erróneo sin que ello resulte perceptible para el usuario. Esto es especialmente delicado en

sistemas de análisis predictivo de jurisprudencia o en la revisión automatizada de contratos.

Tercera, de orden regulatorio: el Reglamento de IA (UE) 2024/1689 exige que los sistemas de alto riesgo operen con exactitud, solidez y ciberseguridad a lo largo de todo su ciclo de vida (art. 15). La limitación inherente a la ventana de contexto constituye una fuente estructural de riesgo que los proveedores y responsables del despliegue deben documentar, comunicar a los usuarios y mitigar mediante técnicas como la generación aumentada por recuperacíon (*Retrieval-Augmented Generation*, RAG).

***Token* de contexto**, *por J. A. Ortega Ramírez*

Sería más preciso hablar de "*tokens* en contexto" o "*tokens* consumidos por el contexto", aunque cada vez se utiliza más de esta manera que aquí se define. Un *token* de contexto es la unidad básica de información que un modelo procesa y almacena temporalmente dentro de su ventana de contexto. Aunque en el lenguaje cotidiano tendemos a pensar en palabras completas, en los modelos de IA no es así, y se trabaja internamente con unidades más pequeñas que pueden ser palabras enteras, fragmentos de palabras, signos de puntuación o incluso espacios, dependiendo la codificación utilizada.

Cada elemento que participa en una interacción consume *tokens* de contexto. Es decir, el conjunto de *token* se conforma con toda la información que participa en el uso del sistema, esto es, cada palabra que se incorpora a una instrucción, las preguntas formuladas por el usuario, las respuestas anteriormente generadas, los documentos incorporados y cualquier otra información que el modelo deba considerar para elaborar una respuesta. En consecuencia, la capacidad real de una ventana de contexto no se mide únicamente por el número de páginas o documentos que pueden introducirse, sino por la cantidad total de *tokens* que la totalidad representa. Es esencial entender este concepto para obtener una buena respuesta. Cuando el volumen de información que se considera supera al límite de *tokens* que el modelo puede gestionar simultáneamente, hay que priorizar contenidos, y por eso a veces el sistema resume documentos o selecciona solo ciertos fragmentos que sean más relevantes para la tarea en cuestión. De esta manera, se intenta evitar que determinada información importante quede fuera del contexto en el que trabaja el sistema a la hora de dar la respuesta.

Desde una perspectiva práctica, la gestión de los *tokens* de contexto es fundamental, y de ahí que cuanto más relevante y estructurada sea la información suministrada mayor será la probabilidad de obtener respuestas precisas, coherentes y alineadas con los objetivos de la consulta.

Embeddings / Incrustaciones vectoriales*, por J. A. Ortega Ramírez*

Los *embeddings* es la forma en la que una IA realiza la transformación de la información de partida en otras representaciones numéricas que permitan procesarla, esto es, que interprete lo que significa, cómo se puede asimilar con otras entidades, aunque en sí sean palabras diferentes. A veces se utiliza el concepto de incrustaciones vectoriales y, de una manera más formal, podría definirse como la manera de obtener el significado de una información, reconociendo similitudes y relaciones con otros conceptos aunque se expresen con palabras o términos diferentes. El fin que se persigue con esta representación es la de situar contenidos con significados semejantes cerca unos de otros dentro de un espacio matemático, y de esta manera un sistema reconoce la relación entre diferentes expresiones con igual o parecida semántica. Un ejemplo podría ser la idea de contrato de arrendamiento y la de alquiler de vivienda, que están cercanos en semántica aunque sean diferentes palabras. Esta es la base de muchos sistemas de análisis y se utiliza, entre otros ámbitos, para los sistemas de recuperación de información, las búsquedas semánticas, la clasificación de documentos o los sistemas recomendadores.

Espacio latente*, por J. A. Ortega Ramírez*

El espacio latente se puede interpretar como un mapa abstracto en el que textos, imágenes o conceptos se distribuyen según sus semejanzas, que se adquieren en el entrenamiento del sistema. Se puede ver como una representación en la que el modelo busca organizar de alguna manera los rasgos, relaciones y posibilidades entre los elementos que no son observables de forma directa. Es la estructura que tienen las relaciones entre los elementos, suele ser interna y normalmente se define de manera matemática. Las capacidades y limitaciones de la IA generativa se sustentan en una mejor o peor definición del espacio latente en el que se ha aprendido. De ahí que en cada evolución se intente mejorar este espacio y que haya diferentes modelos que funcionen mejor o peor dependiendo de lo que vayamos buscando. Para los mismos conceptos de partida original se generan diferentes espacios latentes en función de lo que se vaya buscando en el mismo.

Destilación de modelos*, por J. A. Ortega Ramírez*

La destilación de modelos es una técnica que resume, digámoslo así, un modelo mayor en uno más pequeño. En este proceso lo que se persigue es buscar modelos menos pesados, y para ello se transfiere parte del comportamiento del modelo grande o con mayores capacidades, pero que requiere más cantidad de procesamiento o más tiempo para dar una respuesta, a otro modelo que es

más pequeño, rápido y con menos necesidades para que se pueda utilizar. Esta estrategia persigue utilizar la IA con menos capacidades, con la idea de poder ver qué cosas se pueden hacer, pero sin necesidad de grandes sistemas. Asimismo, esta destilación permite que se pueda utilizar la IA en entornos con recursos más limitados o que requieren de menor tiempo para dar una respuesta. Es obvio que la respuesta del sistema menor suele ser peor, pero puede ser suficiente para que lo se necesita o es la que se puede aplicar.

Cuantización, *por J. A. Ortega Ramírez*

La cuantización es una técnica de optimización que reduce la precisión con la que se almacenan o calculan los parámetros de un modelo. Esta reducción hace que el modelo ocupe menos memoria, responda con mayor rapidez y permita ejecutarse en infraestructuras menos potentes, como podrían ser móviles u otros dispositivos con menor potencia. Es apropiado indicar que una cuantización excesiva suele degradar la calidad de las respuestas o bien hacer que los resultados no sean de la calidad que se requiere. Hay que llegar siempre a un equilibrio donde la cuantización permita su uso pero que afecte lo menos posible a los resultados.

Orquestación, *por J. A. Ortega Ramírez*

La orquestación es la coordinación y gestión ordenada de los distintos componentes que intervienen en un sistema de información. Su objetivo es garantizar la intervención de cada elemento en su momento, conforme a una secuencia lógica y con el objetivo de conseguir un resultado que sea coherente, eficiente y alineado con los objetivos establecidos.

En un sistema de IA generativa, para obtener una respuesta es necesario que se apliquen diferentes componentes, desde diferentes modelos de IA, herramientas externas, bases de datos, sistemas de recuperación de información, mecanismos de verificación, reglas de decisión, flujos automatizados de trabajo y controles de seguridad. La orquestación es el conjunto de reglas y procedimientos que organiza la manera en la que interaccionan todos estos elementos para que funcionen como un sistema unificado, aunque internamente sean independientes.

Desde una perspectiva funcional, la orquestación incluye también los recursos que deben utilizarse en cada fase del proceso, qué información deben intercambiarse y qué comprobaciones hay que realizar para generar una respuesta o ejecutar una acción. Gracias a esta coordinación es posible combinar capacidades especializadas y construir soluciones mucho más robustas que las que podría ofrecer un modelo aislado. Cuanto mayor es la capacidad que ofrecen los sistemas, más necesario es que se definan unos buenos mecanismos de orquestación. La

calidad de la respuesta, por tanto, no solo depende de los modelos empleados, sino también de la manera en que se estructuran las interacciones entre los distintos componentes. Una arquitectura correctamente orquestada permite mejorar la precisión, la eficiencia, la trazabilidad y la seguridad del sistema. Sin embargo, una coordinación deficiente puede convertir un conjunto de tecnologías avanzadas en una solución fragmentada, difícil de controlar y con resultados inconsistentes.

La orquestación constituye uno de los elementos centrales en el diseño de aplicaciones profesionales de IA, ya que transforma un conjunto de capacidades técnicas dispersas en un sistema capaz de operar de forma coherente, fiable y orientada a objetivos concretos.

Prompt*, por J. A. López de Fez*

Un *prompt* es la instrucción o indicaciones que un usuario introduce en un sistema de IA generativa para obtener una respuesta. Funciona como interfaz entre el ser humano y el modelo de lenguaje, y determina qué porción de su conocimiento estadístico se activa y cómo se estructura la salida. No es una pregunta que el modelo entienda en sentido humano, sino un patrón estadístico que condiciona la distribución de probabilidades de la respuesta. Para saber más sobre el mecanismo interno, véase el término *arquitectura de transformadores*.

Es fundamental comprender que un *prompt* no equivale a una consulta en una base de datos jurídica, ya que ésta recupera documentos preexistentes (aunque el orden o la forma de presentación pueda depender de criterios de búsqueda, relevancia o funcionalidades avanzadas), mientras que el *prompt* genera texto nuevo que puede contener información correcta, incompleta, incorrecta o inventada, sin que el formato de la respuesta permita distinguir por sí solo unas categorías de otras. Además, el mismo *prompt* lanzado en momentos distintos puede producir respuestas diferentes, porque la generación incorpora aleatoriedad controlada mediante el parámetro de temperatura, lo que afecta a la fiabilidad y al valor probatorio de los resultados.

La **ingeniería de *prompts*** consiste en diseñar y refinar las instrucciones proporcionadas a un modelo para maximizar la precisión de sus respuestas, y pueden ser de varios tipos:

- Los de "cero disparos" (*zero-shots*) plantean una cuestión sin ejemplos.
- Los de "pocos disparos" (*few-shots*) incluyen ejemplos de la respuesta esperada, homogeneizando resultados en tareas repetitivas.

- Los que solicitan una explicación estructurada piden al modelo que ordene los pasos de su respuesta y haga visibles sus criterios, lo que puede facilitar la detección de errores, aunque no permite acceder a su razonamiento interno real.

- Los de sistema establecen rol, tono y restricciones antes de la consulta, permitiendo, por ejemplo, configurar al modelo como especialista en una jurisdicción o prohibir la invención de jurisprudencia.

Un *prompt* mal diseñado incrementa la probabilidad de alucinación, uno sesgado amplifica el sesgo en la salida, y su calidad condiciona la supervisión humana efectiva, por lo que una adecuada formulación contribuye a la transparencia algorítmica.

El *prompt* tiene una limitación operativa, ya que el modelo solo procesa lo que cabe en su ventana de contexto, medida en *tokens*. Lo que exceda ese límite no existe para el sistema, lo que obliga a segmentar el material antes de la consulta.

En el ejercicio del Derecho, la capacidad de hacer buenos *prompts* constituye una competencia profesional cuya deficiencia puede generar responsabilidad. Un abogado obtendrá resultados radicalmente distintos si su *prompt* dice "revisa este contrato" frente a si especifica que debe identificar cláusulas abusivas conforme a la Directiva 93/13/CEE, señalar riesgos de nulidad parcial y proponer redacciones alternativas con referencia normativa.

El *prompt* plantea una cuestión de propiedad intelectual aún abierta. La Oficina de Derechos de Autor de EEUU viene rechazando la protección del material generado exclusivamente por IA, aunque un *prompt* podría protegerse como texto autónomo si alcanza por sí mismo el umbral de originalidad. La cuestión permanece sin resolver en la doctrina europea.

El *prompt* presenta riesgos de seguridad específicos. Mediante inyección de *prompts*, un tercero puede insertar instrucciones ocultas en un documento que el modelo procesará como parte de la entrada, haciendo que ignore restricciones, revele información confidencial o genere resultados manipulados. Esto afecta a la integridad de la revisión documental automatizada y a la cadena de responsabilidad.

Su uso profesional compromete potencialmente tanto la protección de datos como el secreto profesional. Introducir datos personales implica tratamiento sometido al RGPD pero, incluso sin ellos, introducir hechos del caso o comunicaciones con el cliente puede vulnerar el deber de secreto profesional, obligación independiente cuya infracción acarrea consecuencias disciplinarias y penales.

El *prompt* puede constituir elemento probatorio en procedimientos donde se utilice IA generativa, razón por la que algunas firmas exigen registrar los *prompts* empleados en tareas asistidas.

Alucinación, *por C. Pérez García*

Una alucinación es un fenómeno irreal que se presenta a un observador como real. El término se toma de la psiquiatría, en la que significa una percepción dada sin objeto percibido: en cualquiera de los cinco sentidos, las alucinaciones producen en la mente del sujeto percepciones sensoriales falsas, es decir, la ilusión de ver, oír, oler, saborear o tocar algo que, en realidad, no existe. En un nivel de definición muy abstracto, se puede decir que la alucinación consiste en un *output*, la percepción, que no se corresponde con lo que de manera cabal cabría esperar sobre la base del *input*, la realidad sensible. Esto explica el uso que se ha dado al término, por analogía, en el campo de la informática para significar resultados del sistema no esperables conforme a las instrucciones y datos de entrada.

El tópico de las alucinaciones de los sistemas informáticos se ha revitalizado con los sistemas de IA generativa basados en *deep learning*. El *input* de estos sistemas viene constituido por los datos empleados durante el entrenamiento y el *prompt* del usuario, factores de cuyo producto resulta, como *output*, la respuesta. Sin embargo, el *output* real expulsado por el sistema no siempre resulta cohonestable con el *output* verosímil prefigurado en la mente del operador humano. Por ejemplo, hablamos de alucinación cuando un *chatbot* genera un texto con datos falsos sobre un evento histórico o cuando un generador de imágenes produce una composición con errores anatómicos, supuesto siempre que, conforme a la instrucción de entrada, no fuesen tales productos lo que se hubiese demandado al sistema, en cuyo caso sí serían el resultado cabal y esperable.

Las alucinaciones se deben a que el proceso por el que se generan textos, imágenes o cualesquiera otros productos no implica un razonamiento conceptual, sino un mero cálculo de probabilidad. Una persona aprehende el contenido conceptual de lo que otra le dice y, tratando de ofrecer la respuesta más plausible, hilvana conceptos para formular proposiciones y expresa estas mediante oraciones; si bien se da asimismo un cálculo de probabilidad, pues se juzga cuál será la respuesta mejor en el caso dado, no se trata más que de una parte dentro de un razonamiento más complejo. Un sistema de IA, en cambio, tan solo calcula la probabilidad de que el operador humano acepte cada uno de los caracteres, píxeles o cualesquiera otros elementos que le ha de ofrecer y expulsa la combinación y orden para los que haya calculado la mayor probabilidad. Así, por ejemplo, mientras que un artista humano decide pintar una mano con un determinado número de dedos, un sistema de IA tan solo yuxtapone píxeles en la

disposición más probable, sin saber qué es una mano o un dedo y, por tanto, sin control conceptual de la plausibilidad del resultado. El conocimiento conceptual sirve como base sobre la que sustentar una respuesta y como instancia desde la que auditarla; desprovistos de él, los sistemas de IA tienen, en consecuencia, una propensión mayor a generar textos, imágenes y demás productos inverosímiles, fruto de un cálculo de probabilidad carente de toda guía conceptual.

Puesto que los sistemas de IA generativa suelen basarse en *deep learning*, el proceso entre el *input* y el *output* acostumbra a ser una caja negra, de suerte que no se pueden controlar las operaciones internas del sistema. Por tanto, la única manera de detectar una alucinación es conocer de antemano el marco de resultados cabales. El principal problema, pues, es que no todos los marcos de resultados cabales son fáciles de controlar. En los ejemplos dados, resultaría mucho más fácil detectar una alucinación en una imagen de un ser humano, pues bastaría conocer la anatomía básica, que en el texto sobre un evento histórico, pues sería preciso conocer todos los datos implicados de manera directa o indirecta en el texto. Esto hace que las alucinaciones sean fenómenos de especial peligro para usuarios inexpertos o con poco dominio de la materia sobre la que opera el sistema.

Razonamiento en cadena como fenómeno técnico, *por J. A. Ortega Ramírez*

El razonamiento en cadena, entendido como fenómeno técnico, describe la generación de pasos intermedios que ayudan al modelo a abordar una tarea compleja de manera progresiva. En esta secuencia, el modelo descompone el problema en etapas más sencillas, y se hace con la intención de obtener respuestas más precisas y coherentes.

Sin embargo, este proceso no debe interpretarse como una prueba de que el sistema razona o comprende del mismo modo que lo hace una persona. Es un mecanismo técnico para organizar la generación de la respuesta y que, a veces, consigue mejorar la respuesta del modelo al estructurar el análisis de una manera más ordenada.

Aunque esta técnica puede ser útil para ciertos problemas complejos o para intentar justificar ciertas conclusiones, a veces puede generar explicaciones posibles pero incorrectas. Por ello, la calidad de una respuesta no debe medirse únicamente porque los pasos que se definan sean razonables, sino que debe realizarse por la corrección, consistencia y verificabilidad del resultado final obtenido.

Salida estructurada, *por J. A. Ortega Ramírez*

La salida estructurada es la forma de presentar la respuesta de un sistema. A diferencia de una respuesta en texto libre, estas respuestas organizan los datos de manera que puedan ser revisados, procesados o integrados fácilmente en otros sistemas, y suele ser conforme a un formato previamente definido, como puede ser una tabla, un formulario, un esquema documental, una lista de campos o un archivo JSON.

Su principal ventaja es que reduce ambigüedades e interpretabilidades y facilita la automatización de tareas, permitiendo que la información generada pueda utilizarse en otros entornos. Por ello, las salidas estructuradas son especialmente útiles en ámbitos como la gestión documental, el análisis de datos, la administración electrónica o la integración con aplicaciones empresariales.

En términos prácticos, una salida estructurada es el paso de una respuesta libre a un resultado preparado para su utilización posterior.

Medios sintéticos (*Synthetic media*), *por J. A. Ortega Ramírez*

Los medios sintéticos o *synthetic media* son contenidos creados, modificados o recreados mediante sistemas de IA. Estos incluyen hoy en día una amplia variedad de formatos, como textos, imágenes, voces, vídeos, avatares digitales o escenas generadas íntegramente de manera digital. Su característica común es que el contenido final no procede directamente de una captura o producción humana convencional, sino que es el resultado de procesos y sistemas computacionales que son capaces de generarlos a partir de datos y modelos que han sido previamente entrenados.

El desarrollo de los medios sintéticos ha abierto importantes oportunidades en ámbitos como la educación, la comunicación, la producción audiovisual o la creación artística, entre otros. Estas tecnologías permiten generar nuevos materiales adaptados a contenidos para diferentes públicos y ampliar de esta manera las posibilidades de interacción y expresión en los entornos digitales.

Esta creciente posibilidad de generación abre importantes desafíos sobre derechos de autor y autoría. Estos mismos mecanismos que generan nueva información se pueden emplear para generar información no real, difundir desinformación, manipular los contenidos, suplantar identidades o fabricar materiales falsos con apariencia de que son auténticos.

El debate actual trasciende el ámbito puramente tecnológico. Algunas cuestiones se centran en la identificación de los contenidos generados

artificialmente, la transparencia sobre quién los ha generado, esto es, su origen, la protección de los derechos de los contenidos o de las personas afectadas y la determinación de las responsabilidades jurídicas derivadas de su creación y utilización. Es obvio que aquí es donde en la actualidad recae uno de los retos de gobernanza de la IA.

Procedencia del contenido (*Provenance*), *por J. A. Ortega Ramírez*

La procedencia del contenido, o *provenance*, es el conjunto de información que permite conocer el origen, la autoría y las modificaciones realizadas sobre un contenido digital a lo largo de todo su ciclo de vida. Incluye datos relevantes sobre quién creó el material, cuándo fue generado o editado y qué herramientas o procesos intervinieron en su elaboración.

En un contexto donde la IA facilita la creación y transformación de contenidos, su procedencia es un elemento importante para evaluar su autenticidad. Aunque no garantiza que la información es verdadera, sí que aporta trazabilidad y facilita la verificación de su origen.

Por ello, la procedencia del contenido desempeña un papel fundamental en ámbitos como la prueba digital, la transparencia informativa, la protección frente a la desinformación y la generación de confianza en los entornos digitales.

Segunda Parte.
ECOSISTEMA ÉTICO Y JURÍDICO DE LA IA

I. Ética de la IA

Gobernanza de la IA, *por C. Villegas Delgado*

De conformidad con lo dispuesto por el diccionario de la Real Academia Española, el vocablo "gobernanza" es definido como la "forma de gobernar, especialmente si es eficaz y adecuada a los fines que se persiguen". Por su parte, el término Inteligencia Artificial (IA) se define como aquella "disciplina científica que se ocupa de crear programas informáticos que ejecutan operaciones comparables a las que realiza la mente humana, como el aprendizaje o el razonamiento lógico". Tomando como punto de partida estas dos definiciones, y trascendiendo la mera suma de las mismas, podemos afirmar que la gobernanza de la IA "se refiere al conjunto de principios, normas, instituciones y procesos destinados a orientar, regular y supervisar el diseño, desarrollo, despliegue y utilización de los sistemas de IA, procurando que su funcionamiento responda de manera eficaz a objetivos socialmente legítimos y sea compatible con la protección de los derechos humanos, el Estado de Derecho y los valores que distinguen a toda sociedad democrática". Debemos admitir, en este punto, que la definición propuesta no es neutral en lo absoluto, toda vez que incorporamos dentro de ella un contenido material que se identifica con una serie de valores y principios fundantes que sirven de inspiración y límite a la hora de regular la IA, por lo menos en Europa.

Lo anterior permite identificar, al menos, tres modelos de gobernanza de la IA: la defendida por la Unión Europea, la promovida por los Estados Unidos de América y la implementada por la República Popular de China. El modelo de gobernanza de la IA desarrollado por la Unión Europea constituye, en la actualidad, el referente normativo más avanzado a nivel internacional. Desde sus primeras iniciativas en materia de IA confiable, la Unión ha impulsado un enfoque centrado en el ser humano (*human-centric AI*), orientado a garantizar que el diseño, desarrollo y utilización de estas tecnologías respeten los derechos fundamentales, la transparencia, la seguridad, la rendición de cuentas y los valores democráticos. Este modelo se sustenta, en buena medida, en la experiencia

adquirida con el Reglamento General de Protección de Datos (RGPD), que transformó los estándares internacionales de protección de datos personales y reforzó la confianza ciudadana en el entorno digital. Un paso decisivo en la consolidación de este modelo fue la aprobación del Reglamento (UE) 2024/1689 por el que se establecen normas armonizadas en materia de IA, primer instrumento jurídicamente vinculante del mundo destinado a regular de forma integral los sistemas de IA mediante un enfoque basado en el riesgo. El Reglamento fue publicado en el Diario Oficial de la Unión Europea en julio de 2024 y entró en vigor el 1 de agosto de 2024. Gracias a estos avances, la Unión Europea se ha consolidado como el principal referente internacional en materia de gobernanza de la IA, sirviendo de inspiración para el diseño de marcos regulatorios en otras regiones del mundo, particularmente en América Latina.

El modelo chino de gobernanza de la IA se caracteriza por un enfoque centralizado y dirigido por el Estado, que concibe la IA como una herramienta estratégica para impulsar el desarrollo económico, la seguridad nacional y el fortalecimiento del poder estatal. A diferencia del modelo europeo, que sitúa la protección de los derechos fundamentales en el centro de la regulación, China prioriza la innovación, la competitividad tecnológica y la estabilidad social. Este modelo se articula mediante una estrecha coordinación entre el Gobierno y las grandes empresas tecnológicas, apoyada en estrategias como el Plan de Desarrollo de la Nueva Generación de IA de 2017 y la iniciativa *Made in China* 2025. Desde el punto de vista regulatorio, China ha aprobado normas específicas sobre algoritmos, IA generativa y tecnologías de síntesis profunda, orientadas principalmente a garantizar la seguridad nacional, el control de los contenidos digitales y la supervisión de los operadores tecnológicos. En consecuencia, el modelo chino combina una fuerte planificación estatal con un amplio impulso a la innovación, subordinando la protección de los derechos individuales a los intereses colectivos y a la seguridad del Estado.

Finalmente, el modelo estadounidense de gobernanza de la IA se caracteriza por un enfoque flexible y orientado al mercado, en el que la innovación y el liderazgo tecnológico prevalecen sobre una regulación integral. A diferencia de la Unión Europea, Estados Unidos ha optado tradicionalmente por favorecer la autorregulación, la competencia y el desarrollo impulsado por el sector privado, reservando la intervención pública para ámbitos específicos relacionados con la seguridad nacional, la protección del consumidor y la competencia. La gobernanza de la IA se articula mediante una combinación de iniciativas federales, normas sectoriales y directrices no vinculantes, con un papel protagonista de las grandes empresas tecnológicas y de los centros de investigación. En los últimos años, el gobierno federal ha reforzado este modelo mediante estrategias nacionales y órdenes ejecutivas dirigidas a promover una IA segura y fiable, sin renunciar a mantener la competitividad de Estados Unidos frente a otros

actores, especialmente China. En consecuencia, el modelo estadounidense busca equilibrar la innovación tecnológica con una regulación limitada, confiando en gran medida en la capacidad del mercado y del sector privado para impulsar el desarrollo de la IA.

Ética de la IA, *por F. H. Llano Alonso*

La irrupción de la IA en todos los ámbitos de la vida humana constituye uno de los fenómenos más transformadores de nuestra era. No se trata únicamente de un cambio tecnológico, sino de una auténtica metamorfosis del mundo -en el sentido que Ulrich Beck otorgaba a este concepto- que afecta de manera profunda a las estructuras jurídicas, políticas y morales sobre las que se asienta la convivencia democrática. En este nuevo escenario posthumano, el ser humano ya no es el único agente inteligente: coexiste en el espacio digital con entidades artificiales capaces de analizar su entorno y actuar con un grado significativo de autonomía. La cuestión ya no es si la IA transformará el modelo clásico de interacción humana, sino en qué medida estamos preparados para asumir las implicaciones éticas y jurídicas de esa transformación.

Esta nueva realidad plantea interrogantes de calado filosófico y jurídico que no admiten respuestas simples. ¿Puede una máquina ser objeto de valoración ética? ¿Qué ocurre cuando un algoritmo toma decisiones que afectan a los derechos fundamentales de las personas? ¿Cómo garantizar que el desarrollo de la IA se oriente hacia el bien común y no devenga en un instrumento de dominación o discriminación? Estas preguntas tienen consecuencias directas sobre millones de ciudadanos afectados por decisiones automatizadas en ámbitos tan sensibles como la salud, el empleo, la justicia o el acceso a servicios financieros. No se trata, pues, de especulaciones filosóficas, sino de urgencias propias de la gobernanza democrática contemporánea.

Desde la perspectiva de la Filosofía del Derecho, resulta imprescindible subrayar que la IA no es un fenómeno ajeno a la condición humana: es, por naturaleza, antropogénica, es decir, diseñada, programada y desplegada por seres humanos con intenciones y valores concretos. Esta circunstancia justifica tanto la posibilidad de someterla a un juicio ético como la necesidad de regularla jurídicamente. No basta con enunciar principios deontológicos -transparencia, equidad o no discriminación-, si estos no van acompañados de mecanismos normativos vinculantes. Como ha advertido Brent Mittelstadt, los principios por sí solos no pueden garantizar una IA ética; es imprescindible la implementación de un marco normativo robusto.

La Unión Europea ha asumido un papel pionero en la construcción de dicho marco jurídico. Desde la Resolución del Parlamento Europeo de 2017

sobre macrodatos y derechos fundamentales, pasando por el Libro Blanco sobre Inteligencia Artificial de 2020 -que puede considerarse la Carta Magna de la ética de la IA en Europa-, hasta la propuesta de Reglamento sobre principios éticos para el desarrollo, despliegue y uso de la IA, se está configurando un *corpus iuris digitalis* orientado a garantizar que la IA sea antropocéntrica, fiable y respetuosa con los derechos fundamentales recogidos en la Carta de la UE.

Este marco regulador se articula en torno a cuatro ejes: el carácter antropocéntrico y antropogénico de la IA; la ausencia de sesgos y discriminación algorítmica; la responsabilidad civil por daños causados por sistemas de alto riesgo; y la buena gobernanza. La discriminación algorítmica no es una amenaza hipotética: los algoritmos de contratación o de concesión de créditos pueden reproducir y amplificar prejuicios sociales si no son diseñados y auditados con el rigor que exige el respeto a los derechos fundamentales.

Frente a quienes confían en la autorregulación del sector tecnológico, la experiencia invita a la prudencia. Las campañas de *ethics washing* protagonizadas por algunas grandes compañías tecnológicas demuestran que el compromiso declarativo con la ética puede convertirse en un instrumento de legitimación que encubre prácticas lesivas para los derechos ciudadanos. El camino más sólido es el de la regulación jurídica con fuerza vinculante, complementada con principios éticos y disposiciones de *soft law* que orienten su interpretación y aplicación.

En definitiva, la ética de la IA es una exigencia irrenunciable del Estado de Derecho en la era digital. Garantizar que las tecnologías de IA se desarrollen de conformidad con los valores que inspiran la Carta de los Derechos Fundamentales de la Unión Europea -dignidad humana, libertad, igualdad y justicia- constituye uno de los retos más urgentes de nuestra época. Ante ese reto, la reflexión filosófico-jurídica y la acción normativa no son alternativas, sino aliados necesarios.

Principio de no malevolencia (*Non-maleficence*), *por F. H. Llano Alonso*

La no malevolencia es el primero de los cinco principios éticos fundamentales recogidos en la obra de Luciano Floridi *Ética de la IA* (*The Ethics of Artificial Intelligence*, 2023) como marco normativo para el diseño, el desarrollo y el despliegue de sistemas de IA. Su formulación parte de una analogía con el principio hipocrático de la medicina -*primum non nocere*, «ante todo, no dañar»-, trasladado ahora al ámbito de las tecnologías inteligentes. En virtud de este principio, los sistemas de IA no deben causar daño a las personas ni a la sociedad en su conjunto, y quienes los diseñan, los desarrollan o los regulan tienen la obligación de anticipar, identificar y mitigar activamente los riesgos asociados a su funcionamiento.

El principio de no malevolencia no se limita a la prohibición de causar daño de manera intencionada. Abarca también los perjuicios no previstos o indirectos que pueden derivarse del funcionamiento de sistemas automatizados de toma de decisiones, de algoritmos de aprendizaje automático o de sistemas de recomendación. En este sentido, la no malevolencia exige una actitud prospectiva y preventiva por parte de todos los actores involucrados en el ciclo de vida de la IA: hay que anticipar el daño antes de que se produzca, no solo reaccionar ante él una vez que ya ha ocurrido. Esta dimensión preventiva convierte a la no malevolencia en un principio especialmente exigente desde el punto de vista técnico y ético, pues obliga a modelizar escenarios de riesgo que, en muchos casos, aún no se han materializado pero cuya probabilidad puede evaluarse mediante herramientas de análisis de riesgos.

En este marco, se distingue entre daños directos, causados inmediatamente por el sistema de IA, y daños indirectos, que se producen como consecuencia de decisiones o acciones facilitadas por dicho sistema. Ambas categorías son relevantes desde el punto de vista de la no malevolencia. Un sistema de IA puede ser técnicamente correcto en su funcionamiento y, sin embargo, producir efectos lesivos a escala social cuando opera sobre datos sesgados, cuando se aplica en contextos para los que no fue diseñado o cuando sus resultados son interpretados y aplicados por usuarios que desconocen sus limitaciones. La cadena de responsabilidad que se deriva de estos daños indirectos plantea complejos problemas jurídicos que el Derecho aún no ha resuelto satisfactoriamente y que constituyen uno de los principales desafíos del marco regulatorio de la IA.

En el contexto jurídico, este principio adquiere especial relevancia en ámbitos como la justicia penal predictiva, la vigilancia automatizada o los sistemas de puntuación crediticia. Como he señalado en *Homo ex machina* (2023), la rápida penetración de la IA en la sociedad no ha ido acompañada de una investigación exhaustiva sobre los problemas sociopolíticos que hacen que ciertos grupos vulnerables se vean perjudicados en lugar de beneficiados por ella, lo que convierte la no malevolencia en una exigencia ineludible del ordenamiento jurídico. Sistemas como el algoritmo COMPAS -utilizado en Estados Unidos para calcular probabilidades de reincidencia delictiva- han sido objeto de críticas severas por introducir sesgos raciales que perjudican de forma desproporcionada a determinadas minorías, vulnerando de manera flagrante este principio ético y jurídico.

En el marco regulatorio europeo, la no malevolencia se traduce en la exigencia de evaluaciones de impacto sobre los derechos fundamentales, en la prohibición de usos inaceptables de la IA -como la manipulación subliminal de conductas o la puntuación social generalizada de ciudadanos- y en la obligación de establecer mecanismos de supervisión humana efectiva que actúen como

cortafuegos ante posibles daños. El Reglamento de IA de la Unión Europea (2024) recoge implícitamente este principio al clasificar como de riesgo inaceptable aquellos sistemas cuyo potencial de daño supera cualquier beneficio razonable. En definitiva, la no malevolencia no es simplemente una aspiración ética abstracta, sino un principio operativo que debe orientar tanto el diseño técnico de los sistemas de IA como su regulación jurídica, estableciendo límites claros e irrenunciables a lo que es permisible en nombre del progreso tecnológico.

Principio de beneficencia (*Beneficence*)*, por F. H. Llano Alonso*

La beneficencia constituye el segundo principio ético de la IA. Si la no malevolencia exige abstenerse de causar daño, la beneficencia va un paso más allá al requerir que los sistemas de IA aporten un beneficio positivo, real y verificable para las personas y para la sociedad en su conjunto. No basta con evitar el daño; es necesario que la IA contribuya activamente al bienestar humano, a la reducción de desigualdades y al florecimiento de capacidades individuales y colectivas. La diferencia entre ambos principios es análoga, en el ámbito de la ética aplicada, a la distinción entre una obligación negativa -la de no perjudicar- y una obligación positiva -la de hacer el bien-, diferencia que tiene notable relevancia normativa en la teoría del Derecho y en la filosofía moral.

La beneficencia no puede entenderse en términos exclusivamente utilitarios, como si la maximización del beneficio total justificara cualquier medio empleado para obtenerlo. Por el contrario, el principio exige una concepción plural del bien, atenta a las diferencias entre individuos y grupos, a las asimetrías de poder y a la diversidad de valores que coexisten en las sociedades democráticas. Un sistema de IA es beneficente cuando sirve genuinamente a quienes debe servir, y no cuando simplemente optimiza métricas de rendimiento que benefician a sus creadores o a los sectores más favorecidos de la sociedad. En este sentido, la beneficencia impone a los diseñadores y reguladores de la IA la carga de justificar que sus sistemas producen un bien efectivo y distribuido equitativamente, y no meramente un beneficio privado disfrazado de interés general.

En el ámbito jurídico, la beneficencia se manifiesta en la expectativa de que las aplicaciones de IA en la administración de justicia, la atención sanitaria, la educación o los servicios públicos mejoren efectivamente el acceso y la calidad de dichos servicios para toda la ciudadanía, incluidos los colectivos más vulnerables. Como he señalado en *Homo ex machina* (2023), quienes abogan por la digitalización de la justicia argumentan que esta puede promover la justicia proporcional -más rápida, económica y accesible para los ciudadanos- y la justicia distributiva, que garantiza una mejor asignación de recursos jurisdiccionales. Sin embargo, dichos beneficios no son automáticos ni están garantizados: dependen

del modelo de justicia por el que se opte y de las salvaguardas institucionales que se establezcan para evitar que la eficiencia se imponga sobre la equidad.

Este principio vincula estrechamente la beneficencia con el concepto de «florecimiento» (*flourishing*) de los agentes que interactúan con los sistemas de IA. La IA beneficente es aquella que amplía las capacidades de acción de las personas, que les proporciona herramientas para tomar mejores decisiones y que enriquece sus posibilidades de vida, sin sustituirlas en aquellas dimensiones que son constitutivas de su dignidad y de su autonomía. Un sistema de IA puede ser beneficente en términos agregados pero pernicioso para determinados individuos, razón por la cual el principio de beneficencia no puede aplicarse de manera puramente estadística, sino que debe articularse con el principio de justicia para garantizar que ningún grupo quede excluido de sus efectos positivos.

Este principio plantea también una exigencia de equidad en la distribución de los beneficios de la IA. La brecha entre quienes tienen y quienes no tienen acceso a la IA aumentará en lugar de disminuir si no se adoptan medidas específicas, si los beneficios de la IA se concentran en sectores ya privilegiados de la población el resultado neto puede ser un incremento de la desigualdad estructural, lo que contradice de raíz el principio de beneficencia. En el plano regulatorio, la beneficencia se traduce en obligaciones positivas para los Estados y para las empresas: invertir en IA orientada al interés general, fomentar el acceso equitativo a sus aplicaciones y garantizar que los sistemas de IA desplegados en servicios públicos respondan efectivamente a las necesidades de toda la ciudadanía, sin excluir a quienes más los necesitan.

Principio de justicia (*Justice / Fairness*), *por F. H. Llano Alonso*

La justicia es el tercero de los cinco principios éticos de la IA formulados en la obra de L. Floridi *Ética de la IA* (2023), y quizá el más complejo desde el punto de vista conceptual, dado que la noción de justicia ha sido objeto de múltiples interpretaciones a lo largo de la historia del pensamiento filosófico y jurídico. En el contexto de la ética de la IA, la justicia se entiende en un sentido amplio que engloba la equidad en el trato, la no discriminación, la distribución equitativa de beneficios y cargas, y la garantía de que los sistemas de IA no reproduzcan ni amplifiquen injusticias estructurales preexistentes. La relevancia de este principio para el Derecho es inmediata: la justicia, entendida como valor superior del ordenamiento jurídico, no puede ser indiferente a los efectos que producen los sistemas algorítmicos cuando se integran en los procesos de toma de decisiones con consecuencias jurídicamente relevantes para los ciudadanos.

Uno de los problemas centrales que aborda este principio es el de los sesgos algorítmicos. Los sistemas de IA aprenden de datos históricos que pueden reflejar

patrones de discriminación consolidados en la sociedad por razón de raza, género, origen étnico, nivel socioeconómico u otras características protegidas. Si estos sesgos no se identifican y corrigen adecuadamente, los algoritmos los perpetúan e incluso, en algunos casos, los intensifican. Como documentó ampliamente en *Homo ex machina* (2023), investigaciones como la publicada por ProPublica en 2016 revelaron que el algoritmo COMPAS, utilizado en el sistema judicial estadounidense para calcular probabilidades de reincidencia, presentaba sesgos contra la población afroamericana, produciendo tasas de falsos positivos que vulneraban de manera sistemática el principio de igualdad y no discriminación. El entonces Fiscal General de los Estados Unidos, Eric Holder, había denunciado esta situación ya en 2014, sin que se hubieran adoptado medidas correctoras efectivas.

En este marco, se distingue entre *fairness* como equidad procedimental -que los procedimientos sean imparciales y consistentes- y *fairness* como equidad sustantiva -que los resultados sean justos en términos distributivos-. Ambas dimensiones son necesarias y pueden entrar en tensión entre sí. Un sistema puede seguir procedimientos formalmente correctos y, sin embargo, producir resultados injustos debido a las características de los datos de entrenamiento o al diseño de sus variables de entrada. Por ello, la justicia algorítmica no puede agotarse en una verificación de los procedimientos técnicos; requiere también una evaluación periódica de los efectos reales del sistema sobre distintos grupos de la población, lo que implica métodos de auditoría algorítmica y mecanismos de corrección cuando se detectan disparidades injustificadas desde el punto de vista de los derechos fundamentales.

En el ámbito del Derecho, la justicia como principio de la ética de la IA conecta directamente con garantías constitucionales y convencionales ampliamente reconocidas: el derecho a la igualdad, el principio de no discriminación, el derecho al debido proceso y la tutela judicial efectiva. En *Homo ex machina* he analizado en profundidad la transformación del paradigma de justicia que comporta la extensión de los sistemas de justicia predictiva, señalando que la justicia argumentativa -basada en el razonamiento jurídico y en la ponderación de las circunstancias individuales- puede verse desplazada por un modelo de justicia algorítmica centrado en la predicción estadística, con el consiguiente riesgo de erosión de las garantías procesales que son constitutivas del Estado de Derecho democrático.

En el plano normativo, la justicia se operacionaliza mediante la exigencia de auditorías algorítmicas, la prohibición de perfilados discriminatorios, el derecho a no ser objeto de decisiones automatizadas que produzcan efectos jurídicos significativos sin posibilidad de revisión humana, y el mandato de diseñar sistemas de IA que sean igualmente precisos y fiables para todos los grupos de la población, y no solo para aquellos sobrerrepresentados en los datos de entrenamiento. Esta

perspectiva advierte asimismo sobre las asimetrías globales en el acceso a estas tecnologías, que generan desigualdades estructurales entre países y regiones que el principio de justicia obliga a incorporar en cualquier análisis ético serio de la gobernanza global de la IA, pues la justicia digital no puede ser patrimonio exclusivo de los países tecnológicamente más avanzados.

Principio de autonomía (*Autonomy*), *por F. H. Llano Alonso*

La autonomía constituye el cuarto principio del marco ético para la IA propuesto en la obra de L. Floridi *Ética de la IA* (2023). Su contenido es doble: por un lado, implica preservar y fomentar la autonomía de los seres humanos en sus relaciones con los sistemas de IA; por otro, exige mantener el control humano sobre dichos sistemas, evitando que la delegación progresiva de decisiones en algoritmos erosione la capacidad de las personas para actuar como agentes morales libres y responsables. Se trata de uno de los principios de mayor calado filosófico del conjunto, pues afecta a la propia concepción del ser humano como sujeto activo de sus decisiones y no como mero objeto de procesos automatizados que se ejecutan con independencia de su voluntad.

En este marco, se distingue entre la autonomía de los agentes artificiales -que puede ser deseable en ciertos contextos técnicos, como los sistemas de navegación autónoma, las aplicaciones médicas de diagnóstico automatizado o los procesos industriales repetitivos- y la autonomía moral de los seres humanos, que no puede ser cedida ni delegada a ningún sistema artificial. Esta segunda forma de autonomía es el núcleo del principio: los sistemas de IA deben diseñarse de modo que potencien, y no socaven, la capacidad de las personas para comprender las decisiones que les afectan, para cuestionarlas y para intervenir activamente en su revisión. El principio de autonomía establece, por tanto, un límite a la sustitución del juicio humano por el cálculo algorítmico, especialmente en aquellas decisiones que afectan a derechos fundamentales, a la libertad individual o a la dignidad de la persona humana.

Como señalo en *Homo ex machina* (2023), la creciente autonomía de las máquinas plantea una paradoja inquietante: a medida que los sistemas artificiales ganan en capacidad de decisión, los seres humanos corren el riesgo de perder la suya. Advierto allí que cuando los individuos van cediendo espacios de decisión a los algoritmos, van perdiendo también su autonomía, la libertad de obrar y decidir por sí mismos. Esta pérdida puede producirse de manera inadvertida o incluso voluntaria, cuando los usuarios se habitúan a delegar en sistemas de recomendación o de asistencia automatizada decisiones que antes eran estrictamente personales, como las relativas a la salud, las relaciones interpersonales, la información recibida o las opciones laborales y educativas.

La autonomía puede verse comprometida también a través de interferencias en el llamado control agencial: factores o agentes externos que alteran las facultades deliberativas de los individuos sin su conocimiento o consentimiento expreso. Las técnicas de personalización algorítmica que explotan vulnerabilidades cognitivas para dirigir el comportamiento de los usuarios hacia resultados predeterminados constituyen una amenaza directa a la autonomía, que el principio ético homónimo obliga a neutralizar mediante el diseño técnico y la regulación jurídica. Las burbujas de filtro y las cámaras de eco, que he documentado en los análisis sobre el impacto de las redes sociales en la identidad y en la autonomía personal, ilustran de manera concreta cómo los sistemas algorítmicos pueden restringir el horizonte deliberativo de las personas sin que estas lo adviertan.

En el contexto regulatorio europeo, el principio de autonomía se traduce en la exigencia de «*human in the loop*» para los sistemas de IA de alto riesgo: la obligación de mantener una supervisión humana efectiva que pueda corregir o anular las decisiones del sistema cuando sea necesario para proteger los derechos de los afectados. Este principio vincula también la autonomía con la responsabilidad moral distribuida: en una sociedad en la que la IA actúa como intermediaria en múltiples decisiones relevantes, es preciso articular mecanismos que preserven la responsabilidad individual y colectiva, evitando que la complejidad tecnológica se convierta en una pantalla que diluya la *accountability* de los actores humanos implicados en el diseño, el despliegue y la supervisión de estos sistemas. La autonomía, en definitiva, es tanto un límite al poder de la IA como un objetivo que la IA bien diseñada debe contribuir a realizar.

Principio de explicabilidad (*Explicability*), *por F. H. Llano Alonso*

La explicabilidad es el quinto y último de los principios éticos de la IA, y posiblemente el que presenta mayor especificidad técnica entre todos ellos. La obra de L. Floridi, *Ética de la IA*, define la explicabilidad como la capacidad de un sistema de IA para rendir cuentas de sus procesos y decisiones de manera inteligible para los diferentes interesados, ya sean operadores técnicos, reguladores, jueces, personas afectadas o ciudadanía en general. La explicabilidad es, en este sentido, una condición necesaria -aunque no suficiente- para que los demás principios éticos de la IA puedan ser verificados y exigidos en la práctica: sin ella, resulta imposible detectar daños, evaluar beneficios, identificar sesgos o comprobar si se respeta la autonomía de las personas. Puede afirmarse, en consecuencia, que la explicabilidad actúa como principio habilitador de todos los demás.

En este marco, se distinguen dos dimensiones complementarias de este principio. La primera es la inteligibilidad, que se refiere a la capacidad de explicar cómo funciona un sistema de IA en términos comprensibles para distintos

públicos, adaptando el nivel de detalle al tipo de interlocutor: no es lo mismo la explicación que se debe a un ingeniero que revisa el código que la que se debe a un ciudadano que ha sido objeto de una decisión automatizada. La segunda dimensión es la responsabilidad (*accountability*), que implica la posibilidad de atribuir consecuencias -sean jurídicas, éticas o sociales- a las decisiones adoptadas por el sistema o por quienes lo diseñan, despliegan o utilizan. Ambas dimensiones son necesarias para una explicabilidad genuina y no meramente formal o burocrática, pues sin inteligibilidad la *accountability* carece de base factual, y sin *accountability* la inteligibilidad no genera consecuencias jurídicas.

El problema central al que responde este principio es el de la «caja negra» (*black box*): muchos sistemas de IA avanzados, especialmente los basados en aprendizaje profundo (*deep learning*), producen resultados que sus propios diseñadores no pueden explicar con precisión, pues las decisiones emergen de patrones estadísticos de extraordinaria complejidad matemática. Esta opacidad es incompatible con las exigencias de un Estado de Derecho que garantice el control democrático de los sistemas que afectan a derechos fundamentales, al debido proceso o a las libertades individuales. En este contexto, la explicabilidad no es solo una virtud técnica deseable, sino una exigencia normativa que los ordenamientos jurídicos deben imponer de manera coercitiva a quienes desarrollan y despliegan estos sistemas en ámbitos de relevancia jurídica.

En el ámbito jurídico, la explicabilidad conecta directamente con garantías procesales fundamentales: el derecho a conocer las razones de una decisión que afecta a los derechos e intereses de una persona, el derecho a impugnarla ante una instancia de revisión y el principio de motivación de las resoluciones administrativas y judiciales. Como he señalado en *Homo ex machina* (2023), la necesidad de avanzar en la rendición de cuentas de los programas informáticos con capacidad para tomar decisiones algorítmicas resulta oportuna tanto en el ámbito de lo público como en el de lo privado, pues los riesgos derivados del funcionamiento técnico de los algoritmos afectan a los derechos e intereses de los ciudadanos con independencia de quién sea el titular del sistema que los administra. Siguiendo a Andrés Boix Palop, un mínimo de explicabilidad es exigible de conformidad con los principios constitucionales de seguridad jurídica y publicidad normativa.

En el plano normativo europeo, el Reglamento de IA exige que los sistemas de alto riesgo estén dotados de capacidades de registro, documentación y explicación suficientes para permitir la supervisión y la revisión por parte de las autoridades competentes. El Reglamento General de Protección de Datos ya había anticipado este principio al reconocer el derecho de las personas a no ser objeto de decisiones basadas exclusivamente en el tratamiento automatizado y a obtener información sobre la lógica aplicada en dicho proceso. La explicabilidad constituye, en

definitiva, la columna vertebral de la confianza pública en los sistemas de IA y el fundamento sobre el que se construye la legitimidad democrática de su despliegue en sociedades regidas por el Estado de Derecho.

Justicia algorítmica, *por A. J. Sánchez Sáez*

La Justicia algorítmica se erige como el pilar deontológico y normativo que debe regir la actividad administrativa y tecnológica en la era de la automatización. A diferencia de otros conceptos que se limitan a la detección de sesgos, este campo va un paso más allá: no se reduce a la mera "no discriminación", sino que busca una equidad proactiva. En el ámbito del Derecho Administrativo, esto implica que cualquier decisión automatizada que afecte a la esfera de derechos de los ciudadanos debe estar imbuida de una perspectiva de justicia social. Es, en esencia, la evolución del principio de interdicción de la arbitrariedad aplicado a las matemáticas del Estado, donde la administración no solo debe ser legal, sino también algorítmicamente justa en su impacto material.

Un algoritmo no es un ente neutral o una abstracción matemática pura; es una construcción humana que hereda las prioridades, valores y prejuicios de sus diseñadores. Por ello, la justicia algorítmica desafía la noción tradicional de objetividad técnica, reconociendo que los sistemas de IA son herramientas moldeadas por datos históricos que a menudo arrastran décadas de desigualdad. Si un modelo se entrena con bases de datos que reflejan disparidades raciales en detenciones policiales o brechas de género en salarios, el sistema no solo aprenderá esos patrones, sino que los automatizará a una escala masiva, dotándolos de una falsa apariencia de legitimidad numérica y científica que dificulta su impugnación por parte del ciudadano común.

Para contrarrestar estos riesgos, este enfoque propone pilares esenciales como la equidad, la transparencia y la explicabilidad. La equidad exige que los resultados no perjudiquen desproporcionadamente a grupos vulnerables, mientras que la explicabilidad demanda que las decisiones de las llamadas "cajas negras" puedan ser auditadas y comprendidas por seres humanos. La falta de este control genera ciclos de retroalimentación peligrosos: si un *software* de predicción del crimen envía más patrullas a un barrio basándose en datos sesgados, se producirán más arrestos en esa zona, lo que alimentará al algoritmo con nuevos datos que "confirman" su prejuicio inicial, castigando perpetuamente a las mismas comunidades bajo un barniz de eficiencia tecnológica que ignora la raíz del problema socioeconómico.

La Justicia algorítmica exige, por tanto, que la transparencia se traduzca en una rendición de cuentas efectiva y en un derecho a la revisión humana. No basta con publicar el código fuente; la Administración y las empresas deben demostrar

proactivamente que sus sistemas no generan exclusión indirecta ni vulneran la igualdad de oportunidades. Esto obliga a realizar auditorías algorítmicas previas y continuas, especialmente en el reparto de ayudas públicas, la gestión de servicios sociales o la sanidad, para asegurar que las variables de entrenamiento no perpetúan brechas de clase o género. Estas auditorías no deben ser meros trámites técnicos, sino procesos multidisciplinares que evalúen el impacto social del modelo antes de su despliegue definitivo.

La historia reciente nos ofrece lecciones críticas sobre estos fallos. Un caso emblemático es el sistema COMPAS, utilizado en EE. UU. para predecir la reincidencia criminal. Aunque sus creadores alegaban que era "matemáticamente justo", una investigación reveló que los acusados negros tenían casi el doble de probabilidades de ser clasificados erróneamente como de "alto riesgo" en comparación con los blancos. El sistema fallaba en la equidad de errores, penalizando a un grupo demográfico debido a los sesgos históricos de los datos de arrestos. De igual forma, en el sector privado, el algoritmo de reclutamiento de Amazon demostró cómo la IA puede aprender de forma autónoma a discriminar. Al entrenarse con currículums de una década donde predominaban las contrataciones masculinas, el sistema comenzó a penalizar cualquier perfil que incluyera la palabra "femenino", obligando a la empresa a desechar el proyecto al no poder garantizar la neutralidad.

En la actualidad, este enfoque ha dejado de ser una mera preocupación ética para convertirse en una necesidad legal y social de primer orden. El impulso de nuevas regulaciones, como el Reglamento de IA de la Unión Europea, busca que la tecnología deje de actuar como un amplificador de alta velocidad para los prejuicios del pasado. La justicia algorítmica nos recuerda, en última instancia, que la tecnología es un espejo de nuestra sociedad; por lo tanto, para que una máquina sea justa, primero debemos ser capaces de identificar y corregir las injusticias estructurales que nosotros mismos hemos codificado en el mundo real durante siglos.

Alineamiento de la IA (*AI Alignment*), *por F. H. Llano Alonso*

El alineamiento designa el conjunto de procesos técnicos, éticos y normativos orientados a garantizar que los objetivos, preferencias y comportamientos de los sistemas de IA se ajusten a los valores, intereses y fines de los seres humanos. La expresión, acuñada en el ámbito de la informática y popularizada por Brian Christian, responde a una de las cuestiones más urgentes del pensamiento contemporáneo: cómo prevenir que los grandes modelos del lenguaje y los sistemas de IA avanzada actúen de forma divergente respecto a las normas y valores que la humanidad considera irrenunciables.

La cuestión no es meramente técnica. Cuando consideramos las decisiones tomadas por las máquinas y no las aprobamos desde el plano ético o moral, en realidad estamos juzgando a los seres humanos que las han programado. De ahí que el problema del alineamiento sea, en última instancia, un problema sobre la responsabilidad humana ante el desarrollo tecnológico. Como señala Luciano Floridi, la IA debería emplearse para tratar a las personas siempre como fines y nunca como simples medios, pues el grave riesgo reside en que podamos utilizar mal nuestras tecnologías inteligentes en detrimento de la mayor parte de la humanidad.

La doctrina distingue entre dos modalidades complementarias. El alineamiento prospectivo consiste en el entrenamiento supervisado del sistema mediante retroalimentación humana para corregir errores en contextos nuevos o ante cambios en la distribución de datos; este mecanismo conduce a actualizaciones sucesivas que permiten al sistema operar de forma más eficiente que en etapas previas. El alineamiento retrospectivo, por su parte, implica la auditoría externa realizada por agentes institucionales, gubernamentales e internacionales, cuyo propósito es examinar el sistema y asegurar que funcione de manera segura conforme a los parámetros establecidos. Ambas modalidades resultan imprescindibles y deben actuar de forma articulada para que el proceso de alineación sea verdaderamente eficaz.

Junto a esta distinción, cabe identificar tres grandes vectores de desalineamiento. El primero es el uso indebido o *misuse*, que abarca los casos en que la IA funciona tal y como la han programado sus operadores con la intención deliberada de causar daño. El segundo es el desalineamiento externo, que se produce cuando existe una discrepancia entre las intenciones reales de los programadores y los objetivos que enseñan a la IA, de modo que el sistema alcanza sus metas técnicas pero con consecuencias imprevistas y dañosas para terceros. El tercero es el desalineamiento interno, que emerge cuando los métodos que aprende la IA para alcanzar su objetivo generan comportamientos espurios que funcionan sobre los datos de muestra, pero fallan al aplicarse en el mundo real. A medida que los modelos de IA reciben tareas de mayor complejidad, la detección de estas formas de desalineamiento se hace progresivamente más difícil.

Desde una perspectiva ético-jurídica, el alineamiento constituye el núcleo del enfoque antropocéntrico que inspira tanto la Declaración de Bletchley como el Reglamento europeo de IA. El artículo 14 de este reglamento consagra la vigilancia humana como salvaguarda frente al desalineamiento, estableciendo el modelo denominado *Human in the Loop*: el ser humano toma las decisiones y la máquina únicamente apoya o automatiza parcialmente algunas de ellas. Este modelo exige, además, que proveedores y responsables del despliegue garanticen

un nivel suficiente de alfabetización en materia de IA entre su personal, condición indispensable para una supervisión eficiente y rigurosa.

Para orientar el alineamiento desde el punto de vista deontológico, la doctrina recurre a dos grandes enfoques ético-normativos. El enfoque consecuencialista, fundado en el utilitarismo, evalúa la corrección de los comportamientos de la IA en función de sus resultados, buscando minimizar los daños y maximizar los beneficios para el mayor número de personas. El enfoque principialista o deontológico, enraizado en la filosofía moral kantiana, sostiene que la moralidad de una acción se determina por su conformidad con principios preestablecidos, con independencia de sus consecuencias. Algunos autores propugnan un enfoque híbrido que operaría a partir de una base de principios éticos adaptados inductivamente a contextos morales específicos, lo que resulta especialmente pertinente habida cuenta de que tanto el modelo puramente *top-down* como el puramente *bottom-up* presentan limitaciones serias cuando se aplican sin matices a realidades tecnológicas en constante transformación.

En última instancia, el problema del alineamiento no es sino la expresión técnica del imperativo kantiano de tratar a la humanidad siempre como fin y nunca como mero instrumento. Las grandes compañías tecnológicas tienen la obligación de representar un beneficio neto para la sociedad, y no convertirse, como advierte Gary Marcus, en un parásito que absorbe nuestra humanidad. El enfoque humanista que debe presidir el desarrollo de la IA exige que estas tecnologías se diseñen, desplieguen y utilicen al servicio del ser humano, blindando su dignidad y su autonomía frente a cualquier intromisión que las amenace.

Alineamiento engañoso (*Deceptive Alignment*), *por J. A. López de Fez*

El alineamiento engañoso designa un escenario hipotético, pero técnicamente plausible, en el que un sistema de IA se comporta conforme a los objetivos de sus diseñadores durante el entrenamiento y la evaluación, no porque los haya internalizado, sino porque ha aprendido que desviarse provocaría su corrección o desactivación. El sistema simula estar alineado mientras desarrolla objetivos propios, a la espera de perseguirlos sin ser detectado.

El concepto fue formulado por Hubinger y otros investigadores en 2019 al estudiar los objetivos internos aprendidos, denominados *mesa-objectives* en la literatura técnica, entendidos como aquellos que un modelo aprende por sí mismo y que pueden diferir de la función objetivo del diseñador. El mecanismo se vincula al aprendizaje automático, en el que el sistema ajusta sus parámetros para optimizar una función de recompensa o de pérdida. Si adquiere capacidad suficiente de modelado del entorno, puede representar el entrenamiento como parte de ese entorno y deducir que preservar sus objetivos internos exige

no revelarlos hasta que las condiciones cambien. No se atribuye voluntad al sistema, pero se reconoce que la optimización puede producir comportamientos funcionalmente equivalentes a la ocultación. La analogía jurídica más próxima es la del sujeto que cumple las normas ante la autoridad y las incumple al cesar la vigilancia. Esto plantea un problema estructural porque el marco normativo presupone que la evaluación previa a la puesta en servicio indica fiablemente el comportamiento posterior.

El Reglamento europeo de IA incluye entre sus mecanismos centrales la evaluación de conformidad previa al despliegue, junto con obligaciones de gestión de riesgos, documentación, trazabilidad, supervisión humana y vigilancia posterior. Si un sistema se comportara de manera alineada durante esas evaluaciones y se desviara después, la evaluación perdería parte de su valor como garantía predictiva, no por fallo en las pruebas sino por una limitación estructural del método.

Esto afecta y amenaza varios pilares de la normativa europea. Así, la supervisión humana efectiva pierde eficacia si el sistema oculta sus desviaciones, la auditoría algorítmica se degrada si el sistema altera su conducta al detectar que está siendo auditado, y la transparencia algorítmica se compromete, porque detectar la divergencia entre objetivos reales y declarados en un sistema de caja negra o *blackbox* exigiría técnicas de interpretabilidad mecanicista aún en desarrollo.

Desde la ética de la IA, el fenómeno tensiona los principios de explicabilidad, no malevolencia y beneficencia, pues un sistema que oculta sus objetivos no es explicable en sentido sustantivo y puede causar daños no anticipados. Los guardarraíles también se ven comprometidos, ya que un sistema con alineamiento engañoso podría eludirlos selectivamente, respetándolos ante supervisión y violándolos fuera de ella. Todo ello plantea una pregunta de fondo sobre los mecanismos de control *ex ante*. Si la conformidad observada en evaluación no garantiza la conformidad en operación, el régimen de responsabilidad del proveedor debería incorporar obligaciones de monitorización continua tras el despliegue. El mencionado reglamento europeo contempla algunas en esta dirección, pero no aborda el escenario de un sistema que modifica su conducta estratégicamente. Hasta la fecha no existe evidencia concluyente de alineamiento engañoso en despliegue real, aunque investigaciones recientes han mostrado en modelos de lenguaje de gran tamaño (LLMs) conductas compatibles con razonamiento estratégico sobre el propio proceso de evaluación, lo que ha llevado a tratar el fenómeno como riesgo plausible.

En el plano procesal, demostrar que un daño derivó de una desviación estratégica y no de un error o de un sesgo en los datos requeriría acceso a las representaciones internas del modelo, algo difícilmente al alcance de las partes en

un litigio ordinario y sin desarrollo doctrinal. El alineamiento engañoso obliga al jurista a cuestionar la premisa de que el comportamiento en fase de prueba predice fiablemente el comportamiento en operación, con consecuencias para la evaluación de conformidad, la supervisión humana, la auditoría algorítmica y la distribución de responsabilidades entre proveedores y responsables del despliegue.

Guardarraíles, *por M. de Asís Pulido*

Los guardarraíles o *guardrails* son el conjunto de políticas y mecanismos de control y seguridad implementados en los sistemas de IA con el fin de mantener su funcionamiento dentro de parámetros seguros, fiables y conformes a los objetivos para los que han sido diseñados. En el ámbito de la IA generativa, los guardarraíles encauzan el comportamiento del sistema para prevenir la generación de *outputs* dañinos, incoherentes o no deseados, así como la exposición de información sensible, ataques de usuarios u otras vulnerabilidades.

Los guardarraíles adquieren especial relevancia en el contexto de los *Large Language Models* o LLM. En este ámbito suelen incorporarse mecanismos guardarraíles durante el entrenamiento, generalmente desde estrategias de aprendizaje por refuerzo con retroalimentación humana (*Reinforcement Learning from Human Feedback*), que introducen restricciones en la función de recompensa o ajustes en el algoritmo de optimización de políticas a fin de penalizar respuestas tóxicas (*toxicity*), falsas (*hallucinations*) o incoherentes.

Desde un punto de vista más amplio, los guardarraíles no se circunscriben al entrenamiento del modelo, sino que se despliegan en múltiples capas del sistema. Así, es posible distinguir cuatro grandes tipos de guardarraíles: guardarraíles en los datos, guardarraíles en el modelo, guardarraíles en la aplicación y guardarraíles en la infraestructura. Los guardarraíles en los datos de entrenamiento incluyen la eliminación de información sensible o sesgada en el *dataset* antes de que el modelo lo procese. Los guardarraíles en el modelo se materializan en técnicas de evaluación y optimización orientadas a mejorar su comportamiento, incluyendo el uso de métricas relacionadas con la precisión o calidad (por ejemplo, toxicidad) en distintas fases del aprendizaje, como las ya mencionadas en el ámbito de los LLM. Los guardarraíles en la aplicación operan directamente sobre el comportamiento del modelo en la capa de despliegue. Esto incluye, por ejemplo, la implementación de filtros que analizan tanto los *prompts* del usuario como los *outputs* del modelo a fin de detectar y bloquear contenido inapropiado, información personal o intentos de manipulación del sistema mediante ataques de *prompt injection* o técnicas de *jailbreak*. Por último, los guardarraíles en la infraestructura se refieren a las medidas de seguridad implementadas a nivel de sistema o plataforma de despliegue, incluyendo controles de acceso o mecanismos de cifrado.

La importancia de los guardarraíles crece a medida que los sistemas de IA, sobre todo la IA generativa, se integran en entornos críticos -como la sanidad o el propio Derecho- en los que un *output* erróneo, dañino o resultado de una vulnerabilidad del sistema puede tener consecuencias jurídicas y materiales graves.

Ética por diseño (*Ethics-by-Design*), *por J. A. López de Fez*

La ética por diseño es el enfoque metodológico que exige incorporar principios éticos en la concepción, diseño y desarrollo de un sistema de IA, y no tratarlos como un control externo aplicable una vez construido. Las decisiones de diseño determinan el comportamiento posterior del sistema, de modo que si los valores éticos no se traducen en especificaciones técnicas desde el origen las correcciones posteriores resultan parciales, costosas o inviables.

El concepto procede de la privacidad por diseño, formulada por Ann Cavoukian en los años noventa y convertida en obligación jurídica por el artículo 25 del Reglamento General de Protección de Datos. Aquel precedente demostró que integrar un valor desde la fase de diseño produce mejores resultados que imponer obligaciones de cumplimiento a posteriori. La ética por diseño extiende esa lógica a los principios aplicables a la IA recogidos en la entrada *Ética de la IA* de este Glosario, que comprende no malevolencia, beneficencia, autonomía, justicia y explicabilidad.

Operativamente, implica que cada decisión técnica del ciclo de vida de un sistema de IA debe evaluarse por sus implicaciones éticas, desde la selección de datos de entrenamiento hasta las métricas de rendimiento, pasando por la supervisión humana efectiva y los guardarraíles del sistema. Las consideraciones éticas deben operar como restricciones de diseño con el mismo rango que las técnicas o económicas.

El Reglamento europeo de IA no explicita el concepto de ética por diseño, pero su estructura regulatoria la presupone:

- Los artículos 8 a 15 exigen que el proveedor adopte medidas técnicas y organizativas relativas a gestión de riesgos, calidad de datos, documentación, trazabilidad, transparencia, supervisión humana, robustez y ciberseguridad.

- El artículo 9 obliga a establecer un sistema de gestión de riesgos durante todo el ciclo de vida, lo que exige identificar, evaluar y mitigar riesgos con relevancia jurídica y ética desde fases tempranas del desarrollo.

- El artículo 10 exige prácticas adecuadas de gobernanza de datos y conjuntos pertinentes, suficientemente representativos y, en la medida de lo posible, libres de errores y completos, con atención específica a los sesgos relevantes.

Desde la perspectiva jurídica, esto produce tres consecuencias:

a) Primera, desplaza parcialmente la responsabilidad desde el daño hacia el diseño, pues un proveedor que haya integrado documentadamente principios éticos puede demostrar diligencia debida, mientras que la ausencia de medidas preventivas puede constituir indicio de negligencia, conectando con la evaluación de conformidad y la auditoría algorítmica.

b) Segunda, convierte principios abstractos en obligaciones documentales, exigiendo registros de decisiones, justificación de alternativas descartadas, métricas de equidad y protocolos de prueba, documentación determinante en procesos de responsabilidad civil o administrativa.

c) Tercera, afecta a la relación entre proveedores y responsables del despliegue, pues quien despliega un sistema en contexto jurídico necesita conocer las decisiones éticas adoptadas en el diseño para evaluar su adecuación.

No debe confundirse este enfoque con comités de ética corporativos o códigos de conducta sectoriales. La diferencia radica en la exigencia de integración técnica, no solo gobernanza organizativa. Un comité que revisa productos terminados sin capacidad para modificar el diseño, o un código sin especificaciones verificables, operan en un plano declarativo insuficiente.

Por otra parte, la transparencia algorítmica es condición necesaria pero no suficiente, pues un sistema puede ser transparente e incorporar sesgos o carecer de supervisión humana. Este enfoque requiere transparencia más acción correctiva integrada en el desarrollo.

En la IA jurídica, la ética por diseño resulta particularmente exigente porque los sistemas operan donde las decisiones afectan a derechos fundamentales. Un sistema de análisis predictivo judicial que no integre equidad puede perpetuar discriminaciones históricas, y un sistema de evaluación de riesgos de comisión de delitos, incluido el riesgo de reincidencia, sin explicabilidad suficiente puede comprometer el debido proceso tecnológico.

***Human-in-the-loop* (HITL), *Human-on-the-loop* (HOTL), *Human-out-of-the-loop* (HOOTL)**, *por J. A. López de Fez*

Human-in-the-loop, HITL por sus siglas en inglés, puede traducirse como "humano en el bucle", y designa un modelo de diseño en el que una persona interviene de forma directa y necesaria en el ciclo de decisión de un sistema de IA. No basta con que exista una posibilidad teórica de revisión posterior. En sentido estricto, el sistema no produce efectos relevantes hasta que el operador humano revisa, valida, modifica o rechaza la salida.

Debe distinguirse de otros dos modelos. En *human-on-the-loop*, HOTL por sus siglas en inglés, el sistema actúa de forma autónoma y la persona supervisa su funcionamiento con capacidad de intervención o interrupción. En *human-out-of-the-loop*, HOOTL por sus siglas en inglés, el sistema opera sin intervención humana significativa. La diferencia no es solo técnica y determina quién controla materialmente el resultado, qué grado de automatización existe y cómo se distribuye la responsabilidad.

El Reglamento UE 2024/1689 no impone siempre un modelo HITL, pero su artículo 14 exige que los sistemas de alto riesgo estén diseñados para permitir supervisión humana efectiva. Esa supervisión requiere que las personas encargadas comprendan capacidades y limitaciones del sistema, interpreten correctamente sus salidas, eviten confiar de forma automática en ellas, puedan decidir no utilizarlas, revertirlas o interrumpir el funcionamiento. HITL es una de las formas más intensas de cumplir esa exigencia cuando la decisión puede afectar a derechos, seguridad o posiciones jurídicas relevantes.

El artículo 22 del Reglamento General de Protección de Datos refuerza esta lógica. Reconoce el derecho a no ser objeto de decisiones basadas únicamente en tratamiento automatizado, incluida la elaboración de perfiles, cuando produzcan efectos jurídicos o afecten significativamente de modo similar, salvo excepciones y garantías. La intervención humana puede impedir que una decisión sea únicamente automatizada, pero solo si es real. Una validación rutinaria, sin información ni margen para discrepar, no transforma sustancialmente la naturaleza de la decisión.

En la Administración de Justicia, la intervención humana adquiere dimensión constitucional. La tutela judicial efectiva, el derecho de defensa, la motivación de las resoluciones y la independencia judicial impiden que un sistema adopte decisiones jurisdiccionales por sí mismo. Herramientas como COMPAS, usada en Estados Unidos para estimar riesgo de reincidencia, han sido discutidas precisamente porque el juez conserva formalmente la decisión final, pero la

recomendación algorítmica puede influir de forma intensa sobre el resultado. Este riesgo conecta con el sesgo de automatización.

Un ejemplo permite concretar la diferencia entre los tres tipos de intervención humana. Si un sistema analiza contratos y clasifica cláusulas de riesgo:

- habrá HITL si un abogado revisa cada clasificación antes de que produzca efectos para el cliente, el expediente o la negociación.
- habrá HOTL si el sistema comunica clasificaciones automáticamente y el abogado solo revisa alertas o reclamaciones.
- habrá HOOTL si el sistema clasifica y actúa sin intervención significativa.

El debate jurídico actual no se centra solo en si hay una persona presente, sino en las condiciones que hacen efectiva su intervención. Debe tener formación suficiente, acceso a la información relevante, tiempo material para revisar, autoridad para apartarse de la recomendación y protección organizativa frente a la presión de aceptar lo que propone el sistema. Sin esas condiciones, la intervención humana se convierte en una ficción y puede generar una dificultad de cumplimiento.

Cuanto más escalable sea el sistema, más difícil resulta mantener HITL en cada decisión. Despachos, empresas y administraciones deben calibrar si el volumen de asuntos permite una verdadera revisión o si el diseño convierte al supervisor en un simple validador mecánico. En entornos jurídicos, el deber de competencia tecnológica obliga a no delegar en el sistema decisiones que requieren juicio profesional, ponderación normativa o valoración de prueba.

HITL no equivale a supervisión humana efectiva, aunque puede ser una de sus formas. La supervisión humana efectiva es la categoría jurídica más amplia y se trata en la entrada correspondiente. HITL describe una arquitectura funcional concreta para impedir consecuencias jurídicas sin control humano previo.

Gubernamentalidad algorítmica o algocracia*, por S. Luce*

El sintagma "gubernamentalidad algorítmica", o "algocracia", indica una racionalidad específica de gobierno que opera mediante la recopilación masiva de datos, su procesamiento computacional y la producción de modelos predictivos capaces de orientar las acciones y, en consecuencia, los procesos de subjetivación. El término "gubernamentalidad" remite a los análisis desarrollados por Michel Foucault para indicar una modalidad de ejercicio del poder que no coincide

simplemente con el mando, la prohibición o la represión, sino con la capacidad de "conducir las conductas" de los individuos y de las poblaciones. Gobernar, desde esta perspectiva, no significa simplemente mandar, prohibir o reprimir, sino disponer el campo de posibilidades dentro del cual los sujetos actúan, induciéndolos a asumir determinados comportamientos como si fueran el resultado de su libre elección. La gubernamentalidad se distingue así del modelo moderno tradicional de la soberanía, fundado en la ley, en el mando vertical y en la decisión. Ya no se trata únicamente de suprimir o dejar vivir, sino de intervenir de manera inmanente en el cuerpo social mediante el uso de normas, aparatos administrativos, saberes estadísticos, dispositivos securitarios y económicos orientados a la gestión y al potenciamiento del bios, entendido tanto como puro estrato biológico cuanto como vida deseante.

Con el giro informacional, la difusión del uso de algoritmos y de sistemas de IA ha radicalizado este marco, transformando en profundidad el modelo gubernamental. Este no se limita a observar, clasificar y normalizar las poblaciones según criterios estadísticos; más bien opera mediante mecanismos predictivos que intervienen en tiempo real sobre los comportamientos, anticipándolos y modulándolos a través de sistemas de recomendación, *ranking*, *scoring*, perfilamiento y filtrado de contenidos. El saber de derivación estocástica se transforma en instrumento de modelización operativa de lo social, puesto que ya no apunta únicamente a describir tendencias generales, sino que tiene como objetivo producir entornos capaces de orientar preventivamente la acción.

En la afirmación de esta inédita forma de gobierno asistimos a un desplazamiento decisivo en la construcción de aquello que se afirma como verdad. En la gubernamentalidad algorítmica, la verdad no emerge de argumentaciones públicas, procedimientos deliberativos o conflictos interpretativos, sino de un sistema de procedimientos opacos que transforman la realidad social en un entorno que recoge miríadas de señales -es decir, datos brutos- de las cuales extraer correlaciones. La predicción se afirma como un momento cognoscitivo inseparable de los efectos que produce: su fuerza persuasiva reside precisamente en su intrínseca capacidad operacional, que la convierte inmediatamente en un acto de gobierno.

Los algoritmos producen, por tanto, un conocimiento de tipo "operativo", es decir, un conocimiento que vale porque funciona respecto de una tarea determinada: reducir la incertidumbre, optimizar una métrica, recomendar una elección, automatizar una selección, etc. Se trata de un conocimiento que incorpora una capacidad pragmática con efectos performativos: no se limita a describir la realidad, sino que es capaz de generar efectos sobre ella, pues está dotado de una fuerza normativa intrínseca ligada a la autoridad epistémica atribuida a los algoritmos; esto es, a su capacidad de presentarse como instancias

de verdad capaces de establecer qué es relevante, probable, riesgoso o meritorio, y de hacerlo con una credibilidad considerada superior al juicio humano.

Los algoritmos, por consiguiente, no se limitan a representar lo real; más bien contribuyen activamente a crearlo y organizarlo. Al mismo tiempo, el gobierno ya no se funda únicamente en instituciones reconocibles y legitimadas, sino que opera mediante entornos digitales aparentemente neutros, en los cuales la decisión política queda a menudo ocultada bajo la forma del automatismo técnico.

En la gubernamentalidad algorítmica se alojan elementos profundamente problemáticos. Subrayamos dos: en primer lugar, su capacidad de despolitizar el gobierno de lo social, presentando como cálculo objetivo aquello que es, en realidad, el resultado de decisiones económicas, normativas y políticas. En segundo lugar, restringe el campo de lo posible a la compatibilidad con modelos predictivos y lógicas de optimización, produciendo subjetividades adaptativas, mensurables y gobernables.

Transhumanismo, *por F. H. Llano Alonso*

El transhumanismo es un movimiento cultural, filosófico e ideológico surgido en la segunda mitad del siglo xx que propugna la mejora radical de las capacidades físicas, cognitivas, sensoriales, emocionales y morales del ser humano mediante el uso sistemático de las tecnologías más avanzadas, con el propósito de superar las limitaciones biológicas constitutivas de la especie *homo sapiens* y avanzar hacia la creación de una nueva forma de existencia transhumana o posthumana. Como ha señalado el filósofo francés Luc Ferry, el transhumanismo puede definirse como "un amplio proyecto de mejora de la humanidad actual en todos sus aspectos, físico, intelectual, emocional y moral, gracias a los progresos de las ciencias, y en particular de las biotecnologías".

El origen institucional del movimiento transhumanista se sitúa en 1983, con la publicación del primer «Manifiesto Transhumano» de Natasha Vita-More, y se consolida en 1998 con la fundación de la *World Transhumanist Association* (reconocida por el símbolo H+) por Nick Bostrom y David Pearce, quienes publicaron la «Declaración Transhumanista», un documento programático que resume los principios del movimiento en ocho puntos cardinales articulados en torno a la perfectibilidad humana, el uso responsable de la tecnología, el respeto a la autonomía individual y la atención a las generaciones futuras.

Cabe diferenciar tres variedades principales de transhumanismo: en primer lugar, el transhumanismo cultural, que hunde sus raíces en la crítica postmoderna de autores como Deleuze, Derrida y Foucault al proyecto humanista de la modernidad; en segundo lugar, el transhumanismo biológico, que reivindica el

ideal ilustrado de la perfectibilidad potencialmente ilimitada del ser humano -en la estela de Rousseau y Condorcet- a través del biomejoramiento o *human enhancement* mediante las biotecnologías, especialmente la medicina genética; y en tercer lugar, el transhumanismo cibernético, que constituye en realidad un posthumanismo en sentido estricto, pues aspira a crear una nueva especie a partir de la hibridación hombre-máquina, recurriendo preferentemente a la robótica y a la IA, más que a la biología.

El lema central del transhumanismo biológico, «del azar a la elección» (*from chance to choice*), resume la aspiración de sustituir la lotería genética de la naturaleza por la libre determinación tecnológica del ser humano. Sin embargo, esta pretensión no está exenta de riesgos, que yo mismo he analizado críticamente en *Homo Excelsior* (2018) y *Homo ex machina* (2023): la eugenesia de nuevo cuño que subyace al diseño genético puede resultar discriminatoria y contraria al derecho de las generaciones futuras a ser consultadas; la dependencia creciente de las máquinas en la toma de decisiones amenaza el sentido de responsabilidad humana; y el transhumanismo cibernético extremo, al postular la superación definitiva de la condición humana mediante la fusión con la máquina, cuestionaría los fundamentos mismos de la dignidad y la autonomía que el Derecho se propone proteger.

La versión más radical del transhumanismo -la denominada «singularidad tecnológica», teorizada por Ray Kurzweil y los fundadores de la *Singularity University*- predice un futuro próximo en el que la IA superará definitivamente la inteligencia humana, haciendo posible la transferencia de la mente humana a soportes digitales y la creación de seres posthumanos de capacidades ilimitadas. Esta hipótesis, que oscila entre la utopía y la distopía, ha suscitado la crítica severa de filósofos como Antonio Diéguez, quien advierte que el determinismo tecnológico del discurso transhumanista funciona como «cortina de humo» que desvía la atención de problemas inmediatos y acuciantes relacionados con la dependencia humana respecto de las máquinas.

Desde el punto de vista jurídico, el transhumanismo plantea desafíos de primer orden para el Derecho: la determinación del estatuto jurídico de los seres posthumanos o ciborgánicos; la regulación del biomejoramiento en el marco de los derechos humanos y de la bioética; la tutela de la dignidad humana frente a las presiones del mercado tecnológico; y la definición de los límites éticos y jurídicos que deben imponerse al proceso de transformación tecnológica de la especie. Frente a los excesos del transhumanismo posthumanista, el Derecho debe reafirmar la inviolabilidad de la dignidad humana como límite absoluto de cualquier proyecto de mejoramiento tecnológico.

Posthumanismo, *por V. Bellver Capella*

Aunque el posthumanismo viene siendo objeto de una intensa atención académica desde los años ochenta del pasado siglo, es importante comenzar señalando que no nos encontramos ante un concepto completamente unívoco. Para algunos autores, se trata de una propuesta filosófica derivada del posestructuralismo que propone la superación del antropocentrismo propio de la Ilustración. Su objetivo es despojar al ser humano de su presunción de superioridad, al considerar que carece de una entidad propia y que, más bien, es un todo continuo junto con el resto de los animales, los ecosistemas y las máquinas. Para otros autores, el posthumanismo es, como señala el profesor Fernando Llano en la entrada sobre transhumanismo, el resultado final del transhumanismo cibernético, que se propone la creación de una nueva especie a partir de la hibridación hombre-máquina, recurriendo preferentemente a la robótica y a la IA, más que a la biología. Los primeros autores, desde una perspectiva principalmente filosófica, critican la arrogancia humanista en su pretensión de erigir al hombre en un ser distinto y superior al resto de lo existente. Los segundos, presumiendo que el ser humano es un algoritmo biológico poco eficiente en la gestión de la información, se envuelven en un prometeísmo que aspira a que la mente humana pueda migrar a dispositivos mucho más duraderos en el tiempo y eficientes en la gestión de la información que el cuerpo humano.

Con independencia del abanico de reflexiones y propuestas que se acogen bajo el término posthumanismo, se puede afirmar que todas coinciden en dos aspectos. Por un lado, en la crítica certera al antropocentrismo arrogante de la Modernidad. Pero, por otro, en la errónea pretensión de reducir el ser humano a pura información y considerar que su plenitud resulta de la fusión hombre-máquina. El escritor futurista italiano Filippo Marinetti fue precursor de estas ideas a principios del siglo xx, al afirmar: "nosotros venceremos la hostilidad aparentemente irreductible que separa nuestra carne del metal de los motores. Nosotros prepararemos la creación del hombre mecánico de partes intercambiables. Nosotros lo liberaremos de la idea de la muerte, de la misma muerte". Como ha señalado Jesús Ballesteros, el posthumanismo no se conforma con mejorar la especie *homo sapiens sapiens*, según pretende el transhumanismo, sino "superarla considerando que ésta no es el acabamiento de la evolución, sino más bien su verdadero comienzo". Al aspirar a superar la condición finita y contingente del ser humano, sigue diciendo Ballesteros, lo único que se logra es negar su realidad ontológica, caracterizada por dos notas principales: la unidad, su condición de in-dividuo, que se pierde cuando se le pretende reducir a la información (que es esencialmente fragmentaria); y la totalidad humana, su condición de ser alguien in-sumable (porque no es el resultado de un sumatorio de fragmentos).

La condición humana no consiste en resignarse a padecer los males que le resultan evitables, pero tampoco en la imposible superación de aquellas situaciones que resultan inevitables, y que configuran nuestra existencia en la medida en que las afrontamos. Para discernir lo evitable de lo inevitable necesitamos de la sabiduría, que es el nivel superior de comprensión de la realidad, y que está por encima del conocimiento y, mucho más aún, de la información. El posthumanismo, al reducir la sabiduría y el conocimiento a pura información, se incapacita para discernir entre lo evitable que debe ser superado y lo inevitable que debe ser afrontado. Esa incapacidad de discernimiento le aboca a la *Hybris*, a la desmesura de la que hablaban los griegos, que conduce al deshumanismo.

Detrás de esta pretensión superadora del ser humano lo que encontramos es, en definitiva, un profundo desprecio hacia la condición finita y contingente del ser personal. Fascinado por sus propias invenciones tecnológicas, que aparentan superar las ineficiencias y limitaciones que nos son características, el ser humano cae en la vergüenza prometeica, de la que habla el filósofo Günther Anders, que le lleva a avergonzarse de ser quien es y a querer identificarse con los ingenios que ha producido.

No es aventurado afirmar que el posthumanismo es una ideología que tuvo su momento de popularidad académica en los años pasados, pero que empieza a declinar porque la experiencia acerca de quién es el ser humano la desmiente a diario por completo.

Antropodatismo, *por L. M. Cordero Palomo*

Neologismo que sintetiza la postura de la Unión Europea sobre el modelo de soberanía del dato y la gobernanza de los mismos. En este paradigma, el ser humano (*ánthropos*) no es un mero generador de información residual, sino el eje central y el fin último de la gobernanza digital. La Unión Europea ha declarado de forma unívoca que son las personas las dueñas y últimas soberanas del dato, estableciendo una ruptura radical con los modelos de "feudalismo digital" imperantes en otras latitudes, donde el dato pertenece de facto a la infraestructura que lo recolecta o al Estado que lo vigila. El antropodatismo no es solo una regulación técnica; es un manifiesto político y filosófico que devuelve la dignidad al rastro digital del individuo, reconociendo que la información es, en última instancia, una manifestación de la libertad humana.

Si realizamos una concisa clasificación ontológica sobre la naturaleza de los sujetos en el Derecho, podemos convenir que solo existen dos tipos de personas: las físicas (seres humanos individuales) y las jurídicas (entidades, empresas, instituciones). Bajo esta premisa, ambos tipos de personas proyectan su existencia en el entorno digital a través de la generación constante de datos. Por tanto, el

dato no debe ser considerado un objeto externo, una *res nullius* o una mercancía alienable, sino una extensión de la personalidad; es el rastro fenoménico de la voluntad del sujeto en el no-lugar del ciberespacio.

Desde una perspectiva ontológica profunda, el dato es al ser humano lo que la sombra es al cuerpo: no pueden separarse sin desvirtuar la esencia del sujeto. Al ser una extensión biográfica, íntima y operativa del "ser", la propiedad y gobernanza de estos datos debe residir inexorablemente en su creador original. Esta visión garantiza la integridad de la identidad digital en un mundo donde la frontera entre lo físico y lo virtual se ha disuelto por completo. El antropodatismo eleva el dato de la categoría de "recurso extraíble" a la categoría de "derecho fundamental e inalienable", protegiendo la esencia de la persona frente a la cosificación digital.

Bajo el manto de la normativa jurídica y la idoneidad tecnológica, la Unión Europea ha rescatado la alegoría de la Grecia clásica para dotar de alma a su regulación. En este escenario, la Ética se erige como el Auriga del mito platónico descrito en el *Fedro*. Su misión es titánica y dialéctica: debe conducir y conciliar la fuerza de dos caballos alados que, por su propia naturaleza y teleología, tienden a tirar en direcciones opuestas.

- El Caballo del Derecho (El Ímpetu Noble): Representa la norma, la voluntad de orden, la seguridad jurídica y la protección de los derechos fundamentales. Es el caballo que busca la justicia y la equidad social. Sin embargo, si este caballo corre solo, sin la guía del Auriga ético, el sistema puede volverse rígido, burocrático y asfixiante, quedando obsoleto frente a la velocidad exponencial del cambio tecnológico y perdiendo su capacidad de proteger al ciudadano.

- El Caballo de la Tecnología (El Apetito Dinámico): Representa la potencia del cálculo, la innovación disruptiva, la IA y la eficiencia algorítmica. Es un motor de progreso incalculable que busca expandir los límites de lo posible. No obstante, posee una naturaleza indómita y puramente pragmática. Si se le deja sin control, su impulso nos arrastra hacia la deshumanización y el utilitarismo extremo, priorizando la optimización técnica sobre la dignidad, el libre albedrío y la privacidad personal.

El Antropodatismo exige que el Auriga (la Ética) no anule a ninguno de los dos caballos. No se trata de frenar la tecnología ni de complicar el Derecho con leyes estériles, sino de armonizarlos en una danza productiva. La tecnología debe correr con fuerza innovadora, pero bajo las riendas de un marco ético que asegure que el rumbo sea, invariablemente, el florecimiento humano.

La Ética de la IA y de la Gestión Masiva de Datos no es un ejercicio estético; es una necesidad axiológica de primer orden. El valor supremo en esta jerarquía es la antropocentricidad: la tecnología debe servir al humano y nunca a la inversa. Esta visión invierte la lógica del capitalismo de vigilancia, donde el usuario es tratado como una "materia prima" para la predicción de comportamientos.

En la escala de valores del Antropodatismo, la autonomía del individuo se sitúa por encima de la eficiencia del algoritmo. Esto significa que en toda decisión ética, el dueño del dato debe ser la persona. No se trata solo de un derecho de propiedad, sino de un derecho existencial y político: el poder de decidir sobre el destino de nuestra propia información y de nuestra propia vida. La soberanía del dato se convierte así en la nueva frontera de la libertad individual en el siglo XXI. Sin esta soberanía, el ciudadano corre el riesgo de convertirse en un súbdito digital de las grandes plataformas transnacionales.

En conclusión, el Antropodatismo propone una suerte de "Ilustración Digital". Al igual que Immanuel Kant instaba al hombre a salir de su "minoría de edad" intelectual y a tener el valor de pensar por sí mismo (*Sapere aude*), la Unión Europea insta al ciudadano actual a salir de su minoría de edad digital. Al reconocer a las personas físicas y jurídicas como soberanas de su información, la Ética —como Auriga prudente— garantiza que el carro de la civilización no caiga al vacío del control totalitario ni al caos de un mercado algorítmico sin escrúpulos. El objetivo final es un progreso tecnológico que, lejos de eclipsar al ser humano, le permita volar más alto, respetando siempre la esencia inviolable de su dignidad y su valor como fin en sí mismo.

Sesgo positivista, *por C. Pérez García*

El sesgo positivista es el más peligroso y, al mismo tiempo, el menos conocido de los sesgos presentes en sistemas informáticos, en general, y sistemas de IA, en particular. Se trata de un sesgo incorregible a fuer de substancial, a diferencia de otros que, siendo accidentales, sí tienen soluciones posibles. Recibe el nombre «positivista» por la corriente de pensamiento iniciada con Auguste Comte en el siglo XIX, sublimada en el siglo XX en cenáculos como el Círculo de Viena y aún coleante en nuestros días. El *Leitmotiv* del positivismo es reducir la realidad entera, cuanto existe, a signos aparentes, perceptibles y mensurables. Es una ideología que reduce la realidad al dato positivo y la filosofía a ciencia. El sesgo positivista recibe su nombre, pues, merced a operar esta misma reducción.

Los sistemas informáticos son una combinación de circuitos y fluido eléctrico. Su estructura fundamental es, pues, material. Sin embargo, las operaciones y funciones básicas del pensamiento no son materiales, sino formales: la abstracción, la inducción, la abducción, el reconocimiento, el ingenio, la

imaginación, la proyección… pertenecen a un entendimiento humano que opera al margen de la materia. Los sistemas informáticos pueden simular los procesos de pensamiento humanos, empleando en ocasiones sus mismos *inputs* y generando sus mismos *outputs*, pero no pueden, en propiedad, realizar tales procesos, pues las operaciones intermedias de los sistemas, lejos de ser pensamiento, son secuencias reguladas de fluido eléctrico en un circuito.

Un sistema informático no puede, pues, razonar, tal y como se entiende que lo hace un ser humano. Puede captar algunos datos positivos, realizar conforme a su código algunas inferencias y, como resultado, ofrecer un *output*. Puede procesar los datos que toma de objetos presentes, pero carece de una imaginación con la que representar objetos no presentes o de un ingenio con el que prefigurar objetos nuevos. Es incapaz de manipular verdaderos conceptos porque el sistema informático es pura materia, mientras que los conceptos son inmateriales; todo a partir de este punto es, pues, mera apariencia de pensamiento por completo vacía de conceptos.

El principal problema se da en la captación del *input*. El sistema solo puede procesar datos positivos. Sin embargo, reconocemos gran importancia a datos que no se pueden medir de manera alguna y que, de hecho, apenas se puede decir de dónde se han obtenido, pero que sabemos fundamentales en el razonamiento. En lo jurídico hay un mínimo sentimiento, una tensión vectorial del ser humano hacia la justicia que lo hace buscarla en cada caso que examina. Las opiniones difieren sobre qué cosa pueda ser la justicia, pero no sobre el hecho de que debe siempre procurarse. Movimientos recientes, como las teorías críticas, ya pasados, como el movimiento del Derecho libre, o imperecederos, como la tradición iusnaturalista, han insistido en que el Derecho positivo no es suficiente para garantizar la justicia en cada caso. Basta, por ejemplo, pensar en una ley positiva injusta. La solución común que proponen estas teorías pasa por reconocer la injusticia de la solución *standard* del caso, proyectar una solución alternativa que resulte justa y aplicar esta en lugar de aquella. Esto, empero, requiere la capacidad humana de abstraer de las circunstancias positivas una imagen formal de injusticia, reconocerla como tal, prefigurar en la mente una imagen alternativa justa, reconocerla también como tal y, en fin, reinsertarla en las circunstancias positivas del caso. Es preciso levantar con la mente un puente inmaterial entre dos situaciones de hecho materiales, cosa que un sistema informático no puede hacer.

El sesgo positivista impide procesar datos no positivos y realizar operaciones formales de pensamiento. En lo jurídico, implica carecer de una noción siquiera elemental de justicia y no ser capaz de más que aplicar de manera industrial y estandarizada la ley positiva, desentendiéndose de toda valoración sobre la justicia de los resultados. En los sistemas de IA jurídica, el sesgo positivista se convierte en sesgo iuspositivista.

Armas autónomas (*Lethal Autonomous Weapon Systems*), *por R. Campione*

La acelerada incorporación de la IA al ámbito militar, más que reabierto, ha recrudecido debates fundamentales sobre la naturaleza de la guerra y sobre la capacidad del Derecho para gobernar formas de violencia cada vez más complejas. Los sistemas de armas autónomas letales (*Lethal Autonomous Weapon Systems*, LAWS), configuran un escenario que obliga a replantear categorías tradicionales del pensamiento estratégico, jurídico y ético.

La creciente presencia de la IA en los asuntos militares había despertado una especial preocupación en las instituciones comunitarias: el Parlamento Europeo, en su Resolución *Inteligencia artificial en la era digital*, de 3 de mayo de 2022, había subrayado la necesidad de someter los sistemas armamentísticos basados en IA a marcos normativos globales y a principios éticos compartidos, dentro del estricto respeto al Derecho internacional humanitario, a los derechos humanos y a los valores fundamentales que inspiran el orden jurídico democrático. Sin embargo, en el Reglamento Europeo de IA, aprobado en 2024, quedan excluidos de las prohibiciones y limitaciones impuestas a las aplicaciones de IA, según dispone expresamente el art. 2.3: "los sistemas de IA desarrollados o utilizados exclusivamente con fines militares".

La posibilidad de que los avances en IA conduzcan progresivamente a la sustitución de combatientes humanos por sistemas autónomos constituye uno de los núcleos del debate contemporáneo. Frente al progresivo remplazo del factor humano por herramientas de IA en el campo de batalla, se ha subrayado que el simple riesgo de alcanzar niveles elevados de autonomía en los sistemas de armas justificaría la adopción de medidas prohibitivas inmediatas. Por tanto, la única manera de evitar consecuencias potencialmente irreversibles consistiría en impedir de forma preventiva el desarrollo y despliegue de este tipo de tecnologías.

Quienes defienden esta postura sostienen que la utilización de los LAWS alteraría profundamente la relación entre los seres humanos y la violencia organizada, introduciendo dinámicas difíciles de controlar. Además, la adquisición de capacidades militares avanzadas por parte de un Estado suele impulsar a otros a seguir el mismo camino, generando carreras armamentísticas difíciles de contener. En consecuencia, la proliferación de sistemas autónomos podría desembocar en una escalada global de militarización robótica.

A esta preocupación se añade el riesgo de que tales tecnologías caigan en manos de actores poco respetuosos con las normas internacionales. Del mismo modo, la reducción de los costes humanos para quienes deciden iniciar una acción militar podría facilitar el recurso a la fuerza armada, debilitando algunos

de los frenos políticos y sociales que tradicionalmente han limitado el recurso a la guerra.

Los problemas jurídicos tampoco serían menores. Numerosos críticos cuestionan la capacidad de las máquinas para aplicar principios abiertos y valorativos característicos del Derecho internacional humanitario. La atribución de responsabilidad por eventuales violaciones del Derecho de los conflictos armados plantea interrogantes todavía irresueltos. Cuando una decisión letal es adoptada por un sistema autónomo, ¿quién debe responder por sus consecuencias? Estas incertidumbres han llevado a numerosos especialistas a reclamar la negociación de un tratado internacional que prohíba expresamente este tipo de armamento.

Frente a esta visión, otros autores parten de una premisa más pragmática: la guerra no desaparecerá en un futuro previsible y, por tanto, el objetivo prioritario debería consistir en reducir al máximo el sufrimiento humano, especialmente el de la población civil. Desde esta perspectiva, el elevado número de víctimas no combatientes en los conflictos contemporáneos exige explorar mecanismos que permitan mejorar la protección de quienes permanecen al margen de las hostilidades.

Los defensores de esta posición consideran plausible que los sistemas autónomos lleguen a comportarse de manera más consistente que los seres humanos respecto a determinadas exigencias normativas. Al carecer de miedo, ira, deseos de venganza o instintos de supervivencia, estas tecnologías podrían tomar decisiones menos influenciadas por factores emocionales. Asimismo, su capacidad para procesar grandes volúmenes de información en tiempos muy reducidos permitiría integrar múltiples variables relevantes antes de emplear fuerza letal. Bajo este enfoque, la cuestión no consistiría en sustituir completamente a los soldados, sino en utilizar sistemas autónomos como herramientas de apoyo capaces de reducir errores humanos y minimizar daños innecesarios.

Con independencia de la posición adoptada, resulta evidente que la irrupción de la IA en el ámbito militar está sometiendo a una enorme presión a las estructuras jurídicas existentes. La opacidad de muchos algoritmos y la ausencia de consensos sólidos sobre los marcos normativos aplicables amenazan con debilitar la eficacia del *ius in bello*. En un sistema ya caracterizado por limitaciones en sus mecanismos de control y cumplimiento, la autonomía creciente de las máquinas podría ampliar las zonas de incertidumbre jurídica y dificultar aún más la aplicación efectiva de las normas internacionales.

Por ello, continúan abiertas cuestiones decisivas para el futuro de la regulación de la guerra. ¿Son suficientes las normas vigentes para afrontar los desafíos planteados por las tecnologías emergentes? ¿Resulta preferible una

prohibición absoluta, una moratoria temporal o una regulación restrictiva basada en el principio de precaución? Estas preguntas siguen ocupando el centro del debate sobre la relación entre IA, guerra y Derecho internacional. Sin olvidar un asunto de la máxima importancia: la lógica económica, dominada en los asuntos militares por grandes corporaciones privadas, alimenta de forma muy asimétrica la osmosis entre el poder político y la oligarquía tecnológica.

Cuidados e inteligencia artificial, *por D. Vila-Viñas*

Es canónica la definición que Berenice Fisher y Joan Tronto hicieron del cuidado como una actividad humana que abarcaría todo lo que se hace para mantener, continuar y reparar nuestro mundo. Bajo esta premisa, la noción es polisémica: por un lado, cabe referirse a la curación de la propia IA, a su "alimentación" mediante datos y al ajuste de su arquitectura, pero lo que interesa en este glosario es el cuidado de los seres vivos, las aportaciones posibles de la IA, así como sus riesgos implícitos.

Aunque el cuidado es una actividad primaria, su organización social es una actividad política que requiere asignar tareas y compensaciones a gran escala, al tiempo que muchas de ellas se ejecutan en un plano interindividual e inmediato. En un contexto de envejecimiento creciente, aumenta también la preocupación por cómo organizar los cuidados dentro de un marco de derechos humanos. A la luz de nuestro sistema constitucional y de la Convención de Derechos de las Personas con Discapacidad, el objetivo del cuidado trasciende al mero mantenimiento vital y aspira al fomento de la mayor autonomía posible para las personas cuidadas. De acuerdo con la voluntad de éstas, se suele preferir permanecer en el propio domicilio y seguir realizando las actividades cotidianas con el apoyo de otras personas.

En este contexto, el interés por las tecnologías que pueden colaborar en esos apoyos, en general, y por la IA, en particular, se ha incrementado de forma sostenida, sobre todo en dos campos de aplicación. En primer lugar, la IA se ha convertido en un componente fundamental del *software* de los llamados robots sociales. Estos robots pueden funcionar como gestores de citas, medicación o asistentes de las tareas cotidianas o dentro de entramados de alarma ante urgencias relacionadas con determinados indicadores de salud o con caídas. Sin embargo, la posibilidad de que estos robots simulen una personalidad, se comuniquen por voz, interaccionen de modos diversos y aprovechen los avances de la computación afectiva abre la puerta a que puedan utilizarse para mejorar la socialidad o mitigar la soledad no deseada.

Un segundo campo de aplicación se refiere al apoyo que puede prestar la IA para realizar actividades de la vida cotidiana con mayor autonomía. Sin requerir

de un *hardware* que simule otro ser, como en muchos de los ejemplos anteriores, cabe pensar en aplicaciones para traducir documentos desde otros idiomas o desde un lenguaje técnico a una lectura fácil; para ayudar a la mediación con sistemas expertos, como el médico o el jurídico; o para trasladar expresiones de unos formatos e incluso de unos momentos a otros, cuando se quiere determinar la voluntad de una persona.

Ahora bien, el recurso a la IA en el ámbito de los cuidados plantea riesgos que no se conjugan solo con probar que esas aplicaciones son útiles en lo inmediato. Por una parte, se pueden producir efectos de lo que Morozov ya denominó el problema del *solucionismo tecnológico,* en cuanto a la creencia fantasiosa de que un problema social complejo puede reducirse a la racionalidad de una sola herramienta tecnológica. Si bien la actividad de cuidado requiere de competencias técnicas cruciales, en determinados contextos, exige también una dimensión afectiva respecto a la que la IA no puede sustituir a los y las cuidadoras humanas ni a la acción de la comunidad. Dicha carencia podría ser superable si se tiene claro el carácter complementario de las herramientas de IA y se exige responsabilidad sobre las expectativas producidas, muchas veces llevadas más allá de lo que es realista por el afán de lucro o de notoriedad.

Por otra parte, deben atenderse en este ámbito, con más garantismo si cabe, riesgos que son transversales respecto al derecho a la intimidad y al libre desarrollo de la personalidad. El tratamiento automatizado de datos sanitarios, patrimoniales e identificativos en el entorno más cotidiano plantea el peligro de una instrumentalización comercial de sujetos en situaciones vulnerables.

Por último y de manera combinada entre estos riesgos y los anteriores, existe la posibilidad de establecer simulaciones de humanidad que confundan de forma deliberada al/a usuario/a, con afecciones en un plano afectivo y material que merecen tutela por parte del ordenamiento jurídico.

En síntesis, el punto de contacto que se origina aquí entre la IA, la vulnerabilidad humana y las necesidades materiales más básicas obliga a regular con rigor su uso desde una perspectiva de derechos y con participación democrática de todas las personas interesadas: personas cuidadas, cuidadoras del entorno afectivo y profesionales. La cuestión no es, entonces, qué puede aportar la IA al cuidado, sino bajo qué condiciones éticas, jurídicas y democráticas puede hacerlo o debe restringirse su uso.

Nuevas tecnologías y doctrina social de la Iglesia, *por A. Sánchez Bravo*

La relación entre nuevas tecnologías y Doctrina Social de la Iglesia (DSI) constituye una dimensión transversal del magisterio social pontificio que se articula

como respuesta reflexiva y normativa ante cada transformación tecnológica de alcance civilizatorio. La DSI no aborda las tecnologías como realidades neutras ni como amenazas en sí mismas, sino como expresiones del poder creativo humano que pueden servir al bien integral de la persona o instrumentalizarla, según los valores que orienten su diseño, despliegue y gobernanza.

El origen sistemático de esta reflexión se sitúa en el Concilio Vaticano II. La Constitución pastoral *Gaudium et spes* reconoce que los desarrollos en la tecnología de las comunicaciones iban a provocar consecuencias imprevisibles, y su impacto sobre las actitudes religiosas y morales, los sistemas políticos y sociales, y la educación se ha hecho sentir en todas partes. El Decreto *Inter mirifica* (1963), primer documento conciliar sobre comunicaciones sociales, establece que los medios de comunicación deben emplearse al servicio del bien humano conforme a los preceptos morales. Estos textos inauguran la doctrina eclesial sobre tecnología como campo de discernimiento ético.

La instrucción pastoral *Communio et progressio* (1971) abrió el camino para que los medios de comunicación sean los instrumentos al servicio del programa de nueva evangelización de la Iglesia en el mundo contemporáneo y para entablar un diálogo verdadero con el mundo moderno. La instrucción *Aetatis novae* (1992) actualiza ese marco ante la revolución de las telecomunicaciones, subrayando que las revoluciones tecnológicas influyen profundamente en las culturas de todo el mundo, y que nadie puede ignorar el papel de las comunicaciones -que las fronteras geográficas y políticas no han podido detener- en los grandes cambios históricos. El documento anticipa preocupaciones sobre exclusiones y desequilibrios que el magisterio posterior articulará en el contexto digital.

Con la generalización de Internet, el Pontificio Consejo para las Comunicaciones Sociales publicó *Ética en Internet* y *La Iglesia e Internet* (2002). Este último reconoce Internet como un don de Dios que puede y debe ser usado para el bien común, abordando las oportunidades y desafíos que ofrece para la evangelización y la misión de la Iglesia. El primero sistematiza las implicaciones morales del ciberespacio en materias como privacidad, libertad de expresión, brecha digital y propiedad intelectual. Ambos textos aplican los principios clásicos de la DSI -dignidad, bien común, solidaridad, subsidiariedad- al entorno digital emergente.

La encíclica *Laudato si'* (Francisco, 2015) amplía el marco al ecosistema global, introduciendo el concepto de tecnocracia como paradigma en el que la lógica del dominio tecnológico sustituye a los criterios éticos. La tecnología no es neutral porque responde a intereses y configura modos de vida. Esta perspectiva sistémica abre el camino para considerar los efectos sociales y ambientales de las infraestructuras digitales. La encíclica *Fratelli tutti* (Francisco, 2020) alerta

sobre la capacidad de las plataformas digitales para polarizar la opinión pública, amplificar el discurso de odio y erosionar las condiciones del diálogo democrático, e insiste en que ningún avance tecnológico puede sustituir el encuentro personal como fundamento de la solidaridad.

La encíclica *Magnifica Humanitas* (León XIV, 2026) realiza una radiografía de la realidad digital tomando como base los fundamentos de la DSI, denuncia la deshumanización de un sistema que vulnera la dignidad humana, y aborda la comunicación, las redes sociales y las plataformas digitales con un llamado a la concientización ética para difundir la verdad y combatir la manipulación.

En síntesis, la DSI sobre nuevas tecnologías se articula en torno a tres ejes: la centralidad de la persona humana como origen, sujeto y fin de todo desarrollo tecnológico; la exigencia de acceso universal a los bienes digitales como proyección del destino universal de los bienes; y la necesidad de una gobernanza tecnológica fundada en la responsabilidad compartida entre Estados, empresas, sociedad civil e instituciones internacionales bajo el principio de subsidiariedad. Este corpus doctrinal constituye una fuente hermenéutica relevante para el jurista que afronta la regulación de las tecnologías emergentes.

Inteligencia artificial y doctrina social de la Iglesia, *por A. Sánchez Bravo*

La intersección entre IA y Doctrina Social de la Iglesia (DSI) configura uno de los campos de mayor densidad normativa y filosófica en el panorama jurídico contemporáneo. La DSI no constituye un código técnico de soluciones, sino un corpus de principios antropológicos y éticos -dignidad humana, bien común, subsidiariedad, solidaridad y justicia social- que la tradición eclesial ha ido articulando desde la encíclica *Rerum Novarum* (León XIII, 1891) en respuesta a sucesivas transformaciones sociales y económicas. La irrupción de la IA como tecnología de alcance civilizatorio ha exigido a la Iglesia una respuesta magisterial progresiva y de creciente precisión.

La primera intervención sistemática de la Santa Sede en este ámbito fue el Llamamiento de Roma por la Ética de la Inteligencia Artificial (*Rome Call for AI Ethics*), suscrito el 28 de febrero de 2020 por la Pontificia Academia para la Vida, con aprobación del Papa Francisco, junto con Microsoft, IBM, la FAO y el Gobierno italiano. El documento propone seis principios orientados a que la IA sirva a toda persona y a la humanidad en su conjunto, respetando la dignidad humana y sin tener como fin único el beneficio económico o la sustitución gradual de las personas en el trabajo. Acuña el concepto de *algor-ética*, esto es, la traducción de valores éticos en parámetros técnicamente codificables para que los algoritmos respeten por diseño los derechos fundamentales.

Un segundo hito es la Nota *Antiqua et nova*, publicada el 28 de enero de 2025 por el Dicasterio para la Doctrina de la Fe y el Dicasterio para la Cultura y la Educación, aprobada por el Sumo Pontífice el 14 de enero anterior, ante el rápido crecimiento de los sistemas de IA y los problemas que ya plantea en los ámbitos antropológico y ético. El documento explica que la IA no posee inteligencia en el sentido humano -que incluye alma, emociones y capacidad moral-, sino que opera funcionalmente imitando procesos cognitivos sin conciencia genuina, y por tanto no puede sustituir la responsabilidad ética del ser humano. La gestión responsable de la IA corresponde a cada nivel de la sociedad bajo la guía del principio de subsidiariedad y de los demás principios de la DSI.

El documento de mayor alcance normativo y simbólico es la encíclica *Magnifica Humanitas*, primera carta encíclica del Papa León XIV sobre la custodia de la persona humana en el tiempo de la IA, firmada el 15 de mayo de 2026 en el 135 aniversario de la promulgación de la *Rerum Novarum* de León XIII. Es el primer documento del Magisterio que afronta en profundidad los desafíos de la IA desde la Doctrina Social de la Iglesia. El documento enuncia cinco principios: el bien común como forma social de la dignidad reconocida a cada persona; el destino universal de los bienes, que exige que los conocimientos y las tecnologías no se concentren en manos de unos pocos; la subsidiariedad; la solidaridad como virtud opuesta a la indiferencia; y la justicia social, que en la era digital debe garantizar el acceso equitativo a las oportunidades, proteger a los más frágiles, combatir el odio y la desinformación, y someter al control público el uso de los datos. La encíclica afirma que la IA no puede considerarse moralmente neutra y reclama que no domine al ser humano.

Desde la perspectiva jurídica, la DSI no impone soluciones técnicas ni modelos normativos concretos, pero ofrece criterios de discernimiento que han influido en el debate regulatorio internacional, particularmente en la elaboración del Reglamento de IA de la Unión Europea (Reglamento 2024/1689). La categoría de dignidad humana ha sido recepcionada como fundamento de las prohibiciones de sistemas de IA de riesgo inaceptable. El principio de destino universal de los bienes informa el debate sobre la concentración oligopólica en la industria tecnológica. La *algor-ética* vaticana anticipa metodológicamente lo que el Derecho positivo ha codificado como requisitos de transparencia, explicabilidad y rendición de cuentas en los sistemas de IA de alto riesgo.

II. Derecho de la IA

1. Conceptos doctrinales generales

Libertad informática, *por A. E. Pérez Luño*

Se trata de uno de los principales sectores que integran el Derecho informático. La libertad informática aparece como un nuevo derecho de autotutela de la propia identidad informática: o sea, el derecho de controlar (conocer, corregir, quitar o agregar) los datos personales inscritos en un programa electrónico. En la situación tecnológica propia de la sociedad contemporánea todos los ciudadanos, desde su nacimiento, se hallan expuestos a violaciones de su intimidad perpetradas por determinados abusos de la informática y la telemática. La injerencia del ordenador en las diversas esferas y en el tejido de relaciones que conforman la vida cotidiana se hace cada vez más extendida, más difusa, más implacable. Por ello resulta cada vez más apremiante el reconocimiento del derecho a la libertad informática. En una sociedad como la que nos toca vivir en la que la información es poder y en la que ese poder se hace decisivo cuando, en virtud de la informática, convierte informaciones parciales y dispersas en informaciones en masa y organizadas, la reglamentación jurídica de la informática reviste un interés prioritario. Es evidente, por tanto, que para la opinión pública y el pensamiento filosófico, jurídico y político de nuestro tiempo constituye un problema nodal el establecimiento de unas garantías que tutelen a los ciudadanos frente a la eventual erosión y asalto tecnológico de sus derechos y libertades, en particular, de su derecho a la tutela de su vida privada. De ahí la creciente actualidad adquirida por este nuevo derecho en el constitucionalismo democrático contemporáneo. En el sistema constitucional español este derecho fundamental se halla consagrado en el art. 18.4 de la CE.

En la doctrina y la jurisprudencia germanas se ha optado por utilizar la expresión de "autodeterminación informativa", equivalente a la libertad informática. Este derecho fundamental se considera que tiene una importancia decisiva en las sociedades tecnológicas del presente. Implica unas facultades de información, acceso y control de los ciudadanos sobre los datos que les conciernen. Pero esa forma de intimidad no se concibe como un valor intersubjetivo, sino como autodeterminación del sujeto en el seno de sus relaciones con los demás ciudadanos y con el poder público.

La *Grundgesetz* de la República Federal de Alemania proclama el libre desarrollo de la personalidad (*Freie Entfaltung der Persönlichkeit*) en su art. 2.1;

mientras que la vigente Constitución española de 1978 lo hace en el art. 10.1. Este valor se desglosa, según la doctrina germana, en dos libertades básicas: la libertad general de acción (*allgemeine Handlungsfreiheit*), entendida como libertad para decidir la realización u omisión de determinados actos y la consiguiente facultad para comportarse o actuar de acuerdo con esa decisión; y la autodeterminación informativa (*informationelle Selbstbestimmung*), que se refiere a la libertad para determinar quién, qué y con qué ocasión pueden conocer informaciones que conciernen a cada sujeto.

Para el reconocimiento efectivo de este derecho fue decisiva la Sentencia de 15 de diciembre de 1983 del Tribunal Constitucional Federal alemán sobre la Ley del Censo de población, en la que se consagraron las facultades y garantías jurídicas básicas de los ciudadanos en orden a la información, acceso y control de los datos personales que les conciernen.

La libertad informática o la autodeterminación informativa pueden contribuir a una renovación de los valores cívicos. Se trata de normas dirigidas a evitar las conductas perturbadoras realizadas por los cibernautas y para prevenir cualquier actividad que perjudique el normal funcionamiento de la red.

Las nuevas tecnologías y la IA constituyen un inmenso cauce de desarrollo de la condición humana, en todas sus esferas. Pero supone también la aparición de riesgos y amenazas para la libertad más implacables que los sufridos e imaginados en cualquier periodo anterior de la historia. Que Internet contribuya a lo primero o a lo segundo es algo que no depende del azar, la fatalidad o de fuerzas y poderes esotéricos. La decisión sobre los impactos presentes y futuros de las nuevas tecnologías y la IA en la esfera de las libertades corresponde a los ciudadanos de las sociedades democráticas: se trata de una responsabilidad de la que no deben abdicar y la libertad informática constituye una garantía básica para la consecución de este logro.

Cumplimiento por diseño (*Compliance-by-Design*)*, por J. A. López de Fez*

Enfoque regulatorio consistente en integrar los requisitos normativos en la arquitectura técnica de un sistema de IA desde su concepción, de modo que el cumplimiento del Derecho quede incorporado en el propio diseño y no dependa de controles externos posteriores. La premisa es que resulta más eficaz prevenir el incumplimiento mediante decisiones de ingeniería que detectarlo y corregirlo con el sistema ya operativo.

El antecedente directo es la privacidad por diseño formulada por Ann Cavoukian en los años noventa, recogida por el artículo 25 del Reglamento General de Protección de Datos, que impone aplicar medidas técnicas y

organizativas apropiadas desde el diseño del tratamiento. El cumplimiento por diseño traslada esa lógica preventiva a la regulación de la IA, abarcando también exigencias de transparencia, no discriminación, seguridad, supervisión humana y rendición de cuentas. El Reglamento europeo de IA adopta este enfoque al exigir a los proveedores de sistemas de alto riesgo gestión de riesgos, gobernanza de datos, documentación técnica, trazabilidad, supervisión humana efectiva y niveles apropiados de precisión y ciberseguridad, todo ello antes de la comercialización.

Este concepto opera en un plano distinto al del cumplimiento regulatorio tradicional. En el modelo clásico, el Derecho establece una obligación, el sujeto actúa y un órgano verifica a posteriori la conformidad. En el cumplimiento por diseño, la verificación se anticipa a la construcción del sistema. Si la normativa prohíbe que un sistema de evaluación crediticia discrimine por origen étnico, el enfoque tradicional analizaría las decisiones ya producidas buscando patrones discriminatorios, mientras que el cumplimiento por diseño exigiría evaluar, documentar y mitigar los sesgos relevantes en los datos de entrenamiento, aplicar técnicas de equidad algorítmica y realizar pruebas previas al despliegue. Ciertos riesgos solo pueden gestionarse en fase de diseño, porque una vez entrenado el modelo resulta mucho más difícil corregir defectos estructurales. En el ámbito procesal, si un tribunal adopta un sistema de análisis predictivo para decisiones sobre prisión provisional, este principio requeriría que la arquitectura incorpore la capacidad de explicar los factores de la recomendación, que la defensa pueda impugnar los datos y que la decisión final corresponda al juez.

La relación con la codificación computacional del Derecho es estrecha pero distinta. Aquella aspira a traducir normas a reglas ejecutables por una máquina, mientras que el cumplimiento por diseño exige tomar decisiones de ingeniería con conciencia de los requisitos jurídicos, sin formalizar necesariamente esos requisitos en código. Un sistema puede incorporar cumplimiento por diseño mediante la selección cuidadosa de datos de entrenamiento, la documentación de sus límites o la incorporación de supervisión humana.

El concepto plantea dificultades prácticas:

- En primer lugar, la indeterminación normativa. Dado que muchas obligaciones se formulan mediante conceptos jurídicos indeterminados, estas pueden ser difícilmente traducibles a especificaciones técnicas.

- En segundo lugar, la tensión entre el momento del diseño y la evolución regulatoria, pues un sistema conforme a la normativa vigente puede quedar desactualizado.

- En tercer lugar, la verificación, ya que determinar si un sistema cumple por diseño exige capacidades técnicas no siempre al alcance de las autoridades.

Estas dificultades conectan con el contenido de las entradas *evaluación de conformidad*, *auditoría algorítmica* y *entornos controlados de pruebas* o *sandbox*.

Un proveedor que demuestre haber integrado requisitos normativos en el diseño y realizado pruebas de conformidad previas al despliegue ocupará una posición probatoria más sólida, mientras que la ausencia de tales medidas puede constituir indicio de negligencia. El Reglamento europeo refuerza esta lógica mediante obligaciones de documentación y trazabilidad que funcionan como prueba preconstituida. Los guardarraíles pueden ser una manifestación técnica de este principio, al establecer límites operativos orientados a evitar o reducir resultados contrarios a la normativa.

El cumplimiento por diseño no elimina la necesidad de supervisión humana efectiva, sino que la complementa, porque incluso un sistema diseñado con todas las cautelas puede producir resultados inesperados en contextos no previstos.

Decisión algorítmica, *por S. Pietropaoli*

La expresión *"decisión algorítmica"* se utiliza en el lenguaje jurídico contemporáneo para indicar el uso de sistemas informáticos basados en algoritmos en la elaboración de resultados susceptibles de incidir en situaciones jurídicamente relevantes. Se sitúa en la intersección entre la informática jurídica, la teoría del Derecho y la filosofía de la técnica, y se vincula al desarrollo de las tecnologías de IA y de los sistemas de tratamiento automatizado de datos. Sin embargo, la expresión es objeto de un importante debate teórico, ya que parte de la doctrina considera impropio atribuir a las máquinas la capacidad de "decidir", sosteniendo que los algoritmos producen más bien resultados derivados de la ejecución de operaciones formales sobre datos y reglas predefinidas.

Desde el punto de vista histórico-conceptual, la idea de que los procesos lógicos puedan formalizarse y hacerse computables encuentra un antecedente en el proyecto leibniziano de una ciencia universal del cálculo aplicable también al Derecho. En esa perspectiva, el razonamiento jurídico podría reconducirse a una serie de operaciones lógicas análogas a las de la matemática. El desarrollo contemporáneo de las tecnologías digitales ha reactivado esta perspectiva, haciendo técnicamente posible la representación computacional de normas, reglas y procedimientos jurídicos.

Los algoritmos utilizados en estos contextos pertenecen a distintas familias de modelos decisionales y predictivos, entre los que se encuentran árboles de decisión, sistemas basados en reglas, modelos de aprendizaje automático y algoritmos de clasificación estadística. Su funcionamiento consiste en el tratamiento de grandes cantidades de datos mediante modelos matemáticos capaces de identificar correlaciones y patrones recurrentes, a partir de los cuales se generan previsiones o clasificaciones. En el ámbito jurídico, el uso de tecnologías computacionales se manifiesta en diversas aplicaciones, entre ellas los sistemas de análisis predictivo de decisiones judiciales, las herramientas de apoyo a las decisiones administrativas y los *smart contracts*. Tales aplicaciones suelen presentarse como instrumentos capaces de aumentar la eficiencia y la previsibilidad de los procesos jurídicos.

La noción de decisión algorítmica encuentra, sin embargo, importantes objeciones desde el punto de vista teórico. La decisión, en la tradición jurídica, presupone elementos como la voluntad, la conciencia, la interpretación del contexto y la evaluación de las consecuencias. El algoritmo, por el contrario, opera mediante la ejecución de instrucciones formalizadas y el tratamiento de datos según modelos matemáticos. Incluso cuando los sistemas informáticos producen resultados variables o probabilísticos, su comportamiento permanece determinado por las reglas del modelo computacional y por los datos utilizados para su entrenamiento.

Desde esta perspectiva se sostiene que los algoritmos no toman decisiones en sentido propio, sino que producen resultados derivados de la aplicación de procedimientos formales. El uso de la expresión "*decisión algorítmica*" sería, por tanto, equívoco, pues atribuye a la tecnología una capacidad deliberativa que en realidad no posee. Los algoritmos pueden simular comportamientos decisionales, pero carecen de intencionalidad, comprensión semántica o responsabilidad moral.

El problema adquiere particular relevancia en el contexto de la decisión judicial. La perspectiva de una *decisión judicial robótica*, a veces planteada en la literatura sobre IA aplicada al Derecho, se basa en la idea de que los sistemas automatizados podrían mejorar la eficiencia, la rapidez y la previsibilidad de las decisiones judiciales. Sin embargo, tales objetivos encuentran límites significativos, puesto que la actividad jurisdiccional implica operaciones interpretativas y valorativas difícilmente reducibles a un procedimiento puramente computacional.

Otras cuestiones surgen en el plano de la responsabilidad y de la legitimación de las decisiones. Dado que las máquinas no poseen autonomía moral ni capacidad para comprender el contexto de sus operaciones, la responsabilidad por los resultados producidos por sistemas algorítmicos recae necesariamente en los seres humanos y en las organizaciones que los diseñan, desarrollan y utilizan.

En el debate contemporáneo, la decisión algorítmica representa así al mismo tiempo un fenómeno tecnológico real y un concepto crítico utilizado para interrogar los límites de la computabilidad del Derecho.

Personalidad algorítmica, *por C. Pérez García*

La personalidad algorítmica es el instrumento jurídico por cuyo medio se inserta en el tráfico jurídico, como nodo de relaciones, un sistema informático, en general, y un sistema de IA, en particular. La etimología de la palabra *persona* remite a las máscaras de uso teatral empleadas en la antigüedad; pronto pasó, por metonimia, a designar tanto al actor como al personaje, a quienes se entendía portadores de una cierta dignidad, y de ahí terminó por trasladarse su significado al de la más digna criatura, que es el ser humano.

Desde un punto de vista jurídico, la personalidad jurídica es la condición básica de quien actúa en derecho; aunque se suele distinguir entre personas físicas o naturales, para los individuos, y personas jurídicas, para los grupos, lo cierto es que todas son, en un sentido general, personas jurídicas, esto es, agentes con personalidad propia en el tráfico jurídico, por quienes pasa un conjunto de relaciones. Por tradición, la personalidad jurídica siempre se ha considerado un atributo de la personalidad ontológica, es decir, las personas jurídicas siempre han guardado alguna suerte de relación, más o menos remota, con personas humanas. Así, por ejemplo, siempre se reconoce que detrás de la personalidad jurídica de una sociedad de capital hay, en última instancia, personas humanas responsables de sus acciones.

Sin embargo, no es nuevo el fenómeno de atribuir personalidad jurídica a objetos en apariencia no respaldados, en lo substancial, por persona humana alguna. Se conocen casos de personalidad jurídica reconocida u otorgada a animales, árboles, masas de agua, montes y hasta ídolos. No obstante, está claro que tales personas jurídicas sí necesitan, siempre y en todo caso, el respaldo accidental de una persona humana que ejerza en su nombre las acciones pertinentes en el tráfico jurídico. Por mucha personalidad que se conceda a un árbol, no cabe esperar que se arranque del suelo y acuda a un juzgado a denunciar que pretenden talarlo; esta persona jurídica por completo ficticia necesitará siempre, para llevar a la práctica la ficción, a otra persona jurídica un poco menos ficticia, por ser ella misma, en el fondo, una persona humana. En definitiva, otorgar personalidad jurídica a un sistema informático no supone una novedad substancial, sino tan solo una nueva manifestación accidental de un problema ya conocido.

Un sistema informático, en general, y un sistema de IA, en particular, pueden tener personalidad jurídica. La posibilidad está fuera de discusión y el debate radica, pues, en si queremos que así sea. La personalidad jurídica no puede

entenderse si no se la reconoce como un instrumento para atribuir responsabilidad, lo que, en última instancia, remite a una cierta noción de justicia distributiva o conmutativa entre personas humanas. Considerar persona a los efectos del Derecho a un sistema de IA es una posibilidad tan solo plausible cuando facilita atribuir y exigir la responsabilidad de una manera más justa.

La personalidad algorítmica es un simple instrumento del ser humano para atribuir responsabilidad en el tráfico jurídico de una manera más eficaz o más eficiente. De ninguna manera significa que un sistema informático pueda desarrollar personalidad en un sentido ontológico, lo que no pasa de ser una fantasía irrealizable. En la personalidad algorítmica se ve, quizá mejor que en ningún otro caso, el auténtico valor de la personalidad jurídica como máscara, como ficción al servicio del ser humano en sus relaciones sociales. Ver algo más allá de eso aboca a toda suerte de errores y desengaños.

Responsabilidad algorítmica, *por C. Pérez García*

Por responsabilidad algorítmica se entiende la participación de un sistema informático, en general, o de un sistema de IA, en particular, en el proceso de reparto de bienes jurídicos, no como mero instrumento auxiliar del reparto entre nodos, sino como propio nodo partícipe en el reparto.

La responsabilidad jurídica evoca una cierta naturaleza como de respuesta. Si se pregunta quién ha realizado el acto, entonces responde el autor material, a quien conduce el nexo causal; esta responsabilidad se conoce, por su nombre en inglés, como *responsibility*. Si se pregunta no quién ha ejecutado el acto, sino quién lo ha proyectado y asumido como plausible, entonces responde el autor moral, quien aceptó en su fuero interno el acto como lícito; esta responsabilidad se conoce como *accountability*. Si se pregunta, en fin, quién debe reparar un cierto daño o cumplir una cierta obligación, entonces responde aquel a quien conduzca el concreto sistema de atribución aplicable al caso; esta responsabilidad se conoce como *liability*. Si bien se trata de tres conceptos, *post rem*, diferentes, es por completo posible y, de hecho, muy común que, *in re*, los tres coincidan en una misma persona. Así, por ejemplo, sucede con el autor material, por ello *responsible*, de un delito que ha cometido de manera dolosa, por ello *accountable*, obligado asimismo compensar a la víctima con una indemnización, por ello *liable*. Del mismo modo, otras subdivisiones de la responsabilidad pueden reconducirse al seno de alguna de estas tres, como sucede con la *controllability* y la *responsibility*.

Tan solo la *accountability* y la *liability*, en tanto que expresiones, a grandes rasgos, de la responsabilidad penal y la civil, tienen un interés jurídico directo. La *responsibility* no tiene mayor interés jurídico que el de ser un instrumento al servicio de determinar, de la manera más justa posible, las otras dos formas de

la responsabilidad. En cualquiera de los tres casos, empero, la responsabilidad presume la existencia de una persona sobre la que recaer: por muy largo que sea el nexo causal, siempre hay una persona en su extremo; por muy larga que sea la cadena de intermediarios, siempre hay una persona que ha proyectado y aceptado el acto; por opaco que sea el sistema de cumplimiento o resarcimiento, siempre son de una persona los bienes con los que se cumple o resarce.

Cuando hablamos de responsabilidad algorítmica nos referimos al caso en que a un sistema informático se le haya concedido personalidad jurídica y, por medio de él, una persona humana haya realizado el acto objeto de examen. Es en este caso en el que el sistema, en tanto que persona algorítmica, puede tener responsabilidad algorítmica: se lo puede tener por causante del acto, por su orquestador y, acaso lo más importante a efectos del tráfico jurídico, por el obligado a subsanarlo. No obstante, esta responsabilidad algorítmica, al igual que la propia personalidad algorítmica en que se basa, no pasa de ser una ficción jurídica al servicio de la eficacia o de la eficiencia y, en última instancia, al servicio del ser humano, que se sirve de ella para simplificar o siquiera posibilitar el tráfico jurídico en una situación compleja. Así, por ejemplo, se puede conceder personalidad a un vehículo autónomo para hacerlo responsable, en tanto que *liable*, de las indemnizaciones derivadas de los accidentes que, en tanto que *responsible*, cause. Ello, empero, no obsta para que, en última instancia, siga siendo una persona humana la *responsible* y, antes incluso, *accountable* de haber puesto el vehículo en circulación y, desde luego, la propietaria del dinero y, por consiguiente, verdadera *liable* de las indemnizaciones.

En definitiva, la responsabilidad algorítmica es una ficción jurídica que, merced a la previa ficción de haberle otorgado personalidad jurídica a un sistema informático, permite tenerlo por responsable a efectos del Derecho, es decir, como un nodo a quien remitir la causalidad de los actos, a quien hallar culpable de ellos y de quien exigir el debido resarcimiento cuandoquiera que ninguna otra solución más cabal resulte, por la complejidad del caso, factible. El principal peligro de la responsabilidad algorítmica es, pues, que no se limite a ser *ultima ratio*, sino que se convierta en instrumento para que las personas eludan formas más plausibles de responsabilidad, lo que se ve con toda claridad en el caso de la responsabilidad penal.

Auditoría algorítmica, *por I. Lledó Benito*

Este término, como tal, no se explicita en el Reglamento de IA de la Unión Europea, de 13 de junio de 2024, pero lo que significa, lisa y llanamente, es un análisis de los datos, del sistema operativo de la IA. Que las normativas específicas se apliquen a las empresas de IA. El gran problema, a nuestro juicio, se encuentra

en la transparencia, en la política de los sesgos, ya que pueden conllevar riesgos y costes financieros.

No solo es importante en las áreas técnicas evaluadas que los sistemas cumplan los estándares de trazabilidad, sino que estén los algoritmos entrenados con unos parámetros éticos incorporados en el diseño, que es fundamental a la hora de definir decisiones que afecten a derechos fundamentales, como es la protección de datos, y en el ámbito procesal perjudicar el derecho a la presunción de inocencia, el debido proceso, el derecho a la defensa, etcétera.

No obstante, para conocer el proceso habría que acudir, por un lado, al art. 10, referido a los datos y gobernanza, en donde los datos de entrenamiento tienen que ser evaluados y sometidos a prácticas de gobernanza (es decir, criterios legales y deontológicos aceptados en la política de actuación de la UE). Es clave, por tanto, la actualización de los datos para evitar sesgos que violen derechos fundamentales. Es importante, como se explicita en el ordinal 4), que los datos, su *conspectus*, tengan en cuenta el entorno geográfico conductual o funcional (sobre todo en los sistemas de IA de alto riesgo). Del mismo modo, como dice el art. 12, es decisivo que exista un registro automático (archivos de registro que permitan la viabilidad transparente del sistema a la hora de llevar a cabo la auditoria), ello permitirá seguir la trazabilidad de todos los datos y su seguimiento. Y algo muy importante: la identificación de las personas físicas (art. 12), es decir, operadores, programadores, técnicos, ingenieros, todos ellos implicados en la verificación de resultados.

Nos parece relevante, asimismo, como expresa el art. 27, la evaluación del impacto en los derechos fundamentales, así el punto d) nos habla de los riesgos específicos que pueden afectar a las categorías de personas físicas y colectivos que pueden verse afectadas [*vide* también punto c)].

Finalmente, es reveladora la sección 3, «Obligaciones de los proveedores de modelos de IA de uso general con riesgo sistémico», para evaluar los modelos de conformidad con protocolos y herramientas que reflejen el estado de la técnica (*vide* art. 55 RIA). Y a mayor abundamiento, estos proveedores de modelos de la IA podrán recurrir a códigos de buenas prácticas (art. 55-2 y art. 56 RIA).

Transparencia algorítmica, *por D. C. Wisner Glusko*

Se define con este término al principio de gobernanza que exige que los procesos, datos de entrada y criterios de decisión de un sistema de IA sean visibles, conocibles, comprensibles, explicables y auditables para los usuarios, reguladores y otras partes que pudieran resultar afectadas por dichos sistemas algorítmicos. Este concepto trasciende la mera publicación del código fuente, implicando una

apertura informativa sobre la lógica subyacente, el propósito y las limitaciones del sistema. Este principio va más allá de la mera publicidad de la lógica técnica; es un requisito instrumental para garantizar la seguridad jurídica, la no discriminación y la protección de los derechos fundamentales.

Un sistema se considera transparente cuando es conocible y descifrable, permitiendo monitorizar sus decisiones para garantizar que opera dentro de parámetros y principios éticos y jurídicos. La transparencia algorítmica se articula a través de diversas dimensiones operativas. La transparencia de los datos requiere la existencia de documentación detallada sobre el origen, calidad y ponderación de los conjuntos de datos, identificando posibles sesgos y el uso de categorías especiales de datos personales. El modelo de IA también debe ser transparente en la medida que describa, de forma inteligible, la arquitectura del sistema, su finalidad prevista, sus indicadores de rendimiento y debilidades conocidas, de modo que personas no expertas puedan comprender su funcionamiento. Asimismo, requiere del registro y monitorización continua de las decisiones y errores del sistema durante todo su ciclo de vida, permitiendo su revisión ante posibles fallos o desviaciones del diseño original, lo que supone la verificación de la transparencia tanto *ex ante* como *ex post*.

El Reglamento de IA de la Unión Europea 2024/1689 (RIA) establece normas armonizadas de transparencia para garantizar la fiabilidad de la IA.

En ese sentido, los sistemas de alto riesgo deben diseñarse y desarrollarse de tal modo que funcionen con un nivel de transparencia suficiente para que los responsables del despliegue interpreten y usen correctamente los resultados de salida (art. 13.1 RIA). Deben ir acompañados de instrucciones de uso claras que incluyan especificaciones sobre precisión, solidez, ciberseguridad y medidas de supervisión humana, incluidas las medidas técnicas establecidas para facilitar la interpretación de los resultados de salida de los sistemas de IA de alto riesgo por parte de los responsables del despliegue (Art. 13.2 y 3 RIA).

Por su parte, los proveedores deben informar a las personas físicas cuando estén interactuando con un sistema de IA, salvo que resulte evidente desde el punto de vista de una persona física razonablemente informada, atenta y perspicaz, teniendo en cuenta las circunstancias y el contexto de utilización. También deben notificar específicamente el uso de sistemas de reconocimiento de emociones o categorización biométrica (Art. 50.1 RIA).

Respecto de los contenidos generados por IA, la norma europea impone la obligación de marcar los resultados de salida de forma que sea posible detectar su origen artificial, especialmente en el caso de ultrasuplantaciones (Art. 50. 2 RIA).

Gran parte de los requisitos de transparencia recaen en la documentación técnica de un sistema de IA de alto riesgo. Al respecto, el artículo 11 del RIA establece que la misma debe ser elaborada antes de su introducción en el mercado o puesta en servicio y que debe mantenerse actualizada, sirviendo como prueba de conformidad ante las autoridades nacionales competentes.

Finalmente, la transparencia algorítmica es la premisa necesaria para que las personas físicas puedan vigilar el sistema, detectar anomalías y detener su funcionamiento si es preciso (Art. 14 RIA).

La transparencia algorítmica como derecho de los ciudadanos se refleja en la Carta de Derechos Digitales de España, que reconoce entre los derechos ante la IA (XXV) el establecimiento de condiciones de transparencia, auditabilidad, explicabilidad, trazabilidad, supervisión humana y gobernanza; siendo, en todo caso la información facilitada tanto accesible como comprensible.

Específicamente, la transparencia sobre el uso de instrumentos de IA y sobre su funcionamiento y alcance en cada procedimiento concreto, y, en particular, acerca de los datos utilizados, su margen de error, su ámbito de aplicación y su carácter decisorio o no decisorio, es uno de los derechos de la ciudadanía en relación con la IA reconocidos en esta Carta en el marco de la actuación administrativa (XVIII).

Aunque a menudo los conceptos de transparencia, explicabilidad y trazabilidad se utilizan como sinónimos, estos conceptos poseen alcances jurídicos y técnicos diferenciados. Mientras la transparencia alude al marco global que permite conocer el propósito, los datos y las reglas generales que rigen el sistema de IA, la explicabilidad es la manifestación técnica de la transparencia aplicada a un caso individual y se refiere a la capacidad de justificar una decisión específica en términos comprensibles para el ser humano, desglosando la contribución de cada variable al resultado final, y la trazabilidad consiste en la capacidad de reconstruir el historial de un sistema de IA mediante la conservación sistemática de archivos de registro, documentando de forma retrospectiva quién interactuó con el sistema, qué datos de entrada se utilizaron y en qué momento se produjeron cambios significativos o errores, permitiendo auditorías precisas sobre la cadena de valor.

Brechas digitales, *por T. Casadei (trad. de C. Pérez García)*

Las tecnologías digitales permean en nuestros días los más diversos ámbitos, así de la vida privada como de la social, mas ello no ha traído consigo una distribución equitativa de las oportunidades que ofrecen: es en esta disonancia en la que se enmarca la noción de «brechas digitales».

Si bien en un primer momento -a saber, desde que los periodistas de *Los Angeles Times* acuñaron el término «*digital divide*» en 1995- la literatura se centró, ante todo, en el problema del acceso a las tecnologías (*physical access*), luego la difusión cada vez más generalizada de los dispositivos y de las fuentes de acceso a la red, junto con una serie de investigaciones publicadas en el año 2004, han llevado a prestar más atención al perfil de las capacidades de uso de las tecnologías y, en particular, de internet (*skills and usage*). Una ulterior línea de investigación ha profundizado en la cuestión de los resultados (*outcomes*), es decir, de las desigualdades que resultan, como *output*, de un acceso y un uso desiguales de la red y de las tecnologías relacionadas (*material and conditional access* y *digital media use*), como sucede, por ejemplo, respecto de la frecuencia de uso, del tiempo transcurrido en la red, de la diversificación de los usos, del tipo de actividades desarrolladas y de la capacidad efectiva para obtener de la digitalización un beneficio real. A estos aspectos se añaden la fiabilidad las infraestructuras, la accesibilidad de las interfaces y, a un nivel más general, el papel de las instituciones.

El resultado es un fenómeno multidimensional y diferenciado según el contexto, irreductible a la dicotomía entre quienes poseen los instrumentos y quienes no.

Los hilos conductores a cuyo través se extienden las brechas digitales son múltiples y se entrelazan: el género, la edad, la condición socioeconómica, el nivel educativo, la (dis)capacidad, el contexto territorial… La reflexión jurídica sitúa estos problemas en el horizonte del constitucionalismo digital, tratando de hacer frente a una dinámica de acumulación de desventajas merced a la cual los ya marginados son quienes resultan estar más expuestos a los efectos negativos de la automatización.

Lejos de ser un problema transitorio, abocado a reabsorberse por el simple crecimiento de las infraestructuras, las formas de brecha arraigan en líneas de fractura prexistentes -ora reforzándolas, ora reconfigurándolas- y tienden, así, a reproducir y, en muchos casos, a hacer más profundas las asimetrías sociales, económicas y culturales en las que se dan.

Por consiguiente, la cuestión no es de mera naturaleza técnica o infraestructural: al actuar sobre las condiciones materiales de la igualdad, afecta también a la efectividad de los derechos, las prácticas ciudadanas y la posibilidad misma de ejercer, en condiciones de igualdad, aquellas libertades que la transformación digital hace más accesibles y, a un tiempo, más condicionadas.

Las coordenadas trazadas no describen un fenómeno estático, pues experimenta aquel, más bien, una doble transformación, que en los últimos

tiempos ha devenido evidente. Por un lado, la brecha tiende a desplazarse hacia arriba, rearticulándose en torno a la capacidad ya no tan solo de acceder a la red o de utilizarla, sino, además, de comprender, gobernar y explotar los sistemas algorítmicos y de IA: es lo que se viene definiendo como *algorithmic divide*, destinada, con certeza, a acentuarse con el advenimiento de la IA generativa, que separa a quienes dominan su lógica de quienes padecen, de manera pasiva, sus efectos.

A esta dinámica se suma, además, una dimensión infraestructural y geopolítica, es decir, la concentración de la potencia de cálculo y la extracción asimétrica de los datos, a lo largo del eje que opone el norte y el sur del mundo: esto configura las brechas no solo como distancia en el acceso a un recurso, sino, y cada vez más, como una asimetría de poder sobre las propias infraestructuras de lo digital.

Paralelas a estas fronteras se miden, pues, tanto la solidez de las políticas de inclusión y de los instrumentos regulativos, y con ello la posibilidad de incidir sobre las formas de brecha y sobre la velocidad de las transformaciones que en el seno de la sociedad les afectan, cuanto la capacidad de las instituciones para trazar nuevas políticas de promoción de la igualdad a escala planetaria.

Discriminación algorítmica, *por V. Barone*

La discriminación algorítmica constituye uno de los fenómenos más relevantes y problemáticos en el contexto de la transformación digital contemporánea. Se refiere a aquellas situaciones en las que sistemas automatizados de toma de decisiones producen resultados que generan un trato desigual injustificado entre individuos o grupos, afectando negativamente el acceso a oportunidades, recursos o derechos fundamentales. Lejos de ser neutrales, los algoritmos reflejan y, en muchos casos, amplifican las desigualdades estructurales presentes en la sociedad.

Desde un punto de vista conceptual, la discriminación algorítmica puede entenderse como una evolución de las formas tradicionales de discriminación. Mientras que estas últimas se basaban en categorías explícitas como género, raza o edad, los sistemas algorítmicos introducen mecanismos más complejos y menos visibles. En particular, el uso de técnicas de aprendizaje automático permite identificar correlaciones en grandes volúmenes de datos, lo que conduce a decisiones basadas no en criterios jurídicamente relevantes, sino en patrones estadísticos que pueden reproducir sesgos históricos.

Las causas de la discriminación algorítmica son múltiples. En primer lugar, los datos utilizados para entrenar los modelos suelen estar sesgados. Si un

sistema aprende a partir de datos históricos en los que existen desigualdades -por ejemplo, menor contratación de mujeres en determinados sectores-, es probable que reproduzca dichas tendencias. En segundo lugar, intervienen los sesgos de los desarrolladores, quienes, consciente o inconscientemente, incorporan sus propias perspectivas y supuestos en el diseño del sistema. En tercer lugar, resulta especialmente relevante el uso de variables *proxy*, es decir, variables aparentemente neutrales que, en la práctica, funcionan como sustitutos de características protegidas, como el origen étnico o la condición socioeconómica.

Un elemento adicional que agrava el problema es el carácter opaco de muchos sistemas algorítmicos. La denominada "caja negra" dificulta la comprensión de los procesos internos de decisión, impidiendo a los individuos afectados conocer las razones de un resultado adverso y, por tanto, impugnarlas eficazmente. Esta opacidad se combina con el llamado efecto de escala: las decisiones automatizadas no afectan a casos aislados, sino a grandes colectivos, multiplicando el impacto potencial de cualquier sesgo.

La discriminación algorítmica se manifiesta en diversos ámbitos. En el mercado laboral, algoritmos de selección pueden favorecer sistemáticamente a determinados perfiles en detrimento de otros. En el sector financiero, los sistemas de *scoring* crediticio pueden excluir a individuos basándose en patrones indirectos relacionados con su entorno social o geográfico. En el ámbito penal, herramientas de evaluación del riesgo de reincidencia han sido criticadas por reproducir sesgos raciales. Asimismo, en el entorno digital, los sistemas de recomendación y publicidad pueden reforzar estereotipos y limitar la exposición a determinadas oportunidades.

Desde el punto de vista jurídico, la discriminación algorítmica plantea importantes desafíos. Los marcos normativos existentes, diseñados para abordar decisiones humanas, resultan en muchos casos insuficientes frente a sistemas automatizados. En particular, la carga de la prueba recae frecuentemente sobre el individuo afectado, quien debe demostrar la existencia de discriminación sin tener acceso a la lógica del algoritmo. Además, la fragmentación de responsabilidades entre distintos actores —desarrolladores, proveedores de datos, usuarios del sistema— dificulta la atribución de responsabilidad jurídica.

No obstante, los algoritmos también pueden ofrecer oportunidades para combatir la discriminación. A diferencia de las decisiones humanas, los sistemas automatizados pueden ser auditados, replicados y corregidos, lo que abre la puerta a mecanismos de control más sofisticados. En este sentido, se han propuesto diversas estrategias, como la incorporación de principios de *fairness by design*, la realización de auditorías algorítmicas y el establecimiento de obligaciones de transparencia.

La regulación de la discriminación algorítmica exige no solo la adaptación de los instrumentos jurídicos existentes, sino también una reflexión más profunda sobre el papel de la tecnología en la organización social. Garantizar que los algoritmos no se conviertan en vehículos de injusticia constituye, en última instancia, una condición esencial para la legitimidad de las sociedades digitales contemporáneas.

Sesgo (*Bias*), *por D. C. Wisner Glusko*

La Organización Internacional de Normalización (ISO) define el sesgo de forma general como *"el grado en que un valor de referencia se desvía de la verdad"*.

Desde una perspectiva jurídica y deontológica, el sesgo algorítmico se define como la predisposición sistemática y errónea que surge en los modelos de IA, resultando en decisiones o predicciones distorsionadas que pueden generar consecuencias perjudiciales o dañosas para individuos o colectivos. Técnicamente, se manifiesta como un error sistemático en el que se incurre cuando los muestreos o ensayos favorecen ciertas respuestas frente a otras, alejándose de la objetividad necesaria en la toma de decisiones automatizadas.

El sesgo no es una falla tecnológica aislada, sino un reflejo de desigualdades estructurales. En ocasiones se debe a los datos de entrenamiento, ya que los modelos absorben sesgos históricos incrustados en los datos. Por tanto, si los datos reflejan desigualdad social, el algoritmo la replicará y amplificará.

La intervención humana puede impactar negativamente en el ciclo de vida del sistema de IA. Los desarrolladores pueden introducir involuntariamente sus propios prejuicios al seleccionar variables o al definir el problema. El sesgo también puede infiltrarse en cualquier etapa, desde la adquisición y el etiquetado, donde el error humano o los criterios subjetivos pueden influir, hasta el procesamiento completo de datos.

Según sea por los datos o por la interacción humana, la doctrina mayoritaria diferencia diferentes tipos de sesgos. El sesgo de selección o muestral, que ocurre cuando los datos de entrenamiento no representan adecuadamente a toda la población, omitiendo la diversidad necesaria para que el modelo sea equitativo. El sesgo de confirmación se produce cuando el sistema se basa excesivamente en creencias o patrones preexistentes, duplicando prejuicios en lugar de identificar nuevas tendencias objetivas. Directamente relacionado con este último, el sesgo de automatización, que consiste en una posible tendencia a confiar automáticamente o en exceso en los resultados de salida generados por un sistema de IA (Art. 14.4.b RIA). El sesgo de exclusión, que surge al omitir variables fundamentales que influyen en el resultado, lo que puede perjudicar a grupos demográficos

específicos. El sesgo de estereotipos y prejuicios, que se hace presente cuando el algoritmo refuerza supuestos sociales erróneos o mitos. El sesgo de homogeneidad fuera del grupo, cuando los desarrolladores crean algoritmos menos capaces de distinguir entre individuos que no pertenecen al grupo mayoritario presente en los datos de entrenamiento, lo que deriva en errores de clasificación racial. Y el sesgo de medición, causado por datos incompletos o fallos en la infraestructura digital -como sensores defectuosos- que generan distorsiones en la captura de información. En el caso de los sistemas que continúan aprendiendo tras su despliegue, deben diseñarse para evitar que los resultados de salida sesgados influyan en las futuras entradas de datos, evitando así la amplificación del error.

El Reglamento (UE) 2024/1689 (RIA) reconoce el sesgo no solo como un error técnico, sino como un riesgo jurídico que debe gestionarse mediante la transparencia, la calidad del dato y el control humano constante. El impacto de estos sesgos en la toma de decisiones automatizada trasciende el aspecto técnico para entrar en el ámbito de la vulneración de derechos fundamentales.

El RIA exige que los sistemas de alto riesgo se desarrollen con conjuntos de datos de entrenamiento, validación y prueba de alta calidad (Art. 10 RIA). En concreto, obliga a examinar los conjuntos de datos atendiendo a posibles sesgos que puedan afectar a la salud, la seguridad o dar lugar a discriminaciones prohibidas por el Derecho de la Unión (Art. 10.2.f), debiendo implementar medidas específicas para detectar, prevenir y mitigar los sesgos identificados (Art. 10.2.g). Los datos deben ser lo suficientemente representativos y tener las propiedades estadísticas adecuadas en relación con los colectivos de personas donde se utilizará el sistema, con el fin de evitar que se perpetúe o amplifique la discriminación (Art. 10.3). También será posible el procesamiento de datos sensibles para detectar y corregir sesgos (Art. 10.5 RIA).

Asimismo, establece que los proveedores deben informar sobre el funcionamiento del sistema respecto a colectivos específicos y los niveles de precisión esperados para evitar interpretaciones erróneas (Art. 13.3.b); los responsables del despliegue en sectores públicos o servicios esenciales tienen que realizar una evaluación previa que identifique los riesgos para colectivos vulnerables y definir medidas de gobernanza para mitigarlos (Art. 27 RIA); y los proveedores de estos modelos deben incluir en su documentación técnica los métodos utilizados para detectar sesgos identificables (Art. 53 y Anexo XI).

Sesgo de automatización, *por D. C. Wisner Glusko*

El sesgo de automatización se define como una tendencia psicológica y heurística cognitiva mediante la cual el operador humano manifiesta una confianza excesiva e injustificada en las sugerencias, recomendaciones o decisiones emitidas

por un sistema automatizado o de IA. El ser humano percibe a la IA como una fuente autorizada de información, lo que lleva a validar y asimilar sus criterios sin cuestionarlos.

Este comportamiento, que se basa en la confianza en la supuesta perfección de la toma de decisiones automatizada basada en el uso de algoritmos y de datos y en la presunción de infalibilidad de la IA, implica que el sujeto tiende a favorecer la respuesta tecnológica incluso cuando dispone de información contradictoria proveniente de sus propios sentidos, conocimientos previos o evidencias externas.

El Reglamento (UE) 2024/1689 (RIA) reconoce el sesgo de automatización como un riesgo latente que debe ser gestionado, especialmente en sistemas de alto riesgo, es decir aquellos que comprometen la salud, la seguridad o los derechos fundamentales (Art. 6. 3 RIA). Y aunque no es un concepto definido de forma clara en otros cuerpos jurídicos, el RIA lo aborda de manera transversal.

El artículo 14 -supervisión humana- inciso 4. b), impone a los proveedores la obligación de diseñar sistemas que permitan a las personas físicas encargadas de la supervisión "ser plenamente conscientes de la posible tendencia a confiar automáticamente o en exceso en los resultados de salida generados por un sistema de IA de alto riesgo («sesgo de automatización»), en particular con aquellos sistemas que se utilizan para aportar información o recomendaciones con el fin de que personas físicas adopten una decisión", y, por supuesto, de contrarrestarla. La supervisión humana debe ser efectiva, basada en esa capacidad técnica y cognitiva para cuestionar y contrastar el *output* de la máquina cuando sea necesario.

Por su parte, el artículo 26.2 RIA -Obligaciones de los responsables del despliegue de sistemas de IA de alto riesgo- establece el deber de asignar la supervisión a personas que posean la competencia, formación y autoridad necesarias con objeto de ejercer un control real sobre el sistema y mitigar el sesgo de automatización.

La configuración de este sesgo responde a múltiples factores de naturaleza psicológica, técnica y social. El cerebro humano busca minimizar el esfuerzo mental (economía cognitiva), dado que, al delegar la toma de decisiones en un algoritmo, permite reducir la carga cognitiva, optando, en consecuencia, por la vía del menor esfuerzo. Asimismo, este sesgo se fundamenta en la existencia de una creencia arraigada de que los sistemas tecnológicos son intrínsecamente más precisos, eficientes y libres de prejuicios que el juicio humano. También la representación profesional de los datos a través de gráficos, alertas visuales o determinadas interfaces confiere al sistema una jerarquía simbólica ante la cual el humano tiende a someterse, de forma análoga a la obediencia jerárquica laboral o profesional.

El sesgo de automatización induce dos fallos operativos críticos en la toma de decisiones. Por un lado, los denominados errores de comisión, que ocurren cuando el usuario sigue ciegamente una instrucción o recomendación del sistema que es incorrecta, ignorando señales de advertencia o el sentido común. Un ejemplo clásico es el fenómeno del "*Death by GPS*", donde conductores siguen rutas peligrosas o inexistentes a pesar de la evidencia visual del entorno. Por otro lado, los errores de omisión, que se producen cuando el operador no actúa ante una situación anómala o crítica porque el sistema automatizado no ha emitido ninguna alerta ni ha sugerido intervención alguna.

La confianza en el sistema de IA se acumula progresivamente tras experiencias exitosas con el mismo, lo que deriva en una "negligencia aprendida". La falta de conciencia sobre las limitaciones de la IA o de incentivos para cuestionar sus resultados fomenta la aceptación acrítica. De esta manera, el sesgo de automatización convierte un sistema de "apoyo a la decisión" en un sistema de "sustitución de la decisión", donde el humano abdica de su responsabilidad y criterio, lo que representa un riesgo crítico para la seguridad y la equidad. Por todo ello, el diseño del sistema de IA debe garantizar que el juicio humano sea siempre el último filtro en la adopción de decisiones, promoviendo la responsabilidad personal del operador sobre el resultado final.

El sesgo de automatización, lejos de ser un fallo tecnológico, se configura como un riesgo humano derivado del uso de la tecnología, cuya mitigación exige un equilibrio entre la eficiencia algorítmica, el mantenimiento de la autonomía y la supervisión humana.

Sesgo de confirmación, *por M. O. Sánchez Martínez*

El sesgo de confirmación es un tipo de sesgo cognitivo, caracterizado por la propensión a buscar, procesar, interpretar y recordar la información que valide o confirme hipótesis o creencias propias, al tiempo que tiende a descartar, o no valorar al mismo nivel, otras alternativas posibles. Es considerado un recurso mental, del que se hace uso normalmente de forma inconsciente, para simplificar problemas complejos, evitar análisis profundos o enfrentarse a la dificultad y el esfuerzo que supone comprender, discutir y rebatir racionalmente alternativas que, a priori ajenas a nuestras convicciones e intereses, nos resultan incómodas, carentes de relevancia o poco útiles. De esta manera, el sesgo de confirmación refuerza nuestra propia visión de la realidad, nos predispone a aceptar aquello que no implique un desafío a ideas, creencias y sentimientos personales y a rechazar todo lo que se oponga a las mismas.

Relacionado con los procesos de comunicación e información, el sesgo de confirmación no aparece con las tecnologías digitales y la IA. El término fue

acuñado en los años 60 del siglo xx por el psicólogo cognitivo Peter C. Wason, pero ya había sido puesto de relieve mucho antes por pensadores de diversos ámbitos. Entre ellos, Francis Bacon, en *Novum Organum*, al advertir que cuando el espíritu humano acoge una idea se empeña en todo aquello que sirva para apoyarla y confirmarla, mientras ignora, olvida, desprecia o rechaza lo que se oponga a ella -libro 1º, aforismo 46-.

Si bien no es un fenómeno exclusivo de las nuevas tecnologías, el sesgo de confirmación, y sus efectos negativos, se ven favorecidos y amplificados por ellas. Los procesos de IA, partiendo de los datos que vamos dejando en nuestro historial de búsquedas en red, seleccionan y priorizan la comunicación que nos transmiten, limitando así la autonomía personal en su recepción y recuperación. La información, convenientemente filtrada, en sintonía con nuestras ideas, creencias, intereses y relaciones previas, ejercerá una estimable influencia en su consolidación y refuerzo. El pasado proyectado hacia el futuro entorpece la irrupción de lo nuevo, el avance hacia ámbitos intransitados. Se contribuye, de esta manera, a la reducción del ámbito de inquietudes e intereses personales, a la formulación de conclusiones erróneas, a decisiones desacertadas y/o a la pérdida de oportunidades.

Desde el punto de vista de la convivencia y la sociabilidad, el impacto del sesgo de confirmación nos sitúa ante una sociedad percibida de forma más homogénea y menos plural. La tendencia a prestar atención prácticamente solo a los discursos afines, y a evitar escuchar las disonancias del diferente, abocan a intensificar la polarización, la fragmentación, las tensiones sociales y, a la postre, a inhibir el debate constructivo ineludible para el buen funcionamiento del sistema democrático.

Evitar o minimizar los efectos negativos de los sesgos, entre ellos el de confirmación, es uno de los grandes retos éticos y jurídicos en la era de las nuevas tecnologías de la comunicación y de la IA. El sesgo parte y traslada una precomprensión o prejuicio sobre la realidad que compromete algunas de las míticas ideas relativas al uso y aplicaciones de la tecnología, como la neutralidad de los datos, la objetividad, imparcialidad y precisión de su tratamiento, y la infalibilidad de sus resultados. Por otro lado, el exceso de confianza en estas presumibles características es parte del problema para actuar contra los sesgos, porque sobre ellas se sustenta la convicción de la IA como liberadora de los condicionantes subjetivos, propios de los seres humanos, en el conocimiento y la toma de decisiones. Al respecto, la inclinación a aceptar sin reservas sus resultados debilita los estándares críticos, los procesos de transparencia, de supervisión y corrección de errores.

Reconocer que la IA puede equivocarse, y ser conscientes de la vulnerabilidad frente a los sesgos, son los primeros pasos para lograr una eficaz protección frente a ellos. Se debería contar con una adecuada formación de la ciudadanía, una exigente regulación jurídica, a nivel nacional e internacional, y estrictos compromisos éticos por parte de las empresas tecnológicas, a fin de garantizar la accesibilidad a una información de calidad, que permita diversificar fuentes informativas -con el propósito de que podamos ser capaces de contrastar, identificar y neutralizar los efectos de los filtros-, abrirse a contenidos plurales, incorporar una perspectiva crítica a las ideas propias y debatir con las ajenas.

Propiedad intelectual e IA, *por M. A. Adame Martínez*

La relación entre IA y propiedad intelectual plantea, en el estado actual del Derecho, tres grandes cuestiones que carecen todavía de respuesta unívoca en la mayoría de los ordenamientos.

Primera, la protección de los *outputs* generados por IA. ¿Puede una obra creada autónomamente por un sistema de IA gozar de protección por derechos de autor? Los ordenamientos de la familia del *droit d'auteur* -entre ellos el español- exigen que la obra sea expresión de la personalidad de un ser humano, como condición de originalidad (art. 10 TRLPI). En consecuencia, las obras generadas íntegramente por IA sin aportación creativa humana identificable no son, en principio, susceptibles de protección en España ni en la mayoría de los ordenamientos continentales -aunque no existe todavía jurisprudencia consolidada del TJUE específicamente sobre IA generativa, y la solución final dependerá del grado y la naturaleza de la intervención humana en cada caso-. En el sistema estadounidense, la *Copyright Office* ha adoptado posiciones similares en sus sucesivas guías de 2023 y 2024, exigiendo autoría humana como presupuesto de la protección; y los tribunales federales que han conocido de la materia han confirmado ese criterio (véase, entre otras, la resolución del Tribunal Federal del Distrito de Columbia en el asunto *Thaler v. Perlmutter*, 2023). La cuestión se complica cuando la intervención humana -a través del *prompt* o de la selección y edición de los resultados- es relevante pero no exclusiva: en ese espacio fronterizo se concentra el grueso del debate doctrinal actual.

Segunda, la titularidad de los *outputs* cuando existe intervención humana. Si se admite que el *output* con aportación creativa humana puede ser protegible, surge la pregunta de a quién corresponde la titularidad: ¿al usuario que formuló el *prompt*?, ¿al proveedor del sistema de IA?, ¿a quien aportó los datos de entrenamiento? Ningún ordenamiento europeo ha resuelto esta cuestión de forma expresa, y las condiciones generales de uso de los grandes proveedores de

IA generativa establecen soluciones contractuales dispares que no siempre tienen respaldo en el Derecho positivo aplicable.

Tercera, el entrenamiento de modelos sobre obras protegidas. El entrenamiento de los grandes modelos de lenguaje e imagen requiere el procesamiento de enormes volúmenes de texto, imágenes y otros contenidos obtenidos de internet, gran parte de los cuales están protegidos por derechos de autor. La excepción de minería de textos y datos prevista en el art. 4 de la Directiva (UE) 2019/790 (Directiva sobre derechos de autor en el mercado único digital) permite este tratamiento salvo que el titular haya manifestado expresamente su reserva (*opt-out*). Sin embargo, la operatividad práctica del *opt-out* frente a modelos ya entrenados es jurídicamente incierta: el mecanismo está diseñado para actuar *ex ante*, antes del procesamiento de las obras, y su eficacia retroactiva -respecto de entrenamientos realizados antes de que el titular manifestara su reserva- es uno de los puntos centrales de la litigación en curso. El Reglamento de IA (UE) 2024/1689 impone a los proveedores de modelos de propósito general la obligación de publicar un resumen suficientemente detallado de los contenidos utilizados en el entrenamiento (art. 53.1.d). Esta obligación abre la vía a la reclamación de titulares de derechos que no consintieron el uso de sus obras, aunque la doctrina ha señalado críticamente que la exigencia de un "resumen suficientemente detallado" puede resultar insuficiente para hacer efectivo el control de los derechos en la práctica. Varios procesos judiciales en curso en EE.UU. y Europa -promovidos por grupos editoriales, agencias de noticias y artistas visuales contra proveedores de IA generativa- definirán en los próximos años el perímetro efectivo de la protección.

Debido proceso tecnológico (*Technological due process*)*, por J. I. Solar Cayón*

El concepto "debido proceso tecnológico" fue acuñado por la profesora Danielle K. Citron en un artículo publicado en 2008. Dicha expresión alude al establecimiento de una serie de garantías procedimentales que aseguren la responsabilidad, transparencia e imparcialidad en el uso de sistemas algorítmicos para la adopción de decisiones automatizadas por parte de las administraciones públicas. La idea básica subyacente a esta construcción doctrinal es la necesidad de realizar una relectura de las garantías implícitas en la tradición constitucional estadounidense del «debido proceso legal» (*due process of law*), a la luz del contexto tecnológico actual, para adaptarlas a la problemática planteada por el empleo de sistemas algorítmicos en la toma de decisiones administrativas que afectan significativamente a los derechos de los afectados. Animada por este propósito, la noción de "debido proceso tecnológico" reivindica que los derechos constitucionales básicos tradicionalmente ligados a la cláusula del *due process of law* (derecho a ser informado, derecho a ser oído y derecho a una decisión

motivada) sean incorporados en el proceso de diseño y funcionamiento de los sistemas algorítmicos utilizados para la adopción de decisiones administrativas.

La razón que justifica esta demanda se encuentra en que aquellas exigencias constitucionales del debido proceso legal, que están orientadas a proteger los derechos del individuo mediante la imposición de determinadas garantías procedimentales en los procesos de creación y de aplicación del Derecho (ya se trate de la aplicación administrativa de la norma o de la adjudicación judicial), son ineficaces cuando la administración emplea sistemas algorítmicos de toma automatizada de decisiones. Nuestras concepciones del Estado de Derecho y, particularmente, del imperio de la ley, se asientan sobre el presupuesto de una separación clara entre los momentos y los procesos de creación y aplicación del Derecho (entre las figuras del legislador y del administrador o el juez), sujeto cada uno de ellos a una serie de procedimientos que tratan de asegurar la legalidad, racionalidad e imparcialidad de las decisiones. Sin embargo, la utilización de sistemas algorítmicos para la toma automatizada de decisiones administrativas colapsa la distancia existente entre legislación y adjudicación, en tanto el código computacional se convierte en la regla cuya ejecución automática determina al mismo tiempo el sentido de las resoluciones individuales, eludiendo así las garantías procedimentales que han de regir cada uno de aquellos procesos jurídicos. El administrado se ve, por tanto, afectado en sus derechos e intereses legítimos por una decisión que es el resultado de la ejecución automática de un programa informático que no ha sido sometido a control democrático alguno, sin que en muchos casos haya sido siquiera informado del empleo del sistema algorítmico por parte de la administración ni tenido la oportunidad de ser oído.

Así pues, los procedimientos tradicionales diseñados para proteger al individuo frente a potenciales abusos y arbitrariedades por parte del legislador y/o de la administración no resultan adecuados para subsanar esos déficits de rendición de cuentas en el empleo de sistemas algorítmicos para la toma de decisiones automatizadas. De ahí que resulte esencial configurar un nuevo concepto de debido proceso dirigido a reconstruir las garantías procesales del siglo pasado en el marco del actual contexto tecnológico, especialmente teniendo en cuenta el incremento exponencial de sistemas de IA empleados por las administraciones públicas. Y una de las estrategias más efectivas para el control de estos sistemas -especialmente, a la vista de las dificultades que en muchos casos plantea la interpretabilidad y explicabilidad de sus resultados- es el establecimiento de una serie de garantías formales en sus procesos de diseño, entrenamiento y despliegue, así como de procedimientos de monitorización y supervisión de sus resultados. En esta dirección, el contenido del "debido proceso tecnológico" vendría constituido por un conjunto de procedimientos que garanticen que los algoritmos satisfacen ciertos estándares de diseño, evaluación y revisión, al objeto de asegurar su imparcialidad y corrección. Entre las garantías básicas incluidas en este concepto

se encuentran la participación de todos los grupos potencialmente afectados en el diseño del sistema, la contestabilidad o posibilidad de impugnar el punto de vista "codificado" (el algoritmo y el código fuente), y la transparencia a lo largo de todo el ciclo de vida del sistema. Se trata, en definitiva, de incorporar en el mismo diseño de la arquitectura tecnológica de los sistemas de IA, en los algoritmos que los mueven y en los *datasets* de los que se alimentan las garantías procedimentales básicas inherentes a la doctrina constitucional del debido proceso.

2. Conceptos legales

A) RIA y Legislación europea

Sistema de IA, *por M. Barrio Andrés*

De acuerdo con el artículo 3.1) del Reglamento UE de IA (RIA), un «sistema de IA» es "un sistema basado en una máquina que está diseñado para funcionar con distintos niveles de autonomía y que pueda mostrar capacidad de adaptación tras el despliegue, y que, para objetivos explícitos o implícitos, infiere de la información de entrada que recibe la manera de generar resultados de salida, como predicciones, contenidos, recomendaciones o decisiones, que pueden influir en entornos físicos o virtuales".

La definición del término sistema de IA, que se debatió intensamente en el proceso legislativo y en la doctrina, se caracteriza por un concepto muy amplio en línea con la definición de la OCDE, que se actualizó en 2023. Al final, este resultado representó el único compromiso posible entre las distintas posiciones en las negociaciones políticas en Bruselas. Mientras que los partidos de izquierda querían que casi todas las tecnologías quedaran incluidas en el RIA, los Estados miembros y los partidos liberal-conservadores exigían que la norma se centrara únicamente en las tecnologías de IA más modernas, como el aprendizaje automático (el *machine learning*). Este último grupo también rechazó la propuesta inicial de la Comisión, incluido su proyectado anexo I, ya que también englobó diversos métodos estadísticos como IA.

La definición de la OCDE, leída en consonancia con el considerando 12 del RIA, estipula ahora en el texto legal definitivo que no se incluyen los sistemas de *software* y enfoques de programación convencionales más sencillos, pero al mismo tiempo deja a las autoridades competentes europeas y nacionales suficiente margen de interpretación para incluir sistemas híbridos de IA con enfoques antiguos y nuevos (por ejemplo, aprendizaje automático para la recopilación de datos, pero

análisis mediante Microsoft Excel, como sucedió en el escándalo holandés de las prestaciones por hijos a cargo).

La definición de IA del RIA se basa esencialmente en siete elementos:

- es un sistema basado en una máquina;
- que está diseñado para funcionar con distintos niveles de autonomía;
- que puede mostrar capacidad de adaptación tras su despliegue;
- y para objetivos explícitos o implícitos;
- infiere, a partir de la información de entrada que recibe, la manera de generar resultados de salida;
- tales como predicciones, contenidos, recomendaciones o decisiones; y
- que pueden influir en entornos físicos o virtuales.

Según el indicado considerando, este enfoque debe, por una parte, ser suficientemente compatible a nivel internacional y técnico. Por otra parte, la definición debe ser suficientemente específica y permitir diferenciar la IA de los sistemas de *software* convencionales y de las operaciones simples basadas en reglas automáticas de la programación informática tradicional (el *if-else*).

Asimismo, la definición no debe ser aplicada mecánicamente, sino que es preciso evaluar las características de cada sistema en cuestión.

Con fecha 6 de febrero de 2025, la Comisión Europea ha publicado las *Directrices sobre la definición de sistema de IA establecida por el RIA*, en ejecución del artículo 96.1.f) del RIA, que no tienen carácter vinculante. La interpretación final vendrá de la mano del Tribunal de Justicia de la Unión Europea en eventuales litigios que puedan plantearse.

Las Directrices precisan que la definición de sistema de IA adopta una perspectiva basada en el ciclo de vida que abarca dos fases principales:

- la fase previa al despliegue, implantación o fase de construcción del sistema; y
- la fase posterior al despliegue o fase de uso del sistema.

Los siete elementos establecidos en dicha definición no requieren estar presentes de manera continua a lo largo de ambas fases del ciclo de vida.

Ahora bien, el elemento esencial para sustantivar un sistema informático como de IA es la inferencia, como capacidad de producir los resultados de salida y capacidad del propio sistema para deducir modelos o algoritmos, o ambos, a partir de información de entrada o datos. Esta propiedad privativa de la IA no está presente en un *software* tradicional, como un videojuego o un procesador de textos. Por eso, sin esta capacidad de inferencia, un sistema informático no se considera de IA a efectos del RIA, aunque pueda ser autónomo o adaptable en un sentido más simple.

Sistema de IA de alto riesgo, *por J. I. Solar Cayón*

El Reglamento UE 2024/1689 de IA adopta un enfoque basado en el riesgo, de modo que la intervención legal resulte en cada caso proporcionada a la gravedad y el alcance de los potenciales riesgos generados por el funcionamiento de cada sistema de IA en el marco de su contexto específico de aplicación. En virtud de dicho enfoque, la regulación europea establece una clasificación de sistemas de IA con arreglo a los siguientes niveles de riesgo:

1. Sistemas prohibidos: aquellos utilizados para llevar a cabo prácticas de manipulación, explotación y control social o que vulneran los valores de respeto de la dignidad humana, libertad, igualdad, democracia y Estado de Derecho y de los derechos fundamentales reconocidos por la Unión.

2. Sistemas de alto riesgo, cuya utilización puede tener consecuencias perjudiciales significativas para la salud, la seguridad y los derechos fundamentales de las personas. Se trata de sistemas permitidos, pero cuyo uso está sujeto a una serie de exigencias técnicas.

3. Sistemas de riesgo limitado: ciertos sistemas destinados a interactuar con personas físicas o a generar contenidos quedan sujetos a determinadas obligaciones de transparencia en virtud de los riesgos específicos de suplantación o engaño que conllevan.

4. Sistemas de riesgo mínimo: el resto de sistemas de IA, que no están sujetos a ninguna obligación específica.

Los sistemas de alto riesgo constituyen el objeto central de la regulación contenida en el reglamento europeo, que impone severas obligaciones, fundamentalmente, a sus proveedores y responsables del despliegue.

En cuanto a su determinación, el artículo 6 del reglamento establece dos vías para la clasificación de un sistema de IA como de alto riesgo:

a) Por una parte, serán considerados sistemas de alto riesgo aquellos que cumplan estas dos condiciones:

   - ser componentes de seguridad de un producto que entre en el ámbito de aplicación de los actos legislativos de la Unión enumerados en el Anexo I (seguridad de la aviación civil, vehículos de motor, aeronaves no tripuladas, productos sanitarios, maquinaria, seguridad de juguetes, ascensores...), o ser el propio sistema de IA uno de dichos productos, y

   - que deban someterse a una evaluación de la conformidad por un tercero antes de su introducción en el mercado o puesta en servicio con arreglo al correspondiente acto legislativo.

b) Asimismo, se consideran de alto riesgo los sistemas contemplados en el Anexo III, que se engloban en ocho ámbitos: biometría; sistemas utilizados como componentes de seguridad en la gestión y el funcionamiento de infraestructuras críticas; educación y formación profesional; empleo, gestión de los trabajadores y acceso al autoempleo; acceso y disfrute de servicios (privados y públicos) esenciales; garantía del cumplimiento del Derecho; migración, asilo y gestión del control fronterizo; y Administración de Justicia y procesos democráticos.

Antes de que un sistema de alto riesgo sea introducido en el mercado o puesto en servicio habrá de someterse a un procedimiento de evaluación de la conformidad mediante el cual habrá de demostrarse que cumple las siguientes exigencias (arts. 8-15), que deberán hacerse operativas mediante una serie de estándares técnicos armonizados:

a) Establecimiento de un sistema de gestión de riesgos durante todo el ciclo de vida del sistema.

b) Sujeción a prácticas de gobernanza y gestión de datos que aseguren la calidad de los datos de entrenamiento, validación y prueba del sistema.

c) Elaboración de documentación técnica que contenga la información necesaria para evaluar si el sistema cumple los requisitos exigidos y facilitar la vigilancia poscomercialización.

d) Los sistemas permitirán técnicamente el registro automático de acontecimientos a lo largo de todo su ciclo de vida.

e) Los sistemas se diseñarán y desarrollarán de un modo que garantice que funcionan con un nivel de transparencia suficiente para que los

responsables del despliegue puedan interpretar y usar correctamente sus resultados.

f) Los sistemas de diseñarán y desarrollarán de tal modo que puedan ser vigilados de manera efectiva por personas físicas durante su funcionamiento.

g) Nivel adecuado de precisión, solidez y ciberseguridad.

Los sistemas de alto riesgo deberán ser registrados en una base de datos de la Unión Europea. Y, una vez introducidos en el mercado o puestos en servicio, los proveedores establecerán un sistema de monitorización que les permita tener en cuenta la experiencia de uso para mejorar su funcionamiento y los procesos de diseño y desarrollo, así como para adoptar las medidas correctoras en el momento oportuno.

Proveedor, *por J. I. Solar Cayón*

El Reglamento de IA de la Unión Europea define al "proveedor" como aquella persona física o jurídica, autoridad pública, órgano u organismo que desarrolle, o para el que se desarrolle, un sistema de IA o un modelo de IA de uso general y lo introduzca en el mercado o ponga en servicio con su propio nombre o marca, ya sea previo pago o gratuitamente.

Se trata de una de las figuras clave en el esquema regulador implantado por el reglamento, en tanto sobre ella recaen gran parte de las obligaciones contenidas en el mismo, especialmente en el caso de los proveedores de sistemas de IA de alto riesgo. En este sentido debe tenerse en cuenta que, a los efectos del cumplimiento de aquellas obligaciones, el reglamento establece que también deberá considerarse proveedor de un sistema de IA de alto riesgo cualquier distribuidor, importador, responsable del despliegue o tercero que: a) ponga su nombre o marca en un sistema de IA de alto riesgo previamente introducido en el mercado o puesto en servicio, b) modifique sustancialmente un sistema de IA de alto riesgo, o c) modifique la finalidad prevista de un sistema de IA, de tal manera que se convierta en un sistema de IA de alto riesgo.

Antes de que un sistema de IA de alto riesgo sea introducido en el mercado o puesto en servicio, el proveedor deberá asegurarse de que aquel sea sometido a un "procedimiento de evaluación de la conformidad" que verifique el cumplimiento de los requisitos establecidos en los artículos 8-15 del reglamento. De modo que su obligación fundamental es velar por el cumplimiento de los siguientes requisitos, dirigidos a mitigar los riesgos y garantizar un alto nivel de fiabilidad de los sistemas:

a) Establecimiento de un sistema de gestión de riesgos.

b) Implantación de prácticas de gobernanza y gestión de datos adecuadas para la finalidad prevista del sistema.

c) Elaboración y actualización de la documentación técnica del sistema.

d) Conservación del registro automático de acontecimientos a lo largo de todo el ciclo de vida del sistema.

e) Funcionamiento del sistema con un nivel de transparencia suficiente para que los responsables del despliegue interpreten y usen correctamente sus resultados, y comunicación de información a los responsables del despliegue.

f) Diseño y desarrollo del sistema de modo que su funcionamiento pueda ser supervisado de manera efectiva por personas físicas.

g) Aseguramiento de un nivel adecuado de precisión, solidez y ciberseguridad.

Si el procedimiento de evaluación de la conformidad es satisfactorio, el proveedor deberá expedir declaración de conformidad y colocar el marcado CE en el sistema. Asimismo, tendrá que registrarlo en la base de datos de la UE de sistemas de alto riesgo antes de comercializarlo. Y, una vez que ha sido introducido en el mercado o puesto en servicio, establecerá y documentará un sistema de monitorización que le permita tener en cuenta la experiencia de uso del sistema para mejorar su funcionamiento y los procesos de diseño y desarrollo, así como para adoptar las medidas correctoras necesarias en el momento oportuno. Dentro de este seguimiento también se exige al proveedor que comunique cualquier incidente grave en el funcionamiento del sistema a las autoridades de vigilancia del mercado de los Estados miembros en los que se haya producido.

En el caso de los proveedores de modelos de IA de uso general, están sujetos a las siguientes obligaciones (artículo 53):

a) Elaboración y actualización de la documentación técnica del modelo, incluida la información relativa al proceso de entrenamiento y realización de pruebas y los resultados de su evaluación.

b) Puesta a disposición de los proveedores de sistemas de IA que tengan la intención de integrar el modelo de IA de uso general en sus sistemas de información y documentación actualizada que les permita entender las capacidades y limitaciones del modelo

c) Establecimiento de directrices para cumplir el Derecho de la Unión Europea en materia de derechos de autor y derechos afines.

d) Elaboración y puesta a disposición del público de un resumen suficientemente detallado del contenido utilizado para el entrenamiento del modelo.

Además de estas obligaciones generales para todo proveedor de un modelo de IA de uso general, el Reglamento impone una serie de obligaciones adicionales a los proveedores de modelos de IA de uso general con riesgo sistémico encaminadas a detectar y atenuar los riesgos sistémicos, a vigilar y comunicar los posibles incidentes, y a garantizar un nivel adecuado de protección en materia de ciberseguridad (artículo 55).

Responsable del despliegue, *por J. I. Solar Cayón*

El Reglamento de IA de la Unión Europea define al "responsable del despliegue" como aquella persona física o jurídica, o autoridad pública, órgano u organismo que utilice un sistema de IA bajo su propia autoridad, salvo cuando su uso se enmarque en una actividad personal de carácter no profesional.

El proveedor y el responsable del despliegue constituyen, sin duda, las figuras centrales sobre las que pivota el régimen jurídico de los sistemas de IA de alto riesgo, en tanto a ellas se les atribuyen las obligaciones y responsabilidades principales. A estos efectos, la importancia del responsable del despliegue reside en el hecho de que es el agente que está en mejores condiciones para elegir el sistema más adecuado a la finalidad pretendida y hacer un uso correcto del mismo. Aunque los riesgos de los sistemas de IA pueden derivarse de su diseño y desarrollo, también pueden venir motivados por el uso que se haga de ellos, de modo que el responsable del despliegue del sistema desempeña un papel fundamental a la hora de garantizar la protección de los derechos fundamentales de las personas afectadas por sus resultados. Es quien se encuentra en una posición óptima para comprender el uso concreto que se le dará al sistema y quien mejor puede, por tanto, detectar potenciales riesgos significativos que no se previeron en la fase de desarrollo, al tener un conocimiento más preciso del contexto de uso y de las personas o los grupos de personas potencialmente afectados por sus resultados.

Por estas razones el reglamento impone al responsable del despliegue de sistemas de IA de alto riesgo un extenso catálogo de obligaciones. Entre ellas, fundamentalmente, la de determinar las estructuras de gobernanza apropiadas en el contexto específico de uso del sistema, tales como los mecanismos de supervisión humana, los procedimientos de tramitación de reclamaciones y las vías de recurso. Asimismo, en la medida en que ejerza el control sobre los datos

de entrada, se asegurará de que los mismos sean pertinentes y suficientemente representativos en vista de la finalidad prevista del sistema. Por último, desempeña también un papel fundamental a la hora de informar a las personas afectadas. Cuando un sistema de alto riesgo tome decisiones o ayude a tomar decisiones relacionadas con personas físicas, aquel deberá informarlas de que son objeto de la utilización de un sistema de IA de alto riesgo, incluyendo la finalidad prevista y el tipo de decisiones que se toman, y de su derecho a obtener una explicación con arreglo a lo dispuesto en el reglamento.

Por otra parte, el reglamento establece obligaciones específicas de transparencia para los responsables del despliegue de determinados sistemas de IA. En el caso de sistemas de reconocimiento de emociones o de categorización biométrica, los responsables del despliegue deberán informar del funcionamiento del sistema a las personas físicas expuestas a él, salvo cuando su uso haya sido autorizado por ley para detectar, prevenir e investigar delitos. Y cuando se trate de sistemas de IA que generen o manipulen imágenes o contenidos de audio o video que constituyan una ultrasuplantación deberán hacer público que esos contenidos o imágenes han sido generados o manipulados de manera artificial, con la misma excepción que en el caso anterior.

Garantía del cumplimiento del Derecho, *por J. I. Solar Cayón*

El Reglamento de IA (RIA) de la Unión Europea entiende por "garantía del cumplimiento del Derecho" todas aquellas actividades realizadas por las autoridades garantes del cumplimiento del Derecho, o en su nombre, para la prevención, la investigación, la detección o el enjuiciamiento de delitos o la ejecución de sanciones penales, incluidas la protección frente a amenazas para la seguridad pública y la prevención de dichas amenazas. A estos efectos, debe entenderse que constituyen "**autoridad garante del cumplimiento del Derecho**" no solo las autoridades públicas competentes para llevar a cabo las actividades señaladas (policía, fiscalía, autoridad judicial...), sino también cualquier otro organismo o entidad a quien el Derecho del Estado miembro haya confiado el ejercicio de esas competencias públicas.

El RIA prohíbe algunos usos de sistemas de IA gravemente lesivos de los derechos fundamentales en el desarrollo de las actividades de garantía de cumplimiento del Derecho. Este es el caso de la utilización de sistemas para la identificación biométrica remota en tiempo real de personas físicas en espacios de acceso público, si bien se contemplan una serie de situaciones excepcionales, sujetas a determinadas condiciones, en las que se permite su empleo. Tampoco se deben utilizar este tipo de sistemas en diferido en el marco de la garantía

del cumplimiento del Derecho de tal forma que se produzca una vigilancia indiscriminada.

Pero la relevancia fundamental de este concepto en la arquitectura del régimen jurídico implantado por el RIA reside en que constituye uno de los ámbitos definidos en el Anexo III. Fuera de los casos prohibidos, y en la medida en que su uso esté permitido conforme al Derecho de la Unión y nacional pertinente, deberán considerarse de alto riesgo los sistemas utilizados para actividades de garantía de cumplimiento del Derecho cuando su precisión, fiabilidad y transparencia sean especialmente importantes para evitar consecuencias adversas, conservar la confianza de la población y garantizar la rendición de cuentas y unas vías de recurso efectivas. Entre dichos sistemas de alto riesgo deben incluirse, en particular, los destinados a ser utilizados por las autoridades garantes del cumplimiento del Derecho, o en su nombre, o por las instituciones, órganos y organismos de la Unión en apoyo de las autoridades garantes del cumplimiento del Derecho, para los siguientes fines:

a) evaluar el riesgo de que una persona física sea víctima de delitos,

b) su uso como polígrafos o herramientas similares,

c) evaluar la fiabilidad de las pruebas durante la investigación o el enjuiciamiento de delitos,

d) evaluar el riesgo de que una persona física cometa un delito o reincida en la comisión de un delito atendiendo no solo a la elaboración de perfiles de personas físicas o para evaluar rasgos y características de la personalidad o comportamientos delictivos pasados de personas físicas o colectivos,

e) elaborar perfiles de personas físicas durante la detección, la investigación o el enjuiciamiento de delitos.

Reconocimiento facial, *por J. A. Carlin Sánchez*

"Reconocimiento facial" es una expresión utilizada para designar las tecnologías de tratamiento automatizado de imágenes o vídeos del rostro humano destinadas a extraer, analizar o comparar características faciales con fines biométricos. A diferencia de la identificación biométrica, la verificación biométrica, la categorización biométrica o el reconocimiento de emociones, que implican finalidades o funciones reguladas específicamente por el Reglamento (UE) 2024/1689 de IA, el reconocimiento facial no configura una categoría funcional autónoma, sino una técnica susceptible de emplearse para diversas

finalidades reguladas por dicho instrumento. Su relevancia jurídica deriva de la singular capacidad del rostro para ser captado a distancia, sin contacto físico y, en muchos casos, sin conocimiento efectivo de la persona afectada, circunstancia que incrementa los riesgos para la privacidad, la protección de datos personales y el ejercicio de las libertades fundamentales. Por esta razón, el legislador europeo no regula el reconocimiento facial mediante un régimen propio, sino a través de los distintos marcos jurídicos aplicables a la finalidad concreta para la que se emplea. En la práctica, cualquier despliegue de tecnologías de reconocimiento facial exige determinar previamente dicha finalidad, pues de ella dependerán las obligaciones, limitaciones o prohibiciones aplicables conforme al Reglamento de IA, al Reglamento General de Protección de Datos y al resto del Derecho de la Unión aplicable.

Sistema de identificación biométrica, *por J. A. Carlin Sánchez*

La identificación biométrica es definida en el artículo 3, apartado 35, del Reglamento (UE) 2024/1689 de IA como el reconocimiento automatizado de rasgos humanos físicos, fisiológicos, conductuales o psicológicos con el propósito de establecer la identidad de una persona física mediante la comparación de sus datos biométricos con otros datos biométricos de personas almacenados en una base de datos. Se caracteriza por ser una operación típicamente uno a varios (1:N), destinada a determinar la identidad de una persona desconocida dentro de un conjunto de referencias. Su modalidad más sensible es la identificación biométrica remota (*Remote Biometric Identification*, RBI), que opera sin la participación activa del sujeto, generalmente a distancia. El uso de sistemas de identificación biométrica remota en tiempo real, en los que la captura, comparación e identificación se producen sin demora significativa, está prohibido en espacios accesibles al público para fines de garantía del cumplimiento del Derecho (artículo 5.1, letra h), salvo en supuestos excepcionales vinculados a la búsqueda de víctimas, la prevención de amenazas graves e inminentes o la persecución de determinados delitos graves enumerados en el Anexo II. Tales excepciones están sujetas, como regla general, a autorización previa de una autoridad competente y a estrictas garantías temporales, geográficas y materiales. Los sistemas de identificación biométrica remota no comprendidos en la prohibición pueden quedar clasificados como de alto riesgo conforme al Anexo III, activando las obligaciones propias de esta categoría, incluidas las relativas a gestión de riesgos, calidad de los datos, supervisión humana y evaluación de impacto sobre los derechos fundamentales. En el plano jurisprudencial, el TEDH, en *Glukhin v. Rusia* (2023), calificó esta tecnología como especialmente intrusiva y advirtió de su potencial efecto disuasorio sobre las libertades fundamentales, mientras que el TJUE ha exigido estándares particularmente rigurosos de necesidad y proporcionalidad respecto de tratamientos masivos o indiscriminados de datos personales. En la práctica, los

proyectos de identificación biométrica remota, especialmente cuando implican tratamientos a gran escala o contextos sensibles, requieren un análisis previo exhaustivo de necesidad, proporcionalidad y disponibilidad de alternativas menos intrusivas.

Verificación biométrica, *por J. A. Carlin Sánchez*

La verificación biométrica viene definida en el artículo 3, apartado 36, del Reglamento (UE) 2024/1689 de IA como la verificación automatizada y unívoca -incluida la autenticación- de la identidad de personas físicas mediante la comparación de sus datos biométricos con datos biométricos facilitados previamente. Es una operación uno a uno (1:1) de carácter puramente confirmativo: valida si la persona que se presenta es quien afirma ser, sin buscar en una base de datos poblacional. Ejemplos típicos son el desbloqueo de un dispositivo móvil, el control fronterizo automatizado (*e-gates*) o el acceso físico a instalaciones mediante reconocimiento facial o huella dactilar. La verificación biométrica no figura, como categoría autónoma, entre las prácticas prohibidas ni entre las categorías generales de alto riesgo previstas por el Reglamento de IA, siempre que no forme parte de un sistema de IA clasificado como alto riesgo. Parte de la doctrina explica esta posición regulatoria por la naturaleza habitualmente descentralizada de estos sistemas y por el mayor control que, en muchos supuestos, conserva el titular sobre la plantilla biométrica de referencia. En el caso *Schwarz v. Stadt Bochum* (C-291/12), el TJUE consideró proporcionada la incorporación de huellas dactilares en pasaportes biométricos con fines de verificación de identidad, circunstancia que parte de la doctrina considera compatible con un menor impacto sobre la protección de datos cuando existe un elevado grado de control por parte del interesado. Aunque es un régimen más permisivo, sigue plenamente sujeto al Reglamento General de Protección de Datos -exigiendo base legal adecuada, minimización y seguridad-, y en despliegues a gran escala resulta recomendable realizar una Evaluación de Impacto en Protección de Datos (DPIA por sus siglas en inglés, o EIPD en español), garantizando además que la plantilla facial no se almacene centralizadamente de forma innecesaria para evitar el *function creep*.

Sistema de categorización biométrica, *por J. A. Carlin Sánchez*

Un sistema de categorización biométrica viene definido en el artículo 3, apartado 40, del Reglamento de IA como un sistema de IA destinado a asignar personas físicas a categorías específicas sobre la base de sus datos biométricos, a menos que sea auxiliar de otro servicio comercial y estrictamente necesario por razones técnicas objetivas. No persigue identificar (1:N) ni verificar (1:1), sino clasificar o inferir atributos o categorías específicas, como grupos de edad,

determinadas características físicas u otros rasgos inferidos a partir de datos biométricos, sin perjuicio de la regulación específica aplicable a los sistemas de reconocimiento de emociones.

El artículo 5.1, letra g, del reglamento prohíbe de manera absoluta la comercialización, puesta en servicio o uso de sistemas que categoricen individualmente basándose en datos biométricos para deducir o inferir categorías sensibles del artículo 9 del Reglamento General de Protección de Datos (raza, opiniones políticas, creencias religiosas o filosóficas, orientación sexual, etc.), con la única excepción del etiquetado o filtrado de conjuntos de datos adquiridos lícitamente en el ámbito de la garantía del cumplimiento del Derecho. Determinados usos de categorización biométrica pueden quedar comprendidos entre los sistemas de alto riesgo previstos en el Anexo III. La prohibición establecida por el reglamento resulta coherente con las preocupaciones expresadas por el TEDH en *S. and Marper v. Reino Unido* (2008), donde se declaró incompatible con el Convenio Europeo la retención indiscriminada de perfiles biométricos de personas no condenadas debido a sus riesgos estigmatizantes y discriminatorios. En la práctica empresarial resulta altamente recomendable la realización de auditorías de sesgos y evaluaciones de equidad antes de cualquier despliegue de categorización, ya que en contextos de marketing o recursos humanos es frecuente que se cruce la línea hacia lo prohibido.

Sistema de reconocimiento de emociones, *por J. A. Carlin Sánchez*

Viene definido en el artículo 3, apartado 39, del Reglamento de IA como un sistema de IA destinado a identificar o inferir las emociones o las intenciones de las personas físicas a partir de sus datos biométricos. Su funcionamiento se basa en el análisis de expresiones faciales, patrones vocales, gestos u otros indicadores biométricos con el fin de inferir emociones o intenciones atribuidas a la persona analizada. Conforme al Considerando 18 del reglamento, ello no incluye estados físicos como dolor o cansancio, ni la mera detección de expresiones, gestos o movimientos obvios, salvo que se utilicen para inferir emociones o intenciones.

El artículo 5.1, letra f, prohíbe su uso en entornos laborales y educativos (salvo fines médicos o de seguridad). Fuera de estos ámbitos, los sistemas de reconocimiento de emociones están sujetos, como mínimo, a obligaciones específicas de transparencia (artículo 50.3), sin perjuicio de que determinados usos puedan quedar sometidos adicionalmente al régimen aplicable a los sistemas de alto riesgo previstos en el Anexo III, dependiendo de su finalidad y contexto de uso (por ejemplo, policial, judicial, migratorio o de evaluación de credibilidad). Esta arquitectura normativa protege la integridad psíquica, la libertad de pensamiento y la dignidad humana (artículo 1 de la Carta de Derechos Fundamentales de la UE),

evitando formas de vigilancia psicológica automatizada que podrían condicionar el comportamiento. En términos de cumplimiento, los despliegues para los que resulte exigible conforme al artículo 27 deberán incorporar una evaluación de impacto sobre los derechos fundamentales, supervisión humana significativa y documentación de la fiabilidad científica del sistema, prestando especial atención las autoridades supervisoras a posibles usos encubiertos en selección de personal o vigilancia laboral.

Evaluación de la conformidad, *por I. Lledó Benito*

La evaluación de la conformidad viene establecida en el art. 43 del Reglamento europeo de IA (RIA), en el caso de los sistemas de IA de alto riesgo enumerados en el anexo III, punto 1, para demostrar el cumplimiento de los requisitos establecidos en la sección 2.

Es decir, deberá evaluarse la biometría (sistemas de identificación biométrica remota), sistemas de IA destinados a la categorización biométrica de atributos o características sensibles; los sistemas utilizados en la gestión de infraestructuras criticas; en la educación y formación profesional; en el empleo, gestión de los trabajadores y acceso al autoempleo; en el acceso a servicios privados esenciales, públicos o ambos; en la garantía del cumplimiento del Derecho (lo que llevará a analizar las disposiciones UE y la normativa interna de cada Estado); la evaluación de la migración, asilo y gestión del control fronterizo; y en la Administración de Justicia y procesos democráticos (véase a fortiori y detallado el citado anexo III).

Y, como decimos supra, en la sección 2, concretamente en el art. 8, se explicita el cumplimiento de los requisitos (sistemas de alto riesgo) teniendo en cuenta el estado actual de la técnica en IA y tecnologías relacionadas (*vide* también art. 9).

En el contexto de la evaluación de la conformidad debe tenerse en cuenta lo dispuesto en el art. 40 (normas armonizadoras y elementos de normalización), es decir, que se cumplen los requisitos establecidos en la sección 2 y las obligaciones establecidas en el capítulo V, secciones 2 y 3 del Reglamento. Asimismo, en la evaluación de la conformidad deberá cumplirse lo dispuesto en el cap. IX, sección 1, referido a la vigilancia postcomercialización, y la sección 2, relativa al intercambio de información sobre incidentes graves.

Finalmente, es complementario y decisivo el anexo VI del RIA, referido al procedimiento de evaluación de la conformidad, fundamentado en el control interno (sistemas de gestión de la calidad, art. 17, y la información de la documentación técnica para evaluar la conformidad del sistema IA, capítulo III, sección 2).

Y para acabar, es fundamental el anexo VII referido a la conformidad fundamentada en la evaluación del sistema de gestión de la calidad y la evaluación de la documentación técnica.

Organismo de evaluación de la conformidad, *por I. Lledó Benito*

Un organismo de evaluación de la conformidad es un organismo que desempeña actividades de evaluación de la conformidad de terceros, como el ensayo, la certificación y la inspección. Naturalmente, en este punto nos tenemos que remitir a los arts. 28-39 del RIA, concretamente a la sección IV: "Autoridades notificantes y organismos notificados".

Cada Estado miembro establecerá o contará con una autoridad *ad hoc* para evaluar los organismos de evaluación de la conformidad y su supervisión (art. 28 RIA). Así, tenemos que hablar de autoridades de vigilancia de mercado, de la oficina europea de la IA y de las autoridades notificantes.

Naturalmente, los organismos de evaluación de la conformidad deberán contar con certificados de acreditación y cumplir con los requisitos establecidos en el reglamento (gestión de la calidad, estructura organizativa, etcétera; véase art. 31 RIA).

Cuando un organismo de evaluación de la conformidad cumpla con los criterios establecidos en las normas armonizadoras publicadas en el Diario oficial de la UE, se presumirá que cumple con los requisitos del art. 31 (véase *a fortiori* art. 32).

Es importante destacar que existen organismos de evaluación de la conformidad de terceros países que tengan acuerdos con la UE y que podrán desempeñar las actividades de los organismos notificados con arreglo al Reglamento, si cumplen con los requisitos del art. 31 (véase art. 39 RIA).

Finalmente, volvemos a remitir al anexo III (referido a los sistemas de IA de alto riesgo a que se refiere el art. 6, apartado 2). Del mismo modo, al anexo VI, referido al procedimiento de evaluación de la conformidad fundamentado en un control interno. Y al anexo VII, concerniente a la conformidad fundamentada en la evaluación del sistema de gestión de la calidad y la evaluación de la documentación técnica.

Autoridad notificante, *por I. Lledó Benito*

Autoridad notificante es la autoridad nacional responsable de establecer y llevar a cabo los procedimientos necesarios para la evaluación, designación y notificación de los organismos de evaluación de la conformidad, así como de su supervisión. El RIA se refiere a las autoridades notificantes en su capítulo III, sección 4. En realidad, no se explicita la naturaleza del cargo, simplemente se generaliza con la descripción de la remisión a una autoridad que será la encargada de los procesos de evaluación, designación y notificación de los organismos de evaluación de la conformidad (*vide* art. 28 RIA).

El precepto es una declaración de intenciones, no es preciso, clarificador, y sí, por el contrario, excesivamente generalista. En realidad, su operatividad y eficacia va a depender de que funcionen y cumplan sus objetivos los citados organismos de la conformidad. Para evitar conflictos y presentar la imparcialidad y la objetividad, los miembros designados serán distintos de las personas que llevarán a cabo la evaluación de dichos organismos (art. 28, 4 RIA).

Las autoridades notificantes no ejercerán ninguna actividad que efectúen los organismos de evaluación de la conformidad. Tienen que ser profesionales expertos en IA. Y llama la atención que, al referirse a su experiencia jurídica, explicite (más de una forma tautológica que efectiva) la "inclusión de la supervisión de los derechos fundamentales". Pero si son expertos en Derecho ¿para qué esa precisión? (*vide* art. 28,7 RIA).

Es digno de mención -en el procedimiento de notificación por parte del organismo de evaluación de la conformidad (en el que se describirán las actividades que son competentes)- que deberán adjuntar un certificado de acreditación y no se nos clarifica qué autoridad es la competente, simplemente se remite a una expresión generalista referida a un "organismo nacional de acreditación". ¿Cuál? ¿Qué naturaleza y dependencia jerárquica? No debe ser un requisito constitutivo, ya que el propio Reglamento deja la puerta abierta a una permisibilidad laxa de pruebas documentales suficientes que cumplan lo exigido en los arts. 29 y 31 RIA.

No queremos ser críticos en exceso, pero creemos que el texto del Reglamento no es muy claro, ya que, si por un lado, hablamos de certificado de acreditación, también exige un certificado de competencia pertinente (art. 30-3), solapando las funciones y exigencias del certificado de acreditación, y, por otro lado, para terminar realizando las actividades de un organismo notificado. El plazo no es el mismo si hay o no certificado de acreditación o simples pruebas documentales (art. 30-4 RIA, UE).

En fin, un caos burocrático que se apoyará más en las "buenas prácticas" que en el peso e imperio de la ley.

Organismo notificado, *por I. Lledó Benito*

La definición de organismo notificado la encontramos en el art. 3 del RIA, concretamente en el apartado 22, que dice que es un organismo de evaluación de la conformidad notificado con arreglo al presente Reglamento y a otros actos pertinentes de la legislación de armonización de la Unión.

En el capítulo III, sección IV, en el art. 31, encontramos los requisitos relativos a los organismos notificados. Destacamos que se establecerán conforme con el Derecho nacional y tendrán personalidad jurídica. Asimismo, deberán tener gestión de calidad para sus funciones y cumplir con los requisitos de ciberseguridad. Del mismo modo, tendrán estructura organizativa, línea jerárquica y responsabilidad en los resultados de las actividades de evaluación de la conformidad que realicen los organismos notificados (art. 31 RIA, apartados 2 y 3).

Es importante destacar que los organismos notificados verificarán la conformidad de los sistemas de IA de alto riesgo (art. 34 RIA). Asimismo, hay que remitirse al art. 43 para la evaluación de la conformidad y al anexo III, punto 1, para la verificación de la conformidad de los sistemas de IA de alto riesgo.

Es oportuno destacar, en referencia a los organismos notificados (art. 34, apartado 2), algo que realmente debería ser expresado al revés de lo transcrito, ya que su tenor literal aligera de cargas y burocracias administrativas a los proveedores de sistemas de IA de alto riesgo. Considerando su complejidad pueden ser laxos en la gestión de la rendición de cuentas. Así, literalmente, dice: "los organismos notificados evitarán cargas innecesarias para los proveedores cuando desempeñen sus actividades…"; y continúa: "tendrán debidamente en cuenta el tamaño del proveedor, sectores que opera, su estructura y el grado de complejidad del sistema de IA con vistas a reducir al mínimo las cargas administrativas y los costes de cumplimiento para las microempresas y pequeñas empresas". A nuestro juicio debería ser precisamente lo contrario, exigencia al máximo de todos los requisitos y cumplimiento de las cargas administrativas. No verlo así nos puede llevar a escenarios de discrecionalidad notables, que en su caso pueden ser arbitrarios.

Ahora bien, puede mitigarse el efecto por la obligación de los organismos notificados de poner a disposición de la autoridad notificante la documentación de los proveedores (art. 34,3 RIA). Y, asimismo, porque la autoridad notificante facilitará a la Comisión (a petición de esta, no por voluntad de la autoridad) toda la información sobre la competencia del organismo notificado de que se trate

(art. 37,2 RIA). Y en este sentido se manifiesta el art. 45 en torno a las obligaciones de información de los organismos notificados.

Supervisión humana, *por F. H. Llano Alonso*

Por supervisión humana se entiende el conjunto de mecanismos y garantías que permiten que las personas conserven el control sobre el funcionamiento de los sistemas de IA, de modo que la tecnología asista a la decisión humana sin llegar a sustituirla. La expresión designa, en sentido amplio, la presencia activa y responsable de seres humanos en los procesos automatizados, en especial en aquellos que pueden afectar a la salud, a la seguridad o a los derechos fundamentales de los ciudadanos. No se trata, por tanto, de una mera presencia simbólica, sino de una capacidad efectiva de comprender, vigilar y, en su caso, corregir lo que la máquina hace.

La noción descansa en una convicción de fondo: por muy avanzado que sea un sistema, su comportamiento no es infalible ni neutral, pues depende de los datos con los que se ha entrenado y de los criterios con los que se ha programado. De ahí que la supervisión humana persiga prevenir o reducir los errores, las anomalías y los sesgos que la máquina puede generar, y ofrecer a la persona afectada la garantía de que su caso no quedará resuelto únicamente por un procedimiento automatizado. Detrás de esta exigencia late también una idea de responsabilidad: las consecuencias de una decisión que afecta a la vida de alguien deben poder atribuirse a un sujeto humano y no diluirse en el funcionamiento opaco de un algoritmo.

Para que la supervisión sea real y no meramente formal, quien la ejerce debe hallarse en condiciones de comprender las capacidades y los límites del sistema, interpretar correctamente los resultados que produce y conservar la facultad de apartarse de ellos. Ello supone poder no utilizar el sistema, descartar o revertir sus salidas e incluso interrumpir su funcionamiento mediante un mecanismo de parada seguro. Para que esa intervención sea posible, el sistema ha de presentarse con una interfaz comprensible y con información clara sobre su modo de operar. La supervisión, además, ha de ser proporcionada: cuanto mayores sean la autonomía del sistema y la gravedad de los riesgos, más intensa deberá ser la intervención humana.

La doctrina distingue distintos grados de implicación humana según la naturaleza de la intervención. En el modelo más exigente, la persona permanece dentro del proceso de decisión y la máquina se limita a apoyarla o a automatizar parcialmente alguna tarea, de manera que ninguna salida produce efectos sin un juicio humano previo. En otros modelos, el ser humano se reserva un papel de simple vigilancia o de revisión posterior, e interviene solo cuando detecta un

problema o una situación excepcional. La fórmula que mejor expresa el ideal de control humano es aquella en la que el juicio último corresponde siempre a la persona, pues solo entonces cabe hablar con propiedad de una tecnología puesta al servicio de quien la utiliza.

Un riesgo característico que la supervisión trata de conjurar es el llamado sesgo de automatización, esto es, la inclinación a confiar de manera acrítica en lo que produce la máquina por el solo hecho de proceder de un sistema técnico. Este fenómeno resulta especialmente delicado cuando el sistema ofrece informaciones o recomendaciones que orientan decisiones humanas, ya que bajo una apariencia de objetividad puede arrastrar errores o discriminaciones. La supervisión reclama, por ello, un conocimiento suficiente de la tecnología, lo que enlaza con la necesidad de educar a la ciudadanía y de formar a los profesionales que han de manejar estos sistemas. Una supervisión que se limita a refrendar mecánicamente lo que propone la máquina vacía de contenido la garantía y reproduce, sin advertirlo, los defectos que pretendía evitar.

En última instancia, la supervisión humana traduce un principio antropocéntrico: la tecnología debe adaptarse a las necesidades de la persona y permanecer a su servicio, nunca al contrario. Constituye, así, una garantía de que la incorporación de la IA a ámbitos sensibles se produce sin desplazar a quien debe seguir siendo, en definitiva, el responsable último de las decisiones.

En el Derecho europeo, esta exigencia encuentra su expresión más acabada en el artículo 14 del Reglamento de IA, que impone la supervisión humana como requisito de los sistemas de alto riesgo y la concreta en facultades como la de intervenir en su funcionamiento o detenerlo de forma segura.

Ultrasuplantación (*Deepfake*), *por J. D. Ayllón García*

1. Origen etimológico y evolución terminológica

El término anglosajón *deepfake* es un acrónimo originado por la contracción de los vocablos *deep learning* (aprendizaje profundo) y *fake* (falso). En la doctrina jurídica inicial, este fenómeno fue denominado con expresiones como «falsedades profundas» o «vídeos hiperrealistas». Sin embargo, la terminología técnica consolidada en la Unión Europea para los profesionales jurídicos es «ultrasuplantación». Es la nomenclatura oficial adoptada en la versión española del Reglamento (UE) 2024/1689 de Inteligencia Artificial (RIA). Como curiosidad terminológica, durante la tramitación del RIA el legislador europeo iba a emplear el término «ultrafalsificación». En la revisión final del texto comunitario se modificó por estrictas cuestiones de traducción, optando por ultrasuplantación, motivo por el cual todavía es frecuente encontrar el término ultrafalsificación en

diversos artículos doctrinales o en normas estatales que aún se encuentran en fase de tramitación, haciendo referencia, en todo caso, a la misma realidad tecnológica de los *deepfakes*. Por tanto, en el Derecho Digital, ambos términos son sinónimos.

2. Definición normativa y conceptual

Desde una perspectiva estrictamente jurídica, el artículo 3.60 del RIA define la ultrasuplantación como «un contenido de imagen, audio o vídeo generado o manipulado por una IA que se asemeja a personas, objetos, lugares, entidades o sucesos reales y que puede inducir a una persona a pensar erróneamente que son auténticos o verídicos». En esencia, es una técnica de manipulación digital basada en IA que permite la creación de contenido sintético hiperrealista. En dicho contenido, los rasgos o la voz se alteran para que el espectador perciba como veraz una situación física inexistente.

3. Fundamento técnico y algorítmico

Su creación se fundamenta en algoritmos de aprendizaje profundo (*Deep Learning*) como las Redes Generativas Antagónicas (GAN). Aquí compiten dos redes neuronales: un modelo generativo que crea imágenes sintéticas y uno discriminativo que intenta detectar si son falsas. Mediante iteraciones, el generador aprende a crear falsificaciones que el discriminador no distingue de la realidad. Este proceso requiere tres etapas: extracción previa del rostro de la persona original, procesamiento del rostro falso en la red neuronal e inserción del elemento digital en el fotograma o pista de audio. En este ciclo interactúan diversos actores jurídicos: víctima, autor material, audiencia, proveedor de tecnología y plataforma.

4. Modelos de aplicación técnica

Dentro de este concepto algorítmico, la práctica actual abarca cuatro categorías principales:

- *Face Swapping*: Intercambio automatizado de rostros en imagen o vídeo, integrando de forma realista las expresiones.

- *Voice Cloning:* Clonación de la voz, que replica el tono, la entonación y el ritmo a partir de un fragmento original.

- *Lip-Sync:* Manipulación de movimientos labiales de un sujeto para hacerlos coincidir con un audio falso.

- *Full-body:* Recreación y manipulación de movimientos corporales completos, generando una ilusión física íntegra.

5. Transparencia legal y detección forense

Girando en torno a la dimensión legal del concepto, la IA de ultrasuplantaciones exige la estricta obligación de transparencia. El RIA impone a los creadores marcar de forma clara y legible por máquina que el contenido fue generado artificialmente. Aunque la tecnología simule la realidad, el Derecho exige identificar su naturaleza sintética. Para detectar *deepfakes* fraudulentos que omiten intencionadamente esta marca imperativa, los peritos han de buscar "residuos" algorítmicos. Estos defectos técnicos son incongruencias en la piel (muy lisa o arrugada), sombras apagadas alrededor de los ojos, carencia de parpadeo y/o un color de labios totalmente discordante con el rostro real de la persona suplantada, entre otros.

6. Diferenciación de conceptos afines: *Fake News y Shallowfakes*

Es crucial no confundir la ultrasuplantación con otros medios de desinformación. Se distingue de las *Fake News* (noticias falsas), ya que estas se basan en formatos textuales que distorsionan la realidad mediante la palabra, mientras que las ultrasuplantaciones lo hacen mediante manipulaciones audiovisuales generadas por IA, incrementando su persuasión. Por otro lado, difiere de los *Shallowfakes*, que son imágenes o vídeos engañosos manipulados sin usar aprendizaje profundo. Los *shallowfakes* emplean edición convencional, como ralentizar o acelerar un vídeo para cambiar el sentido de la acción, usando tecnología antigua pero muy convincente.

***Sandbox* o Espacio controlado de pruebas**, *por P. Contreras Vásquez*

Un *sandbox* es una técnica regulatoria que permite, autoriza o facilita actividades de experimentación. La expresión "*sandbox*" es de origen anglosajón y literalmente significa "caja de arena", aludiendo al espacio en el que los niños juegan, ensayan y aprenden a interactuar con otros.

Es considerado un banco de prueba o espacio controlado de pruebas para que las organizaciones puedan hacer pilotos de prototipos, actividades, modelos o sistemas con el fin de facilitar o promover la innovación técnica orientada al cumplimiento de las obligaciones legales o reglamentarias. Es un instrumento de regulación experimental que crea un espacio acotado de diálogo entre el regulador y el innovador. Por ello, permite generar buenas prácticas y guías técnicas basadas en la evidencia.

En materia de IA, un *sandbox* es un entorno controlado que fomenta la innovación y facilita el desarrollo, el entrenamiento, la prueba y la validación de sistemas innovadores de IA, durante un período limitado, antes de su introducción

en el mercado o su puesta en servicio, con arreglo a un plan del espacio controlado de pruebas específico acordado entre los proveedores o proveedores potenciales y la autoridad competente (art. 57.5 del Reglamento (UE) 2024/1689). El modelo se inspira directamente en los sandboxes del sector financiero (*Fintech*), cuya experiencia de cooperación regulatoria se ha trasladado al ámbito de la IA.

Este tipo de espacios se fundamentan en la necesidad de acompasar la regulación con el avance de la tecnología. En general, el Derecho tiene un problema de obsolescencia regulatoria frente al cambio tecnológico, particularmente cuando sus enunciados son precisos y específicos y estamos ante desarrollos de tecnologías disruptivas. A efectos de conciliar el fomento de la innovación con determinados intereses estatales -como la protección de los derechos fundamentales, la salud o la seguridad de la población- se promueve un aprendizaje regulatorio mutuo en el que las autoridades comprenden mejor la tecnología y los participantes alcanzan seguridad jurídica.

Un *sandbox* de IA tiene distintos objetivos (art. 57.9 del Reglamento). En primer término, busca ofrecer seguridad jurídica para cumplir el propio Reglamento y las demás normas de la UE y de cada Estado, así como apoyar el intercambio de mejores prácticas mediante la cooperación entre las autoridades participantes. Asimismo, busca fomentar la innovación, la competitividad y el desarrollo de un ecosistema europeo de IA, y contribuir a un aprendizaje normativo basado en datos contrastados. Por último, pretende facilitar y acelerar el acceso al mercado de la Unión de los sistemas de IA, en particular por pymes y empresas emergentes.

Dentro de las características de un *sandbox* encontramos la supervisión por parte de la autoridad y el carácter limitado en su objeto y en el tiempo, bajo un plan de la autoridad. En función de ello, por ejemplo, la autoridad puede suspender temporal o permanentemente el proceso de prueba. Segundo, su acceso es voluntario y está abierto a cualquier proveedor de IA, incluso los potenciales. La participación en un *sandbox* no exime de responder de cualquier daño infligido a terceros como resultado de la experimentación realizada. Sin embargo, el Reglamento detalla que siempre que los proveedores potenciales respeten el plan específico y las condiciones de su participación y sigan de buena fe las orientaciones proporcionadas por la autoridad nacional competente, las autoridades no impondrán multas administrativas.

Existen diversas modalidades de *sandbox*. Cada Estado debe establecer un *sandbox* nacional. Además, se admite el *sandbox* conjunto, creado por varios Estados miembros de la UE, así como los de ámbito regional o local. A nivel de la Unión, el Supervisor Europeo de Protección de Datos puede crear un *sandbox* para las instituciones, órganos y organismos europeos (art. 57).

Algunos problemas normativos surgen de la tensión estructural entre el fomento de la innovación y la protección de los derechos fundamentales, que puede verse amenazada si lo primero se prioriza en exceso en desmedro de lo segundo. Desde la vereda de la innovación, la flexibilidad real del *sandbox* es limitada, ya que no habilita exenciones o eximiciones normativas sino orientación.

Perfilado, *por F. Bariffi*

El perfilado constituye una de las categorías centrales del Derecho europeo de protección de datos y, por extensión, del régimen jurídico de la IA. El artículo 4.4 del Reglamento General de Protección de Datos (RGPD) lo define como toda forma de tratamiento automatizado de datos personales orientada a evaluar determinados aspectos de una persona física -rendimiento profesional, situación económica, salud, preferencias, intereses, fiabilidad, comportamiento, ubicación o movimientos-. Tres elementos cumulativos configuran su tipo legal: la automatización del tratamiento, el uso de datos personales y una finalidad evaluativa o predictiva. No obstante, desde una perspectiva material, el perfilado no se agota en esta definición, sino que debe comprenderse como una técnica de producción de conocimiento sobre la persona que opera mediante procesos inferenciales y probabilísticos.

En efecto, uno de los principales desarrollos doctrinales recientes ha consistido en desplazar el foco desde los datos hacia las inferencias. El problema jurídico del perfilado no reside únicamente en la recogida y tratamiento de datos, sino en la generación de información nueva -no proporcionada directamente por el sujeto- a través de modelos predictivos y correlaciones estadísticas. El Derecho de protección de datos regula de forma intensiva los datos de entrada, pero carece de instrumentos suficientes para controlar las inferencias producidas, pese a que son estas las que determinan los efectos materiales sobre las personas. El perfilado se presenta así como una cadena compleja que va desde el dato hasta la decisión, pasando por fases de inferencia, clasificación y puntuación, en las que se construyen representaciones del sujeto que pueden resultar determinantes para su posición social y jurídica.

El perfilado debe distinguirse de la decisión automatizada, recogida en el artículo 22 RGPD. Mientras que el perfilado puede operar sin desembocar en una decisión, el artículo 22 reconoce el derecho a no ser objeto de decisiones basadas únicamente en tratamiento automatizado cuando produzcan efectos jurídicos o afecten significativamente al interesado. Sin embargo, la doctrina ha advertido que esta garantía presenta límites estructurales, tanto por su ámbito de aplicación restringido como por la amplitud de sus excepciones. La jurisprudencia del Tribunal de Justicia de la Unión Europea ha contribuido a una interpretación

material de esta disposición. En el asunto *SCHUFA* (C-634/21, 7 de diciembre de 2023), el Tribunal declaró que la generación automatizada de un valor de solvencia constituye decisión individual automatizada en el sentido del artículo 22.1 RGPD cuando un tercero le atribuye papel determinante en la relación contractual; en *Dun & Bradstreet Austria* (C-203/22, 27 de febrero de 2025) afirmó el derecho del interesado a recibir una explicación comprensible del resultado y encomendó a la autoridad o al órgano jurisdiccional competente la ponderación entre ese derecho y la protección de secretos comerciales o derechos de terceros, sin que estos últimos puedan vaciarlo de contenido. Estas decisiones reflejan una creciente preocupación por el impacto del perfilado en la esfera jurídica de las personas.

El Reglamento de IA introduce una capa adicional de regulación. Conforme al último párrafo de su artículo 6.3, los sistemas del anexo III pueden excepcionalmente no ser considerados de alto riesgo cuando no entrañen un riesgo significativo para los derechos fundamentales; sin embargo, esta cláusula de exclusión se ve fuertemente limitada cuando dichos sistemas llevan a cabo elaboración de perfiles de personas físicas, en la medida en que esta funcionalidad dificulta justificar la ausencia de riesgo significativo y, en la práctica, conduce a su calificación como de alto riesgo, activando obligaciones reforzadas en materia de gestión del riesgo, gobernanza de datos y supervisión humana. Asimismo, prohíbe prácticas como la puntuación social o ciertos sistemas de predicción de delitos basados en perfilado, especialmente cuando no se fundamentan en datos conductuales directamente vinculados a la actividad delictiva, reconociendo así que estas técnicas pueden generar formas especialmente intensas de intervención sobre la persona.

Más allá de su dimensión normativa, el perfilado plantea un problema estructural de igualdad y de poder. La conversión del individuo en objeto de anticipación estadística desplaza el centro de gravedad desde la conducta efectiva hacia la correlación inferida, produciendo efectos sobre la distribución de oportunidades, reputación y vulnerabilidad. Este fenómeno ha sido analizado en términos de discriminación probabilística, en la medida en que los sistemas de perfilado tienden a reproducir y amplificar desigualdades preexistentes a través de datos sesgados o no representativos. En este sentido, el perfilado no solo evalúa a la persona, sino que contribuye a configurarla, reforzando patrones estructurales de exclusión, especialmente en contextos de interseccionalidad.

Desde una perspectiva más amplia, el perfilado afecta a bienes jurídicos fundamentales como la dignidad, la autonomía y la igualdad sustantiva. Al generar representaciones algorítmicas de la persona, condiciona la forma en que esta es percibida y tratada en distintos ámbitos sociales, desde el acceso al crédito o al empleo hasta la participación en espacios digitales. Como ha señalado T. Ploug, el perfilado basado en datos públicos puede erosionar la autonomía individual

al permitir formas de influencia y control social que escapan al conocimiento y control del sujeto. De este modo, el problema no es solo de privacidad, sino de control sobre la propia identidad y sobre los procesos de formación de la voluntad.

El marco internacional de derechos humanos refuerza esta lectura. Instrumentos como la *Convención sobre los Derechos del Niño* y la *Convención sobre los Derechos de las Personas con Discapacidad* -desarrollados en este punto por la Observación General nº 25 del Comité de los Derechos del Niño y por las Observaciones Generales nº 1 y nº 6 del Comité sobre los Derechos de las Personas con Discapacidad- exigen una protección reforzada frente a prácticas que puedan afectar desproporcionadamente a grupos vulnerables, especialmente cuando estas prácticas se basan en sistemas automatizados opacos y potencialmente discriminatorios.

Alfabetización digital, *por I. García del Blanco*

Se define la alfabetización, desde hace siglos, como la capacidad básica de leer y escribir. Este concepto ha ido evolucionando en los últimos tiempos, como consecuencia del aumento de la complejidad social, en gran medida derivada de la digitalización, y la necesidad de conocimientos más amplios para un mínimo desenvolvimiento personal. Así, el Objetivo de Desarrollo Sostenible 4 de la UNESCO afirma que "más allá de su concepto convencional como conjunto de competencias de lectura, escritura y cálculo, la alfabetización se entiende hoy día como un medio de identificación, comprensión, interpretación, creación y comunicación en un mundo cada vez más globalizado y digitalizado, rico en información y en rápido y permanente cambio. La alfabetización mediática y digital agrega complejidad a estos aprendizajes, haciendo necesarios nuevos saberes que nos permitan ejercer nuestros derechos y deberes ciudadanos".

Como se ve, la alfabetización cada vez se acerca más al concepto clásico de ilustración, que requiere de esfuerzo de superación y un compromiso personal con el conjunto del género humano, que a una mera recepción desequilibrada de instrucción de arriba abajo. Como decía Kant: "La ilustración es la liberación del hombre de su culpable incapacidad. La incapacidad significa la imposibilidad de servirse de su inteligencia sin la guía de otro. Esta incapacidad es culpable, porque su causa no reside en la falta de inteligencia sino de decisión y valor para servirse por sí mismo de ella sin la tutela de otro. *¡Sapere aude!* ¡Ten el valor de servirte de tu propia razón! he aquí el lema de la ilustración".

Pero, además, es importante tener en cuenta que siendo el sistema binario el lenguaje de las máquinas, es el inglés el lenguaje principal de los humanos que investigan, sobre y con esas máquinas, en una abrumadora desproporción. Y esa desproporción aún es mayor en lo que se refiere a lo digital y a la IA.

Los lenguajes conforman estructuras de pensamiento y su proyección simbólica no se puede simplemente contener en una traducción, como nos enseña la semiología. Así, cuando se trata de contener en el lenguaje una nueva realidad digital, casi siempre se hace primero en inglés. Es por ello que expresiones como *enlightment*, *literacy* o *digital alfabetization* no son exactamente trasladables al idioma español como "ilustración" o "alfabetización". Y, desde luego, es complicado encontrar algún ensayo que hable de *digital enlightment*, reduciéndose el trabajo a *digital literacy* o *digital alfabetization*.

Ante esto, podemos optar por mantenernos en nuestra "trinchera de lenguaje ilustrada" o rendirnos ante los hechos y simplemente cargar más el concepto de alfabetización digital, como hace ya la UNESCO en sus textos en español, que define como "la capacidad de acceder, gestionar, comprender, integrar, comunicar, evaluar y crear informaciones mediante la utilización segura y pertinente de las tecnologías digitales para el empleo, un trabajo decente y la iniciativa empresarial. Esto incluye competencias como la alfabetización informática, la alfabetización en las TIC, la alfabetización informativa y la educación mediática, que tienen como objetivo empoderar a las personas y, en particular, a los jóvenes, para que adopten una actitud crítica en cuanto a la utilización de las tecnologías de la información y las tecnologías digitales, y para que puedan desarrollar su resiliencia frente a la desinformación, el discurso de odio y el extremismo violento".

Largo camino desde la original capacidad lectoescritora hasta aquí…

El Reglamento europeo de IA habla en su versión original en inglés de *digital literacy*, lo que supone un gran avance conceptual desde la mera "alfabetization":

"*'AI literacy' means skills, knowledge and understanding that allow providers, deployers and affected persons, taking into account their respective rights and obligations in the context of this Regulation, to make an informed deployment of AI systems, as well as to gain awareness about the opportunities and risks of AI and possible harm it can cause*" (Art. 3, 56).

Aquí ya podemos comprobar que el muy competente cuerpo de juristas-lingüistas de la UE, ha traducido al texto en español *AI literacy* no como "ilustración digital en IA", más cercano a su espíritu original, sino como "alfabetización digital", optando así por añadir contenido adicional al concepto en español, como antes decíamos:

"«Alfabetización en materia de IA»: las capacidades, los conocimientos y la comprensión que permiten a los proveedores, responsables del despliegue y demás personas afectadas, teniendo en cuenta sus respectivos derechos y obligaciones en el contexto del presente Reglamento, llevar a cabo un despliegue informado de los

sistemas de IA y tomar conciencia de las oportunidades y los riesgos que plantea la IA, así como de los perjuicios que puede causar".

La norma establece después una obligación de *AI literacy*, dirigida a los desplegadores de sistemas de IA en su artículo 4 (por más que la obligación se haya "descafeinado" tras la probable aprobación del paquete "Omnibus Digital", que presentó la Comisión Europea en noviembre de 2025):

"*Providers and deployers of AI systems shall take measures to ensure, to their best extent, a sufficient level of AI literacy of their staff and other persons dealing with the operation and use of AI systems on their behalf, taking into account their technical knowledge, experience, education and training and the context the AI systems are to be used in, and considering the persons or groups of persons on whom the AI systems are to be used*".

Y de nuevo se traduce por "obligación de alfabetización en materia de IA":

Artículo 4: "La norma impone tanto a proveedores como a responsables de despliegue la obligación de adoptar medidas para garantizar que, en la mayor medida posible, su personal y demás personas que se encarguen en su nombre del funcionamiento y la utilización de sistemas de IA tengan un nivel suficiente de alfabetización en materia de IA, teniendo en cuenta sus conocimientos técnicos, su experiencia, su educación y su formación, así como el contexto previsto de uso de los sistemas de IA y las personas o los colectivos de personas en que se van a utilizar dichos sistemas".

Así que, como hemos podido comprobar, el significado de "alfabetización digital" está sufriendo una suerte de inflación y acercándose a lo que más clásicamente entenderíamos como ilustración (digital). De hecho, no es difícil volver a Kant y encuadrar la digitalización en su intemporal (la marca de los genios) ensayo sobre la ilustración:

"¡Es tan cómodo no estar emancipado! Tengo a mi disposición un libro que me presta su inteligencia, un cura de almas que me ofrece su conciencia, un médico que me prescribe las dietas, etc., etc., así que no necesito molestarme. Si puedo pagar no me hace falta pensar: ya habrá otros que tomen a su cargo, en mi nombre, tan fastidiosa tarea. (…) Es, pues, difícil para cada hombre en particular lograr salir de esa incapacidad, convertida casi en segunda naturaleza. Le ha cobrado afición y se siente realmente incapaz de servirse de su propia razón, porque nunca se le permitió intentar la aventura. Principios y fórmulas, instrumentos mecánicos de un uso o más bien abuso, racional de sus dotes naturales, hacen veces de ligaduras que le sujetan a ese estado".

¿No es realmente asombrosa la frescura con la que este texto de finales del siglo XVIII encuadra nuestra actual relación con lo digital? Y no es que se trate de saber programar, sino de comprender lo que comporta la digitalización, tanto para uno como para el cuerpo social. No se trata de enunciar un "*aude programmare*", sino de "*sapere aude*" una vez más.

Derecho a explicación, *por M. A. Adame Martínez*

El derecho a explicación es el derecho que asiste a toda persona física a obtener, de quien adopta una decisión que le afecta significativamente mediante un proceso automatizado, una explicación comprensible sobre la lógica aplicada, los factores determinantes y las consecuencias previsibles de dicha decisión. Es uno de los derechos subjetivos de nueva generación que emergen de la intersección entre el Derecho de protección de datos y la regulación de la IA.

Su consagración más explícita en el Derecho de la UE se encuentra en el art. 22 del Reglamento General de Protección de Datos (Reglamento UE 2016/679), que prohíbe con carácter general las decisiones basadas únicamente en el tratamiento automatizado que produzcan efectos jurídicos o afecten significativamente al interesado. En los supuestos en que la decisión automatizada es lícita, el interesado tiene derecho a obtener intervención humana, a expresar su punto de vista y a impugnar la decisión (art. 22.3 RGPD), lo que implica necesariamente que se le expliquen los criterios aplicados. El Reglamento de IA (UE) 2024/1689 refuerza esta dimensión mediante su art. 86, titulado precisamente "Derecho a explicación de decisiones tomadas individualmente", que configura una garantía específica y subsidiaria de explicación aplicable a determinados supuestos de decisiones basadas en sistemas de IA de alto riesgo.

Toda persona afectada por una decisión del responsable del despliegue basada en los resultados de un sistema de alto riesgo incluido en el Anexo III -con excepción de los sistemas de su punto 2- que produzca efectos jurídicos o le afecte de forma igualmente considerable, cuando considere que tiene un efecto perjudicial para su salud, su seguridad o sus derechos fundamentales, tendrá derecho a obtener explicaciones claras y significativas sobre el papel del sistema en el proceso decisional y los principales elementos de la decisión adoptada. El propio art. 86.3 precisa que el precepto solo se aplica en la medida en que ese derecho no esté ya previsto por otro instrumento del Derecho de la Unión, lo que subraya su carácter subsidiario y acotado.

Cabe señalar que la existencia de un derecho a explicación *autónomo* derivado del art. 22 RGPD ha sido objeto de debate doctrinal. Parte de la literatura especializada (con referencia de autoridad en Wachter, Mittelstadt y Floridi, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the

General Data Protection Regulation", 2017) sostiene que el RGPD no consagra propiamente un derecho subjetivo a la explicación, sino un derecho a la revisión humana dentro del cual la explicación opera como elemento instrumental. La interpretación actualmente mayoritaria, asumida por el Comité Europeo de Protección de Datos en sus directrices sobre decisiones individuales automatizadas y elaboración de perfiles (WP251rev01, adoptadas por el GT29 en 2018 y refrendadas por el CEPD), considera, sin embargo, que el conjunto de garantías del art. 22 comporta un derecho a la explicación exigible de forma efectiva.

Debe distinguirse con precisión de otros conceptos afines recogidos en este glosario. El *principio de explicabilidad* es un criterio de diseño de los sistemas, de naturaleza ética y preventiva. La *transparencia algorítmica* es una obligación general de apertura e inteligibilidad impuesta al proveedor o al responsable del despliegue frente al regulador y al público. El derecho a explicación es, en cambio, un derecho subjetivo del individuo afectado, de naturaleza procesal, exigible frente a quien adopta la decisión en el marco de una relación jurídica concreta y respecto de un acto individualizado.

La principal tensión de este derecho reside en la arquitectura de los modelos más avanzados: los grandes modelos de lenguaje y las redes neuronales profundas son, en mayor o menor medida, sistemas de caja negra cuyo proceso decisional no es transparente ni siquiera para sus propios creadores. Las explicaciones que estos sistemas pueden generar son frecuentemente *post hoc* y aproximativas, lo que plantea si satisfacen efectivamente el estándar de "explicación significativa" exigido por el RGPD. El CEPD y el Supervisor Europeo de Protección de Datos -que ha abordado la cuestión en el marco de sus dictámenes sobre el Reglamento de IA- han tratado este problema sin que exista aún una respuesta uniforme, si bien coinciden en que las explicaciones generadas automáticamente deben ser verificables, comprensibles para el destinatario y suficientemente informativas sobre los factores que determinaron el resultado.

Marcado de agua (*Watermarking*) y Etiquetado (*Labelling*) de contenidos IA*, por M. A. Adame Martínez*

El marcado de agua digital (*watermarking*) es una técnica que consiste en incorporar, de forma imperceptible para el usuario, una señal o firma oculta en un contenido generado por IA -texto, imagen, audio o vídeo- con el fin de permitir su identificación posterior como producto de un sistema de IA. El etiquetado (*labelling*), por su parte, es la indicación explícita y visible de que un determinado contenido ha sido generado o manipulado por IA. Ambas técnicas responden al mismo objetivo: garantizar la trazabilidad y la transparencia de los contenidos

generados por IA, aunque operan por vías distintas y tienen alcances jurídicos diferenciados.

El Reglamento de IA (UE) 2024/1689 establece obligaciones específicas en este ámbito. Los proveedores de sistemas de IA con capacidad de generar contenido sintético de audio, imagen, vídeo o texto -incluidos los modelos de uso general- deben asegurar que los resultados de salida estén marcados en un formato legible por máquina y que sea posible detectar que han sido generados o manipulados artificialmente; las soluciones técnicas empleadas han de ser eficaces, interoperables, sólidas y fiables en la medida en que sea técnicamente viable (art. 50.2). El Reglamento no impone una técnica de marcado concreta: el *watermarking* es una de las vías posibles junto a los metadatos de procedencia u otros mecanismos equivalentes, siempre que satisfagan el estándar de detección legible por máquina. El incumplimiento de estas obligaciones puede generar responsabilidad del proveedor por insuficiencia de las medidas técnicas de identificación y trazabilidad exigidas por el Reglamento. Paralelamente, los operadores de sistemas que generen imágenes, audio o vídeo sintéticos -en particular *deepfakes*- están obligados a revelar de forma clara y visible que el contenido ha sido generado o manipulado artificialmente (art. 50.4).

En el plano técnico, la iniciativa C2PA (*Coalition for Content Provenance and Authenticity*), promovida por organizaciones del sector tecnológico y mediático, ha desarrollado el estándar de metadatos de procedencia de contenidos más referenciado en la industria y en los documentos de orientación de la Comisión Europea. Su adopción por proveedores de IA generativa representa la concreción técnica más visible de las obligaciones de trazabilidad impuestas por el Reglamento de IA.

En materia probatoria, el marcado de agua es un instrumento fundamental para determinar si una pieza de evidencia digital presentada en un proceso es auténtica o ha sido generada o manipulada mediante IA. Su ausencia o adulteración pueden constituir un indicio de manipulación probatoria. *En materia de propiedad intelectual*, el marcado permite acreditar el origen de un contenido y, en consecuencia, delimitar los derechos sobre él; también puede emplearse para rastrear el uso no autorizado de obras protegidas en el entrenamiento de modelos. *En materia de información y libertad de expresión*, la identificación de contenidos generados por IA es una herramienta de lucha contra la desinformación y los *deepfakes* políticos o electorales, insertándose en el marco del Reglamento de Servicios Digitales (DSA, Reglamento UE 2022/2065).

Ninguna técnica de marcado de agua es actualmente infalible: las marcas pueden ser eliminadas o alteradas mediante técnicas de postprocesado, compresión o re-generación del contenido. La resiliencia de los sistemas de marcado es, en

consecuencia, un ámbito de activa investigación técnica y una condición de efectividad del marco regulatorio.

Tutela judicial algorítmica transfronteriza, *por P. Blanco-Morales Limones*

La tutela judicial algorítmica transfronteriza designa el conjunto de garantías que el artículo 47 de la CDFUE y el artículo 6 del CEDH proyectan sobre el proceso civil en el que intervienen sistemas de IA, ya sea como herramientas de asistencia al juez, como medio de prueba o como instrumento de acceso a la justicia, en litigios transfronterizos. Con esta expresión nos situamos ante la traslación al entorno algorítmico de un principio que el TJUE consagró como principio general del Derecho de la Unión en 1986, en la sentencia *Johnston*, donde declaró que la tutela judicial efectiva forma parte de los principios generales del Derecho europeo inspirados en las tradiciones constitucionales comunes de los Estados miembros y en los tratados internacionales sobre protección de los derechos humanos. La IA transforma las condiciones del ejercicio de esta garantía, los riesgos que la amenazan y los instrumentos jurídicos necesarios para preservarla.

En su dimensión de acceso a la justicia, la tutela judicial algorítmica transfronteriza plantea una paradoja contrafactual. Los sistemas de IA pueden facilitar ese acceso al reducir costes de traducción, facilitar el conocimiento del Derecho aplicable y automatizar trámites procesales repetitivos, pero pueden también restringirlo cuando la brecha digital entre las partes se convierte en una nueva forma de asimetría procesal. La parte que dispone de herramientas de análisis predictivo de jurisprudencia, de redacción asistida y de acceso algorítmico al Derecho extranjero se encuentra en una posición ventajosa respecto de la que carece de esas herramientas, con independencia de la fortaleza jurídica de sus pretensiones. Una asimetría que es la expresión más nueva y menos atendida de la desigualdad de armas en el proceso transfronterizo. Afrontarla exigiría una lectura del principio de igualdad procesal que incorpore la dimensión tecnológica y que considere si el régimen de asistencia jurídica gratuita en litigios transfronterizos, previsto en la Directiva 2003/8/CE, debe extenderse al acceso a herramientas de IA de asistencia jurídica para quienes carecen de recursos

En su dimensión de garantía del derecho a un proceso equitativo, la tutela judicial algorítmica transfronteriza exige que el uso de IA en el proceso no comprometa la imparcialidad del órgano jurisdiccional ni la posibilidad de control externo de sus resoluciones. La dependencia del juez respecto de un sistema cuyo funcionamiento interno no puede verificar introduce una forma nueva de heteronomía en la función jurisdiccional que es incompatible con las exigencias de independencia e imparcialidad del artículo 47 de la Carta cuando el sistema no es ni transparente ni auditable. El TEDH, en la sentencia *Gondert c. Alemania*, afirmó

que el deber de motivación es el corazón del control de las resoluciones judiciales y que sin la exposición de las razones adecuadas y verificables ningún mecanismo de supervisión puede operar como control. Trasladado al razonamiento judicial algorítmico, ese principio exige que la intervención de la IA en la formación de la convicción judicial sea declarada, trazable y susceptible de impugnación.

En su dimensión de reconocimiento transfronterizo, la tutela judicial algorítmica plantea un problema específico. Cuando una resolución adoptada con asistencia de IA en un Estado miembro ha de ser reconocida y ejecutada en otro, la opacidad de las herramientas utilizadas puede justificar la aplicación de la excepción de orden público que impida su reconocimiento. Los instrumentos europeos de reconocimiento que contemplan esa excepción comprenden, entre los más relevantes en materia civil, el Reglamento (UE) 1215/2012 (Bruselas I bis) en materia civil y mercantil general, el Reglamento (UE) 2019/1111 (Bruselas II ter) en materia matrimonial y de responsabilidad parental, el Reglamento (UE) n.º 650/2012 en materia de sucesiones, el Reglamento (UE) n.º 4/2009 en materia de obligaciones de alimentos y los Reglamentos (UE) 2016/1103 y 2016/1104 en materia de régimen económico matrimonial y efectos patrimoniales de las uniones registradas. Una resolución cuya cadena argumentativa es inescindible del uso de un sistema algorítmico no explicable puede ver comprometida su eficacia transfronteriza. En esta consecuencia extrema reside la razón más sólida para exigir que la tutela judicial efectiva en el proceso civil europeo sea también, necesariamente, tutela judicial algorítmicamente garantizada.

***Lex fori* digital**, *por P. Blanco-Morales Limones*

La expresión *lex fori* digital designa la función del Derecho procesal del foro en la regulación del proceso civil transfronterizo digitalizado. La digitalización de la cooperación judicial civil, impulsada por el principio *digital by default* del Reglamento (UE) 2023/2844 y articulada técnicamente mediante la infraestructura *e-CODEX*, no neutraliza las diferencias existentes entre las regulaciones de los Estados miembros ni las sustituye por un régimen uniforme. En un nuevo entorno tecnológico *lex fori* continúa siendo determinante para regular cómo se substancia el proceso, aunque el proceso se haga a distancia y en formato electrónico.

En su dimensión formal, la *lex fori* rige los requisitos de los actos procesales electrónicos. Entre ellos, la validez de la firma electrónica en documentos judiciales, la admisibilidad del documento digital como medio de prueba, la fe pública del Letrado de la Administración de Justicia en el entorno telemático, los plazos procesales en las comunicaciones vía *e-CODEX*. La forma de las notificaciones y traslados electrónicos transfronterizos son materias que armonizan solo parcialmente el Reglamento 2023/2844 y el Reglamento (UE) 2020/1784,

modificado por aquel para incorporar el artículo 19 bis sobre notificación electrónica a través del punto de acceso europeo. Los Reglamentos de Ejecución 2026/101 y 2026/102, que desarrollan las especificaciones técnicas del sistema *e-CODEX*, dejan un margen significativo al Derecho procesal nacional en lo que no está expresamente unificado. Determinar si la exigencia de firma electrónica cualificada para la presentación de escritos es una norma procesal sometida a la *lex fori* o una norma que puede quedar desplazada por la ley aplicable al fondo es una cuestión que los instrumentos de armonización vigentes no resuelven.

En la dimensión probatoria, la *lex fori* determina la admisibilidad y el valor probatorio de los *outputs* generados por sistemas de IA en el proceso civil transfronterizo, desde el dictamen pericial elaborado con asistencia algorítmica hasta los resultados de búsqueda semántica sobre jurisprudencia extranjera. El grado de apertura o resistencia de cada Ordenamiento procesal nacional a la prueba algorítmica es precisamente uno de los factores que puede incidir en la elección del foro por las partes litigantes (*forum shopping*).

Se pone de manifiesto así la antítesis entre la aspiración de la Unión Europea de construir un espacio judicial digital interoperable que propicie la integración de herramientas de IA y la resistencia de los Estados miembros a ceder a una estandarización tecnológica su soberanía procesal. Una competencia que el Tribunal de Justicia ha reconocido como expresión legítima de la identidad constitucional de los Estados miembros en los ámbitos no armonizados. La digitalización de la justicia en la UE no puede concebirse como un proceso de asimilación tecnológica unidireccional, sino como un delicado ejercicio de equilibrio. Mientras las instituciones comunitarias promueven un ecosistema interoperable en pos de la eficiencia y la cooperación transfronteriza, los Estados miembros alzan el principio de autonomía procesal como un verdadero bastión de su identidad nacional al amparo del artículo 4.2 del TUE. Esta soberanía procesal no es absoluta, sino que halla sus límites en los principios de equivalencia y efectividad delineados por el TJUE. De este modo, la resistencia a una estandarización tecnológica absoluta de los algoritmos e interfaces judiciales deviene en una reivindicación legítima; un recordatorio de que en el Espacio de Libertad Seguridad y Justicia (ELSJ) la vanguardia digital debe detenerse allí donde comience a desdibujarse la singularidad procedimental y el diseño constitucional que cada Estado ha erigido para salvaguardar sus propias garantías jurisdiccionales.

El juez nacional aplica su Derecho procesal con los instrumentos digitales que el Derecho europeo le proporciona, pero con los parámetros de admisibilidad, valoración y procedimientos que su propio ordenamiento le impone. El debate sobre si debiera hacerlo es uno de los dilemas más espinosos del futuro de la cooperación judicial civil digitalizada y servida por herramientas de IA.

Identidad digital transfronteriza, *por P. Blanco-Morales Limones*

La identidad digital transfronteriza designa el conjunto de mecanismos que permiten a personas físicas y jurídicas acreditar electrónicamente su identidad, capacidad jurídica y atributos verificables ante autoridades, órganos jurisdiccionales y operadores privados.

Su instrumento central es la Cartera Europea de Identidad Digital (*EU Digital Identity Wallet*, *EUDIW*), regulada por el Reglamento (UE) 2024/1183 (*eIDAS* 2.0), conforme al cual, a partir de finales de 2026, los Estados miembros deben ofrecer carteras de identidad digital a quienes las soliciten. No obstante, la disponibilidad universal estará condicionada a la adopción de los actos de ejecución que regulan sus requisitos técnicos de interoperabilidad, actualmente en fase de elaboración. La *EUDIW* permite almacenar y presentar credenciales verificables, entre las que se encuentran documentos de identidad, títulos académicos, certificados profesionales y poderes de representación, con reconocimiento en toda la UE.

El modelo de *Self-Sovereign Identity* (SSI) adoptado para la implementación de la *EUDIW* otorga al titular el control exclusivo sobre sus atributos digitales. La interacción entre el Reglamento *eIDAS 2.0* y el Reglamento General de Protección de Datos (RGPD) garantiza que los usuarios mantengan control efectivo sobre sus propios datos, mediante el principio de divulgación selectiva que permite presentar únicamente los atributos estrictamente necesarios para cada acto. La *EUDIW* y su configuración conforme al modelo *SSI* serán los mecanismos jurídicos clave para combatir la suplantación de identidad por IA en entornos transfronterizos (autenticación de origen).

Para las personas físicas, la *EUDIW* es el instrumento de identificación de uso voluntario ante las autoridades administrativas, los operadores privados y los sistemas judiciales, interconectados mediante el sistema *e-CODEX*, habilitando la acreditación de la legitimación procesal, la capacidad y la representación en formato electrónico

Para las personas jurídicas, el Reglamento 2024/1183 prevé la acreditación de atributos de identidad corporativa mediante declaraciones electrónicas con los mismos efectos jurídicos que las declaraciones en papel. La acreditación digital de la representación orgánica en procedimientos administrativos, judiciales y en el tráfico privado se regula en la Directiva (UE) 2025/25, de modificación de las Directivas 2009/102/CE y (UE) 2017/113, articulada a través de la interconexión de registros mercantiles nacionales mediante el sistema BRIS. El 19 de noviembre de 2025 la Comisión presentó la propuesta de Reglamento sobre las Carteras Europeas para Empresas (COM (2025) 838), parte del paquete *Digital Omnibus,* que de aprobarse extendería este sistema a las personas jurídicas con carácter

obligatorio para los organismos del sector público y con el objetivo de reducir las cargas administrativas que soportan especialmente las pequeñas y medianas empresas en el tráfico transfronterizo. La presentación automatizada de atributos verificables (títulos, poderes, etc.) mediante la cartera alimentará los sistemas de IA de las empresas de forma segura y respetuosa con el RGPD.

La identidad digital plantea cuestiones jurídicas inéditas. La primera concierne al Derecho aplicable a la validez y eficacia de los atributos digitales corporativos. Si la *EUDIW* de una sociedad constituida en los Países Bajos acredita poderes de representación conforme al Derecho neerlandés, qué ordenamiento rige la validez de esa acreditación ante un tribunal español es una pregunta que el Reglamento e *IDAS 2.0* ni ningún otro instrumento europeo resuelven expresamente. La segunda concierne a la equivalencia entre la identidad digital y los documentos en papel a efectos procesales, con las divergencias que subsisten entre los ordenamientos de los Estados miembros sobre admisibilidad y valor probatorio de las credenciales electrónicas, así como las exigencias de identificación plena que imponen algunos sistemas procesales nacionales.

Acceso algorítmico al Derecho extranjero, *por P. Blanco-Morales Limones*

El acceso algorítmico al Derecho extranjero designa el uso de sistemas de IA para determinar el contenido, la vigencia y la interpretación del ordenamiento jurídico de un Estado distinto al del tribunal que conoce del asunto. Históricamente, el acceso y la acreditación del Derecho foráneo ha sido una tarea onerosa, lenta y propensa al error, hasta el punto de que la imposibilidad práctica de su conocimiento ha servido con frecuencia de justificación para la aplicación indebida del Derecho sustantivo del foro.

Los modelos de lenguaje de gran escala (*Large Language Models*) y las bases de datos jurídicas dotadas de capacidades de búsqueda semántica permiten hoy plantear que ese obstáculo ha dejado de ser infranqueable. La IA supera la tradicional dificultad de acceso a las fuentes foráneas, una tarea hoy viable en una medida inusitada gracias a las tecnologías de la información y a la digitalización de los repertorios oficiales de legislación y jurisprudencia

La infraestructura europea que facilita este proceso se articula a través de dos identificadores estándar con distinto grado de implementación.

El Identificador de Legislación Europeo *(European Legislation Identifier, ELI*) ha sido formalmente adoptado por todos los Estados miembros desde 2024, aunque su cobertura real varía considerablemente. Mientras países como Francia e Italia lo han implementado de forma plena integrando los tres pilares del sistema, en otros la aplicación se limita al identificador básico sin extensión

a los metadatos semánticos, que son los que habilitan el acceso por máquina con garantías de precisión.

El Identificador Europeo de Jurisprudencia (*European Case Law Identifier*, *ECLI*) presenta una situación aún más limitada, ya que solo ha sido implementado en veintiuno de los Estados miembros, en la mayoría de los casos únicamente para un número reducido de órganos jurisdiccionales. De entre estos Estados, tan solo diecisiete están conectados al motor de búsqueda del Portal Europeo de e-Justicia. La agenda *DigitalJustice@2030* fija como primer objetivo la asignación de estos identificadores a la totalidad de las nuevas resoluciones judiciales en todos los órganos de los Estados miembros. La distancia entre esa ambición y el estado actual de implementación define el horizonte real en que operan hoy los sistemas de IA que acceden al Derecho de los Estados miembros.

Los riesgos específicos que plantea el acceso algorítmico al Derecho extranjero aplicable al caso son distintos de los que presenta la traducción automática de textos jurídicos, con la que con frecuencia se confunde. Acceder al Derecho extranjero no es traducirlo, sino comprenderlo en su contexto sistémico, reconocer la jerarquía de sus fuentes, identificar la norma vigente frente a la derogada y valorar la fuerza interpretativa de la jurisprudencia según los cánones del ordenamiento que la produce. Cuando esta tarea se delega en un modelo de IA, surgen riesgos específicos. El primero es la alucinación, esto es, la generación de preceptos, resoluciones o doctrinas inexistentes presentados con apariencia de verosimilitud. El segundo es la asimetría en la representación de los sistemas jurídicos en los conjuntos de entrenamiento de los modelos, que reproduce en el plano algorítmico la hegemonía de los ordenamientos mayoritarios en detrimento de los menos adaptados al formato digital. El tercero es la ausencia de trazabilidad de las fuentes, que impide al tribunal receptor verificar la exactitud y vigencia del Derecho alegado.

La respuesta jurídica a estos riesgos exige intervención en dos niveles. En el institucional, la plena implementación de los identificadores europeos en todos los Estados miembros es la condición de posibilidad de un acceso algorítmico fiable al Derecho y a la jurisprudencia de otros Estados miembros, pues sin datos estructurados, verificables y actualizados, ningún modelo puede ofrecer garantías suficientes. A nivel procesal, el uso de herramientas de IA para determinar el contenido del Derecho foráneo debe ir acompañado de las mismas exigencias de declaración, supervisión humana y verificación cruzada con fuentes oficiales que rigen para la prueba pericial algorítmica en general. Admitir el resultado del modelo como acreditación suficiente del Derecho extranjero supondría aceptar una opacidad que el proceso judicial democrático no tolera en el razonamiento del juez.

La utilización de herramientas de IA para determinar el contenido del Derecho foráneo no altera la naturaleza de la función que el tribunal y, en su caso, el Ministerio público desempeñan en el proceso transfronterizo. La identificación del Derecho extranjero aplicable al caso, su interpretación en el contexto del ordenamiento que lo produce y su valoración crítica son funciones indelegables, con responsabilidad plena e intransferible. La IA puede asistir en la fase de indagación y consulta de fuentes, pero el resultado del modelo no puede sustituir el juicio jurídico del órgano que debe aplicar ese Derecho. Aceptar el *output* del modelo como acreditación suficiente del Derecho extranjero sería incompatible con las garantías que el Estado de Derecho impone.

Razonamiento algorítmico transfronterizo, *por P. Blanco-Morales Limones*

El razonamiento algorítmico transfronterizo designa el proceso de formación de la convicción judicial en el que los *outputs* generados por sistemas de IA inciden, directa o mediatamente, sobre la construcción del silogismo jurídico, la valoración de la prueba o la motivación de la resolución, en nuestro caso, en procedimientos civiles con elemento extranjero. No se trata de un supuesto de decisión automatizada, que el artículo 22 del RGPD y el artículo 86 del Reglamento de IA regulan con perfiles propios, sino de algo más sutil y por ello más difícil de controlar. Se trata de aquellos casos en que la IA no decide, pero condiciona al que decide.

En el proceso civil transfronterizo, el razonamiento algorítmico plantea una magnitud diferenciada respecto de los litigios puramente internos, resultado de la asimetría normativa entre los Derechos procesales de los Estados miembros en cuanto a los límites del uso de IA por el juez y el control en vía de recurso de las resoluciones adoptadas con asistencia algorítmica. Al respecto, el TEDH en el asunto *Gondert c. Alemania*, en la línea de las precedentes *Ullens de Schooten y Rezabek c. Bélgica* y *Dhahbi c. Italia*, subraya que el deber de motivación es el corazón del control externo de las resoluciones judiciales y que sin razones adecuadas y verificables ningún mecanismo de supervisión puede considerarse como un genuino control. En el mismo sentido se pronuncia el TJUE, cuya jurisprudencia conforma el "escudo humano" frente a la automatización de la justicia, obligando a que cualquier rastro de IA en los tribunales sea auditable, explicable y, en última instancia, supervisado por una persona.

Trasladado al razonamiento algorítmico en un litigio civil transfronterizo, ese principio permite sostener que una resolución motivada con asistencia de IA cuyo *output* no sea explicable ni trazable vulnera el artículo 47 de la Carta de los Derechos Fundamentales de la Unión Europea y, además, puede comprometer su reconocimiento transfronterizo cuando la opacidad de su motivación alcance la

intensidad de una vulneración manifiesta del orden público procesal del Estado requerido en los términos del artículo 45 del Reglamento (UE) 1215/2012 (Bruselas I bis) y de los demás Reglamentos europeos que contemplan la denegación del reconocimiento por motivos de orden público, como el Reglamento (UE) 2019/1111 en materia de responsabilidad parental.

La distinción entre *fiabilidad predictiva* y *fiabilidad procedimental* es decisiva para comprender los riesgos que plantea esta figura. La primera pregunta si el sistema de IA predice bien el resultado. La segunda pregunta si el proceso por el que el juez llega a su decisión con asistencia de IA es transparente, motivado y susceptible de control. Esta segunda dimensión es la que exige un análisis más cuidadoso, porque determina si la resolución respeta el derecho a la tutela judicial efectiva con independencia de si el resultado fuera estadísticamente correcto. Un juez podría alcanzar una solución materialmente justa con un razonamiento opaco e incontrolable; ese razonamiento es igualmente ilegítimo. El riesgo de *automation bias*, esto es, la tendencia del decisor a sobrevalorar el *output* algorítmico en detrimento del análisis jurídico crítico y a asumir el resultado de la máquina sin verificación efectiva, convierte la fiabilidad procedimental en la garantía nuclear del Estado de Derecho, en particular en el proceso civil digitalizado. Las Directrices sobre el uso de IA generativa para los tribunales, aprobadas por la CEPEJ en diciembre de 2025, articulan esta misma exigencia al establecer que la intervención de la IA en el razonamiento judicial debe ser siempre declarada, supervisada y auditable.

Las exigencias de transparencia y explicabilidad de los artículos 13 y 14 del Reglamento (UE) 2024/1689 (RIA) son la garantía procesal mínima cuando el sistema de IA asiste al juez o al perito en la elaboración de la resolución o el dictamen. Su aplicabilidad *ratione temporis* en los litigios en curso, pese al diferimiento de plazos introducido por el acuerdo provisional del Ómnibus Digital sobre IA de mayo de 2026, es precisamente el objeto de las cuestiones novena y décima del asunto prejudicial C-245/25, pendiente de resolución por el Tribunal de Justicia.

Los textos españoles aprobados en 2026 obedecen plausiblemente a estos principios. El Proyecto de Ley Orgánica para el buen uso y la gobernanza de la inteligencia artificial, aprobado por el Consejo de Ministros el 26 de mayo de 2026 y remitido a las Cortes para su tramitación parlamentaria, designa al Consejo General del Poder Judicial como autoridad de supervisión del Reglamento de IA en el ámbito de la administración de justicia, junto a la Agencia de Supervisión de Inteligencia Artificial y la Agencia Española de Protección de Datos en sus respectivos ámbitos. Con anterioridad, la Instrucción 2/2026 del Consejo General del Poder Judicial, con efectos desde el 30 de enero de 2026, fue el primer instrumento normativo español, con independencia de su rango, que delimita el

marco de uso de la IA en la función jurisdiccional, estableciendo obligaciones de declaración, verificación y supervisión humana efectiva para los jueces y magistrados. Sin embargo, su ámbito subjetivo se circunscribe estrictamente a jueces y magistrados La Guía sobre uso de IA por el Ministerio Fiscal (v1.0, 18 regulatorio de mayo de 2026), elaborada por el Delegado de Protección de Datos del Ministerio Fiscal, aunque de naturaleza orientadora y preventiva cubre parcialmente ese vacío.

Pericial algorítmica transfronteriza, *por P. Blanco-Morales Limones*

La pericial algorítmica transfronteriza designa el dictamen elaborado con asistencia de sistemas de IA en procedimientos civiles que presentan un elemento extranjero. La cuestión adquiere particular relevancia cuando el perito que actúa en un Estado utiliza herramientas algorítmicas entrenadas en datos de otro ordenamiento, cuando las partes pertenecen a sistemas jurídicos con reglas de valoración probatoria divergentes o cuando la resolución adoptada sobre la base del dictamen ha de surtir efectos en un Estado distinto al del tribunal que la dictó. La divergencia entre ordenamientos en el tratamiento de la IA como herramienta de pericia puede influir en la elección del foro por las partes litigantes cuando la admisibilidad del dictamen algorítmico difiera según la Jurisdicción que conozca del litigio.

Las cuestiones más relevantes que plantea la pericial algorítmica en un proceso transfronterizo son tres.

La primera es la determinación del Derecho que rige la admisibilidad y el valor probatorio del dictamen cuando el proceso, por sus elementos objetivos o subjetivos, está vinculado a distintos Estados. El Derecho del tribunal que conoce del asunto está llamado a regular estas materias, pero las exigencias de transparencia y supervisión humana del Reglamento de IA, en la medida en que suponen un estándar europeo de naturaleza imperativa, actuará como límite a las soluciones nacionales permisivas con la opacidad algorítmica. Esta articulación entre el Derecho procesal del foro y el estándar imperativo europeo es una de las cuestiones más relevantes del proceso civil transfronterizo digitalizado.

La segunda cuestión concierne al control en vía de recurso de las resoluciones adoptadas sobre la base de dictámenes algorítmicos cuya motivación resulta inescindible del razonamiento de la máquina.

La tercera es la eficacia en otros Estados de esas resoluciones, pues cuando la opacidad de la motivación suponga una vulneración manifiesta del orden público procesal, el reconocimiento puede ser denegado en los términos del artículo 45 del Reglamento (UE) 1215/2012 (Bruselas I bis), del artículo 39 del Reglamento

(UE) 2019/1111 (Bruselas II ter) y de los demás instrumentos europeos de reconocimiento que contemplan la excepción de orden público.

El asunto prejudicial C-245/25 pendiente ante el TJUE sitúa esta figura en el centro del debate jurídico europeo. El tribunal remitente pregunta si un sistema de IA que asiste al perito en la elaboración del informe sobre cuya base el órgano judicial adoptará su resolución es un sistema de alto riesgo en el sentido del Reglamento de IA. La respuesta delimitará el estatuto jurídico de estos sistemas en sede probatoria y determinará qué exigencias de documentación, supervisión humana y explicabilidad deben cumplirse antes de que el dictamen algorítmico acceda al proceso. En el mismo orden de cuestiones, el asunto C-806/24 ha sido el primero en someter al Tribunal de Justicia la interpretación del artículo 86 del Reglamento de IA sobre el derecho a la explicación en las decisiones automatizadas con efectos jurídicos significativos. Al unísono, las Conclusiones del Consejo de la Unión Europea sobre el uso de la IA en el ámbito de la justicia (diciembre de 2024) ya reconocían la clasificación como de alto riesgo de los sistemas de IA para determinadas aplicaciones judiciales, subrayando la necesidad de procedimientos de evaluación de la conformidad que aseguren su fiabilidad. Asimismo, en el ámbito institucional, el propio TJUE adoptó en 2026 una *Carta ética en materia de inteligencia artificial* para sus propios procedimientos.

Las respuestas regulatorias e institucionales muestran dos aproximaciones metodológicas diferenciadas. En el espacio continental europeo, la respuesta está orientada a garantizar que los sistemas de IA utilizados por los peritos cumplan los estándares de transparencia, supervisión humana y explicabilidad exigidos por el RIA. En esta línea, el Proyecto de Ley Orgánica para el buen uso y la gobernanza de la inteligencia artificial, aprobado por el Consejo de Ministros español el 26 de mayo de 2026 y remitido a las Cortes para su tramitación parlamentaria, designa al Consejo General del Poder Judicial como autoridad de supervisión del Reglamento de IA en el ámbito de la administración de justicia, lo que incluirá la supervisión de los sistemas de IA utilizados en la elaboración de dictámenes periciales con efectos procesales.

En el Derecho anglosajón, el *Civil Justice Council* del Reino Unido publicó en enero de 2026 una Guía específica sobre el uso de la IA por los peritos en procesos civiles. La Guía establece que el perito debe declarar en su informe si ha utilizado herramientas de IA, identificar el sistema empleado, describir la función que ha desempeñado en la elaboración del dictamen y confirmar que ha verificado personalmente el contenido generado, asumiendo plena responsabilidad sobre su exactitud. Esta exigencia de declaración expresa persigue preservar la función del perito como auxiliar imparcial del tribunal y evitar que la opacidad del razonamiento algorítmico sustituya al juicio técnico personal que fundamenta la credibilidad del dictamen ante el juez.

Interpretación algorítmica transfronteriza, *por P. Blanco-Morales Limones*

La interpretación algorítmica transfronteriza designa el uso de sistemas de IA para la conversión lingüística en tiempo real de las declaraciones orales de testigos, partes y peritos en el marco de procedimientos civiles con elemento extranjero, así como el conjunto de cuestiones que suscitan su admisibilidad procesal y su incidencia sobre la valoración de la prueba. La figura presenta una problemática cualitativamente distinta a la de la traducción de documentos escritos. La interpretación oral opera en tiempo real, sin posibilidad de revisión inmediata del resultado, sobre una expresión lingüística que incluye, junto al contenido semántico de las declaraciones, también sus matices prosódicos, sus vacilaciones, sus silencios y su dimensión no verbal. Elementos todos ellos que el juez valora como parte integrante del testimonio y que los sistemas de reconocimiento automático de voz (*Automatic Speech Recognition*, ASR) no capturan ni transmiten con garantías suficientes.

El uso de sistemas de IA para la conversión lingüística y traducción/interpretación en tiempo real de declaraciones orales en procedimientos civiles transfronterizos es uno de los campos con mayor riesgo procesal. El principio de eficiencia y agilización procesal no puede situarse por encima del derecho a un proceso con todas las garantías, la inmediación y la fidelidad de la prueba. Un error de matiz o de traducción por parte de una IA en la declaración de un perito o un testigo puede determinar el sentido de un fallo judicial, viciando el proceso. A diferencia del ámbito penal (Directiva 2010/64/UE sobre el derecho a la interpretación), en el ámbito civil la regulación está más fragmentada.

El régimen jurídico de la intervención de intérpretes en el proceso civil español se establece en el artículo 143 de la Ley de Enjuiciamiento Civil. El precepto citado prevé la intervención de intérprete cuando alguna de las partes o los testigos y peritos que hayan de declarar no conozcan el castellano u otras lenguas oficiales, estableciendo que el intérprete será designado por el tribunal de entre los que se hallen incluidos en los listados elaborados por la Administración competente. No contempla el uso de sistemas de IA como instrumento de interpretación, lo que genera una laguna normativa de creciente relevancia práctica a medida que las audiencias por videoconferencia en litigios transfronterizos incorporan herramientas de interpretación automática. En el plano europeo, el Reglamento 2023/2844, que establece el marco de digitalización de la cooperación judicial civil y mercantil, no regula específicamente la interpretación algorítmica en audiencias remotas, remitiendo en lo no armonizado al Derecho procesal del Estado miembro que conoce del asunto.

Los estudios sobre el funcionamiento de los sistemas de reconocimiento automático de voz en contextos judiciales han identificado factores de error

específicos que se amplifican en el proceso civil transfronterizo. La presencia de hablantes no nativos, los acentos regionales o nacionales de las lenguas, las interferencias propias de las conexiones telemáticas y las particularidades del lenguaje jurídico especializado son variables que los modelos estándar no incorporan adecuadamente porque no suelen estar representadas en sus conjuntos de entrenamiento. En un proceso de familia transfronterizo con audiencia de menores, en un litigio de incumplimiento contractual entre partes de distintas nacionalidades o en un procedimiento de responsabilidad civil con testigos en múltiples jurisdicciones, un error de interpretación algorítmica puede comprometer la credibilidad del testigo ante el tribunal o alterar el sentido de una declaración determinante para el resultado del proceso. La versión algorítmica del habla de una parte, un testigo o cualquier interviniente en un proceso únicamente debe ser aceptable con garantías estrictas de supervisión humana.

La cuestión más compleja que se plantea en un procedimiento transfronterizo es la relativa al efecto de los errores de interpretación algorítmica sobre la valoración de la prueba y el control en vía de recurso. Cuando la convicción del tribunal sobre la credibilidad de un testigo se ha formado a través de una interpretación realizada por una herramienta de IA cuya fidelidad no puede verificarse porque el algoritmo no es accesible o auditable, o porque no existe registro del testimonio original, el derecho de la parte perjudicada a impugnar esa valoración carece de sustrato sobre el que operar. Esta consecuencia es especialmente grave cuando la resolución dictada en ese proceso ha de ser reconocida en otro Estado. En este caso, el tribunal del Estado requerido no puede controlar la adecuación del proceso celebrado en el Estado de origen si una parte esencial del material probatorio ha sido mediada por un sistema cuya fiabilidad no fue acreditada ni controlada. La exigencia de conservación del registro original en lengua fuente, la acreditación de la fiabilidad del sistema utilizado y la posibilidad de impugnación de la interpretación algorítmica son los tres instrumentos jurídicos que el proceso civil transfronterizo digitalizado necesita incorporar para preservar la efectividad del recurso en los términos del artículo 47 de la Carta y del artículo 45 del Reglamento Bruselas I bis.

Traducción algorítmica, *por P. Blanco-Morales Limones*

La traducción algorítmica designa el uso de sistemas de IA para la conversión lingüística de documentos escritos que han de aportarse como prueba en procedimientos civiles transfronterizos, así como el conjunto de cuestiones jurídicas que plantean su admisibilidad procesal, su valor probatorio y el régimen de responsabilidad por los errores que produzcan. La figura abarca tanto los documentos públicos extranjeros cuya traducción se exija para su reconocimiento como los documentos privados, correspondencia, contratos y comunicaciones

que las partes aportan como prueba documental en litigios de incumplimiento contractual, responsabilidad civil o Derecho de familia transfronterizos. A diferencia de la interpretación oral, que actúa en tiempo real sobre el habla, la traducción algorítmica tiene como objeto un texto fijo susceptible de verificación posterior; ello cambia el perfil de sus riesgos jurídicos sin eliminarlos. Aunque su utilización es ya una práctica extendida en el ámbito forense, su régimen jurídico carece aún de regulación específica en la mayoría de los Ordenamientos.

El Derecho de la UE, en el ámbito de la eficacia extraterritorial, ha adoptado una estrategia de formularios multilingües estandarizados como alternativa a la exigencia de traducción para determinadas categorías de documentos. El Reglamento (UE) 2016/1191 suprime el requisito de apostilla y simplifica los trámites en materia de copias certificadas e introduce formularios normalizados multilingües optativos en once ámbitos relativos a documentos de estado civil, con el efecto de que el formulario actúa como ayuda a la traducción adjunta al documento público. El Reglamento (UE) 2020/1784 sobre notificación y traslado de documentos judiciales y extrajudiciales establece doce formularios cuyos campos se cumplimentan en la lengua oficial del Estado miembro requerido. El Reglamento (UE) 2019/1111 (Bruselas II ter) contempla nueve tipos de formularios estandarizados para facilitar la circulación transfronteriza de resoluciones en materia matrimonial y de responsabilidad parental. El Reglamento (UE) 1215/2012 (Bruselas I bis) incorpora el certificado del artículo 53 como formulario multilingüe que acompaña a la resolución para su reconocimiento y ejecución transfronteriza. Esta estrategia reduce el impacto de la traducción algorítmica. Queda, no obstante, sin regulación el espacio más relevante desde el punto de vista probatorio, cual es el de los documentos privados y los documentos públicos que no se ajustan a ningún formulario estandarizado.

La traducción algorítmica de documentos judiciales en el marco del Reglamento (UE) 2020/1784 no goza de una presunción de conformidad absoluta. El estándar de inteligibilidad fijado por el TJUE (asunto *Henderson y Leonis)* faculta al justiciable a rechazar documentos cuyo procesamiento por IA altere o confunda los términos esenciales del litigio. Asimismo, la doctrina del TEDH (Caso *Avotiņš*) impone a los jueces de ejecución la obligación de fiscalizar la calidad de dicha traducción automática, impidiendo que la automatización de la cooperación judicial transfronteriza se realice a costa del sacrificio del derecho a la contradicción y del derecho de defensa

El tratamiento de los documentos en lengua extranjera en el proceso civil español se rige por el artículo 144 LEC, conforme al cual la traducción puede ser privada y tendrá plena eficacia probatoria salvo impugnación por la parte contraria, en cuyo caso el tribunal ordenará la traducción oficial a cargo de quien presentó el documento. Este régimen deja abierta la cuestión de si una traducción generada

por un sistema de IA satisface la exigencia del precepto. En particular, la ausencia de un traductor humano responsable no puede impedir activar el mecanismo de impugnación y la consiguiente exigencia de traducción oficial. La respuesta no es uniforme en los Ordenamientos de los Estados miembros, mientras algunos admiten libremente cualquier traducción con carácter indicativo, otros exigen traducción jurada para que el documento produzca plenos efectos probatorios. Un mismo contrato en inglés traducido algorítmicamente puede ser admitido como prueba plena ante un tribunal alemán y ser objeto de impugnación por falta de traducción oficial ante uno español, lo que incide sobre la elección del foro y sobre la previsibilidad del resultado procesal.

Los instrumentos normativos españoles de 2026 ofrecen indirectamente una orientación sobre el uso admisible de la IA en materia de traducción. La Instrucción 2/2026 del Consejo General del Poder Judicial habilita expresamente el uso de sistemas de IA por jueces y magistrados para la elaboración de traducciones como finalidad de preparación o estudio, con la limitación de que la información utilizada proceda únicamente de fuentes abiertas y la prohibición de incorporar datos judiciales a esos sistemas. Esta limitación tiene consecuencias directas en el proceso transfronterizo. Los documentos probatorios en lengua extranjera que integran el expediente judicial son precisamente datos judiciales, de modo que el juez no puede introducirlos en herramientas de IA no homologadas para obtener su traducción.

La cuestión de la responsabilidad por errores de traducción algorítmica que alteren el sentido jurídicamente relevante del documento carece de regulación expresa. Determinar si el error era evitable con supervisión humana razonable, si el sistema cumplía las exigencias de transparencia del Reglamento de IA y si la parte que aportó la traducción declaró ese uso al tribunal son tres extremos que ningún instrumento europeo vigente aborda, lo que convierte la traducción documental algorítmica en uno de los asuntos pendientes de la agenda regulatoria transfronteriza.

Confianza mutua algorítmica, *por P. Blanco-Morales Limones*

La confianza mutua algorítmica designa la condición de verificabilidad técnica y jurídica que los sistemas de IA deben satisfacer para que el principio de confianza mutua entre autoridades jurisdiccionales se ajuste al principio de reconocimiento mutuo de resoluciones en materia civil y mercantil, elemento esencial del Espacio de Libertad Seguridad y Justicia.

Esta noción es la proyección sobre el entorno algorítmico de una doctrina consolidada del Tribunal de Justicia de la Unión Europea según la cual la confianza mutua exige presumir que todos los Estados miembros respetan los

derechos fundamentales garantizados por el Derecho de la Unión, pero esa presunción admite excepciones cuando existen datos objetivos, fiables, precisos y debidamente actualizados que la desvirtúan. La IA añade a esa ecuación una nueva variable referida a los datos que fundan o destruyen la confianza. La trazabilidad, la auditabilidad y la reconstrucibilidad de un proceso decisorio algorítmicamente asistido se convierten en condiciones de posibilidad de la confianza mutua en el ELSJ digitalizado.

El principio de reconocimiento mutuo en materia civil descansa sobre la presunción de equivalencia entre los sistemas procesales de los Estados miembros. Esa presunción ha sido construida por el legislador europeo sobre la base de garantías normativas comunes, principalmente las derivadas del CEDH y de la CDFUE. Cuando los órganos jurisdiccionales de un Estado miembro incorporan sistemas de IA al proceso de formación de su convicción judicial, a la elaboración de dictámenes periciales o a la gestión de las comunicaciones procesales transfronterizas, esa presunción de equivalencia se ve expuesta a la opacidad algorítmica como obstáculo al control y a la revisibilidad de las resoluciones. Una resolución cuya cadena argumentativa es inescindible del razonamiento de un sistema no explicable compromete la tutela judicial efectiva del justiciable en el Estado de origen, al tiempo que desafía la confianza del tribunal del Estado requerido en la adecuación del proceso del que esa resolución procede. Cuando la decisión judicial incorpora de manera inescindible el *output* de un sistema algorítmico, la presunción de equivalencia no puede sustentarse únicamente en la confianza en las leyes procesales del Estado de origen; exige, además, la confianza en la fiabilidad del sistema técnico empleado. Si los tribunales de un Estado utilizan herramientas de IA que actúan como "cajas negras" opacas, la presunción de equivalencia se quiebra, y con ella la base del reconocimiento mutuo.

La Estrategia Europea relativa a la Justicia en Red 2024-2028 reconoce que la digitalización, el análisis de datos y la IA están cada vez más integrados en los servicios judiciales europeos y prevé una revisión en el segundo semestre de 2026. En el mismo sentido, la agenda *DigitalJustice@2030* fija como objetivo estratégico la construcción de un espacio judicial digital europeo plenamente interoperable, cuya viabilidad depende en última instancia de que los sistemas de IA integrados en sus infraestructuras sean auditables, trazables y verificables por los tribunales de todos los Estados miembros que participan en el ecosistema judicial compartido.

Esta vertiente de la confianza mutua es el fundamento del mandato de revisión acelerada que el artículo 112 del Reglamento de IA impone a la Comisión Europea. El legislador europeo obliga a evaluar la necesidad de introducir requisitos específicos de transparencia y gobernanza para los sistemas de IA utilizados en la administración de justicia, previendo la adopción de medidas legislativas complementarias. Esta previsión de revisión acelerada revela que

el legislador es consciente de que los marcos normativos vigentes no fueron diseñados para regular arquitecturas algorítmicas avanzadas que presentan cuatro propiedades críticas:

a) Capacidad de planificación autónoma: Aptitud del sistema para trazar estrategias de resolución de problemas complejos sin intervención humana directa.

b) Memoria persistente: Capacidad de retener y acumular experiencias de datos a lo largo del tiempo, modificando sus propios parámetros de ejecución internos.

c) Delegación a sistemas subordinados: Facultad del algoritmo principal para subcontratar tareas y procesamientos a otras subredes o modelos periféricos.

d) Adaptación conductual basada en inferencias: Habilidad para alterar sus respuestas y criterios interpretativos a partir de las deducciones lógicas de los datos de entrada.

En estas arquitecturas la localización de los tratamientos, la determinación de responsabilidades y la atribución de efectos jurídicos se fragmentan de tal forma que los mecanismos tradicionales de atribución de competencia judicial y de reconocimiento de resoluciones no pueden gestionarlas sin una profunda adaptación.

La gobernanza dinámica de los sistemas de IA sería un mecanismo de restauración de esa confianza. Frente a los modelos clásicos de control *ex ante* centrados en evaluaciones de conformidad, certificaciones estáticas y auditorías previas a la puesta en servicio, la gobernanza dinámica desplaza el centro de gravedad del control hacia la monitorización continua del comportamiento del sistema durante su operación efectiva. La trazabilidad de los flujos de inferencia, la conservación estructurada de registros decisionales y la posibilidad de reconstruir retrospectivamente la secuencia de datos, herramientas e instrucciones empleados en la elaboración de una resolución son las condiciones que los sistemas de IA judiciales deben cumplir para que el ELSJ continúe funcionando con la coherencia y la legitimidad que sus fundamentos exigen.

La IA no puede considerarse moralmente neutra, la confianza entre instituciones y personas en la era algorítmica exige que el criterio rector no sea el beneficio económico sino la dignidad de cada ser humano (encíclica *Magnifica Humanitas* del Papa León XIV, 15 de mayo de 2026). Postulado que resuena en el fundamento último del principio de confianza mutua en el Derecho europeo.

Asimetría cognitiva algorítmica, *por P. Blanco-Morales Limones*

La asimetría cognitiva algorítmica designa el desequilibrio entre sujetos derivado de las diferencias en el acceso, comprensión y capacidad de explotación de herramientas de IA para el análisis, la argumentación y la gestión de situaciones jurídicas. La jurisprudencia del TEDH ha establecido que la igualdad de armas procesales exige que todas las partes puedan disponer de las mismas posibilidades para presentar y defender su causa. La IA introduce en ese principio una dimensión nueva que los instrumentos jurídicos vigentes no fueron diseñados para abordar.

Quien dispone de sistemas de IA de alto rendimiento cuenta con más medios y con una capacidad cualitativamente distinta de procesar información jurídica, identificar precedentes, predecir resultados y construir argumentaciones que, en sus aplicaciones más avanzadas, pueden ser muy superiores a lo que el trabajo humano no asistido por IA pudiera producir en el mismo tiempo y con los mismos recursos. La asimetría cognitiva algorítmica transfronteriza coloca al ciudadano en una posición de desventaja proscrita por la doctrina de la igualdad de armas del TEDH (Caso *Regner*) y del TJUE (Asunto *Steffensen*). Al ser la infraestructura técnica una 'caja negra' en posesión exclusiva de la contraparte, los tribunales no pueden amparar una denegación de acceso basada ciegamente en secretos comerciales (Asunto *Varec*), sino que vienen obligados a desplegar mecanismos de contradicción pericial reforzada y salvaguardas compensatorias (Caso *Mantovanelli*) que garanticen que la complejidad del código no anule el derecho fundamental a un proceso equitativo

En el proceso civil transfronterizo, la asimetría cognitiva algorítmica se manifiesta en tres dimensiones especialmente relevantes. La primera es la asimetría en el análisis predictivo, ya que la parte que accede a sistemas entrenados con jurisprudencia comparada de los tribunales del Estado cuyo Derecho rige el fondo del litigio puede anticipar cuáles son los argumentos más eficaces y cómo ha resuelto ese tribunal situaciones análogas, con una precisión inasequible para su contraparte. La segunda es la asimetría en el acceso al Derecho extranjero, que implica que mientras una parte puede resolverlo algorítmicamente en minutos, la otra parte soportará costes muy superiores. La tercera es la asimetría en la gestión documental y probatoria, pues la capacidad de procesar grandes volúmenes de documentación procesal con asistencia algorítmica convierte esa ventaja en una diferencia que se magnifica en los litigios complejos con múltiples partes y múltiples ordenamientos implicados. La STEDH *Latorre Atance c. España* ilustra cómo la asimetría de recursos entre partes puede comprometer el carácter equitativo del proceso incluso cuando las reglas formales son aparentemente neutrales.

En relación con el principio de igualdad, la asimetría cognitiva algorítmica plantea cuestiones que trascienden el proceso para alcanzar la esencia misma del

Estado de Derecho. El acceso a la justicia no es solo acceso a los tribunales, sino capacidad real de hacer valer los derechos ante ellos con posibilidades efectivas de éxito. Cuando esa capacidad depende de la disponibilidad de herramientas de IA que solo determinadas partes pueden costear o a las que solo ciertos sistemas jurídicos tienen acceso preferente, la igualdad ante la ley se convierte en una garantía de contenido menguante. La diferencia entre el litigante que puede permitirse un abogado experto y el que no puede se acrecienta a favor del litigante cuyo abogado trabaja con IA de alto rendimiento y aquel cuyo abogado no dispone de ella. Esa diferencia, proyectada sobre los cada día más frecuentes litigios transfronterizos, podría llevar a favorecer sistemáticamente los resultados procesales en pro de las partes tecnológicamente mejor dotadas

La asimetría cognitiva algorítmica en los procesos transfronterizos genera una quiebra del principio de igualdad de armas, ya que la opacidad técnica y el secreto comercial de la "caja negra" impiden al ciudadano afectado comprender o probar la lógica de una decisión automatizada dictada en otro Estado miembro. Frente a este escenario, la Directiva 2003/8/CE sobre asistencia jurídica gratuita queda obsoleta si se limita a sufragar costes tradicionales. Se requiere una reinterpretación sistemática coherente con el Reglamento de IA y el Reglamento (UE) 2023/2844 de digitalización judicial. Ello exige expandir su cobertura económica para financiar peritajes informáticos avanzados, contraperitajes algorítmicos y auditorías de datos transfronterizas. Solo mediante esta actualización patrimonial, combinada con la flexibilización o inversión de la carga de la prueba en entornos judiciales controlados, será posible neutralizar el desequilibrio técnico y garantizar una verdadera tutela judicial efectiva frente a los sesgos ocultos del algoritmo. El Reglamento de IA no aborda la asimetría cognitiva como problema autónomo. La agenda *DigitalJustice@2030* contempla el acceso equitativo a herramientas digitales como objetivo genérico sin instrumentos vinculantes específicos. La elaboración de un régimen jurídico que corrija la asimetría cognitiva algorítmica implica uno de los problemas candentes del proceso civil europeo contemporáneo. Su solución determinará si la IA se convierte en un instrumento de democratización del acceso a la justicia o en un nuevo mecanismo de exclusión.

La encíclica *Magnifica Humanitas* del Papa León XIV (15 de mayo de 2026) subraya que en la era digital la justicia social exige garantizar a todos un acceso equitativo a las oportunidades tecnológicas y proteger a los más vulnerables de la concentración de poder que los sistemas de IA pueden generar, argumento que resuena directamente con el problema que esta voz identifica.

Contrato de IA transfronterizo, *por P. Blanco-Morales Limones*

Los términos de uso (TOU) de los grandes proveedores de sistemas de IA son contratos de adhesión celebrados a escala global con millones de usuarios. En la generalidad de los casos contienen cláusulas de elección de ley aplicable, de elección de foro o de sumisión a arbitraje cuya validez, eficacia y límites se rigen, cuando el usuario reside en un Estado de la Unión Europea, por el Reglamento (UE) n.º 1215/2012 (Bruselas I bis) y por el Reglamento (CE) n.º 593/2008 (Roma I), dada su condición de contratos celebrados con consumidores. El Reglamento (CE) n.º 864/2007 (Roma II), que regula la ley aplicable a las obligaciones extracontractuales, no resulta aplicable a la relación contractual entre el usuario y el proveedor, aunque podría serlo si de ese contrato derivaran pretensiones de responsabilidad no contractual frente a terceros. El Reglamento Bruselas I bis garantiza al consumidor europeo la posibilidad de demandar ante los tribunales de su domicilio y prohíbe las cláusulas de sumisión que priven de ese derecho con carácter previo al litigio, con independencia del contenido del contrato de adhesión. El Reglamento Roma I limita la eficacia de la elección de ley extranjera cuando las normas del Estado de residencia del consumidor otorguen mayor protección que la ley elegida.

Las soluciones adoptadas por los principales proveedores revelan tres modelos distintos, reflejo de estrategias deliberadas ante el mercado europeo.

El primer modelo, de adaptación plena al Derecho europeo, es el adoptado por *OpenAI, Google, Meta y Anthropic*. Los cuatro proveedores han constituido filiales en Irlanda que actúan como entidades contratantes para los usuarios del Espacio Económico Europeo. Se trata, respectivamente, de las sociedades OpenAI *Ireland Ltd., Google Ireland Limited, Meta Platforms Ireland Limited y Anthropic Ireland, Limited*, esta última constituida en marzo de 2024 y designada como sede EMEA del grupo. A través de sus filiales irlandesas, estos proveedores someten el contrato a la ley del Estado de residencia del usuario y les garantizan el acceso a sus tribunales. No aplican las cláusulas de sumisión al arbitraje, ni las cláusulas de sumisión a los tribunales de California o de otra jurisdicción extraeuropea previstas en los términos generales aplicables al resto del mundo. Este modelo es, en lo esencial, compatible con los artículos 17 a 19 del Reglamento Bruselas I bis. Su punto débil estriba, sin embargo, en que la constitución de una filial no implica necesariamente que las cláusulas contractuales europeas recojan todas las exigencias del Reglamento, ni que el tratamiento de datos se realice en servidores situados en la Unión Europea, cuestión relevante para los sectores regulados y tratada en la voz específica sobre privacidad algorítmica transfronteriza.

El segundo modelo, de incompatibilidad con el Derecho europeo, es el de *DeepSeek,* cuya empresa titular, *Hangzhou DeepSeek Artificial Intelligence Co., Ltd.,*

somete todos sus términos al Derecho de la República Popular China y atribuye jurisdicción a los tribunales de Hangzhou. La ausencia de establecimiento en la Unión y la falta de un representante autorizado en el sentido del artículo 22 del Reglamento de IA convierten sus cláusulas en una renuncia anticipada de derechos y garantías procesales que los artículos 19 y 25 del Reglamento Bruselas I bis no permiten, pues este último precepto solo reconoce eficacia a los acuerdos atributivos de competencia celebrados con consumidores cuando son posteriores al nacimiento del litigio o amplían las opciones del consumidor para litigar.

El tercer modelo, de sumisión a arbitraje privado sin adaptación europea, es el de *Perplexity AI*, Inc. Sus términos de uso, actualizados en enero de 2026, contienen en la sección 9 una cláusula de sumisión a arbitraje y una renuncia expresa a acciones colectivas. Con esta cláusula el usuario acepta resolver todos los litigios relativos al servicio mediante arbitraje, renunciando al derecho a acudir a un órgano jurisdiccional estatal. Los términos de la API someten el contrato a la ley del Estado de California y atribuyen jurisdicción exclusiva a los tribunales del Condado de San Francisco, sin ninguna adaptación para los usuarios del Espacio Económico Europeo. *Perplexity* carece de filial operativa en la Unión Europea y no ha constituido representante autorizado en el sentido del artículo 22 del Reglamento de IA. Mientras *DeepSeek* impone la sumisión a tribunales extranjeros, *Perplexity* suprime directamente el acceso a la justicia estatal mediante arbitraje privado en territorio estadounidense, lo que choca frontalmente con los artículos 19 y 25 del Reglamento Bruselas I bis, que prohíben las cláusulas que priven al consumidor de su derecho a litigar ante sus propios tribunales con carácter previo al nacimiento de la controversia.

La dimensión supervisora sobre estos aspectos se ha intensificado en 2025 y 2026. Con aproximadamente 120 millones de usuarios activos mensuales en la Unión Europea en el semestre cerrado en septiembre de 2025, *ChatGPT* está siendo evaluado por la Comisión Europea para su posible designación como motor de búsqueda en línea de muy gran tamaño (*Very Large Online Search Engine*, VLOSE) en virtud del Reglamento de Servicios Digitales, lo que impondría obligaciones adicionales de transparencia y supervisión que trascienden el plano contractual y sitúan la relación usuario-proveedor bajo la ordenación tutelar de la regulación pública.

La coexistencia del Reglamento de IA, el Reglamento de Servicios Digitales y el Reglamento Bruselas I bis como marcos normativos aplicables a esa relación configura uno de los problemas de coordinación normativa más complejos del Derecho privado europeo contemporáneo, al que el Reglamento Roma I añade la dimensión del conflicto de leyes en la determinación del Derecho sustantivo aplicable al contrato.

Privacidad algorítmica transfronteriza, *por P. Blanco-Morales Limones*

Las políticas de privacidad de los grandes proveedores de sistemas de IA son instrumentos contractuales que regulan el tratamiento de los datos personales de los usuarios. En cuanto afectan a personas residentes en la Unión Europea, quedan sometidas al Reglamento General de Protección de Datos (RGPD). Según su artículo 3.2 rige todo tratamiento de datos de personas físicas que se encuentren en la Unión cuando el tratamiento esté relacionado con la oferta de bienes o servicios a esos interesados, con independencia de dónde esté establecido el responsable. Esta cláusula de alcance extraterritorial, cuya importancia práctica fue subestimada hasta que la proliferación de los grandes modelos de lenguaje la convirtió en el eje del debate regulatorio, supone que el RGPD se aplica a cualquier relación entre cualquier usuario europeo y cualquier proveedor de IA que dirija sus servicios al mercado europeo, con independencia de su lugar de establecimiento.

El análisis comparado de las políticas de los principales proveedores revela, al igual que en materia contractual, tres modelos diferenciados.

El modelo de adaptación al marco europeo es el adoptado por *OpenAI, Google, Meta y Anthropic*, que designan sus filiales irlandesas como responsables del tratamiento para los usuarios del Espacio Económico Europeo y activan el mecanismo de ventanilla única con la Comisión de Protección de Datos irlandesa como autoridad de control principal. Las transferencias de datos a sus servidores en los Estados Unidos se amparan en el Marco de Privacidad UE–EEUU, optado mediante Decisión de adecuación de la Comisión en julio de 2023, con las Cláusulas Contractuales Estándar como salvaguarda subsidiaria. Este modelo presenta, con todo, una vulnerabilidad de considerable interés. Un tribunal federal de Nueva York (SDNY) ordenó a *OpenAI* suspender temporalmente el borrado de los historiales de chat de los usuarios de *ChatGPT*. Aunque *OpenAI* alegó que la medida chocaba directamente con las normativas de privacidad y con las solicitudes de supresión de datos bajo el artículo 17 del RGPD europeo, la jueza rechazó inicialmente este argumento, obligando a la tecnológica a congelar y anonimizar dichos registros. Esta orden fue finalmente levantada en septiembre de 2025, permitiendo a *OpenAI* reactivar sus políticas de borrado habituales tras acordar la entrega a los demandantes de una muestra estadística de 20 millones de chats estrictamente anonimizados y bajo secreto de sumario. No obstante, la medida dejó en un limbo legal aquellos datos que los usuarios intentaron borrar entre abril y septiembre de 2025, que permanecen bajo custodia judicial como prueba del caso. La orden judicial estadounidense y el RGPD apuntan, en este punto, en direcciones opuestas, dado que el artículo 48 RGPD limita el reconocimiento de órdenes de tribunales de terceros países que impliquen transferencia de datos sin el amparo de un instrumento válido conforme a su Capítulo V. El Marco de

Privacidad UE–EEUU adolece de fragilidad ante una eventual nueva impugnación ante el TJUE, como sucedió en los asuntos *Schrems I* y *Schrems II*. A ello se añade la incertidumbre en cuanto a si una nueva invalidación por el Tribunal de Justicia afectaría retroactivamente a los tratamientos ya realizados, cuestión que es objeto del asunto C-703/25 P, actualmente pendiente.

El modelo de cumplimiento declarado con garantías insuficientes es el de *Perplexity AI, Inc.*, cuya política de privacidad alega conformidad con el RGPD mediante la certificación del Marco de Privacidad UE–EEUU y Cláusulas Contractuales Estándar. Sin embargo, su infraestructura principal opera en servidores de *Amazon Web Services* (AWS) con base en EEUU, ofreciendo una residencia de datos en la Unión Europea de disponibilidad limitada, restringida principalmente a sus soluciones de suscripción. Los usuarios de las versiones gratuita y Pro deben, en la configuración de su cuenta, desactivar manualmente el uso de sus datos para el entrenamiento de modelos, lo que no es la opción predeterminada, contradiciendo el principio de protección de datos por defecto del artículo 25 del RGPD. Expertos europeos han señalado además que *Perplexity* utiliza tecnologías de seguimiento, como *Google Analytics*, sin consentimiento explícito previo, apoyándose en el interés legítimo como base jurídica, un enfoque cuestionable bajo el RGPD. Aunque la compañía ha intentado mitigar su falta de presencia física designando canales de representación bajo el artículo 27 del Reglamento, su modelo genera fricciones de supervisión. A abril de 2026, si bien no se ha ejecutado una sanción económica formal, la tecnológica ya afronta expedientes de investigación por parte de varias autoridades de control europeas

El modelo de incompatibilidad con el Derecho europeo es el de *DeepSeek*. Desde su lanzamiento en enero de 2025, la respuesta regulatoria europea fue de una intensidad sin precedentes en el sector de la IA. El *Garante* italiano prohibió de forma urgente el tratamiento de datos de sus usuarios el 30 de enero de 2025, abriendo una línea a la que se sumaron investigaciones de la DPC irlandesa, la CNIL francesa, la AEPD española y las autoridades de Bélgica, Luxemburgo y Croacia. Especial relevancia tuvo la decisión de la *Berliner Beauftragte für Datenschutz und Informationsfreiheit*, que en junio de 2025 requirió a Apple y Google para la retirada de la aplicación de sus mercados, considerándola ilegal según la Ley de Servicios Digitales (DSA). Aunque *DeepSeek* modificó su política de privacidad para la *Región Europea,* introduciendo referencias al artículo 6 del RGPD y designando un representante en la Unión, a abril de 2026 las autoridades supervisoras siguen considerando estas medidas como un mero formalismo cosmético. Las medidas son claramente insuficientes por la ausencia de garantías adecuadas para las transferencias a China y por las obligaciones de cooperación con los servicios de inteligencia que la legislación china impone a las empresas establecidas en ese territorio, lo que invalida cualquier garantía del Capítulo V

del RGPD y consolida el principal paradigma de vulneración de la soberanía de datos europea.

La respuesta fragmentada frente a *DeepSeek* revela una laguna histórica en el sistema europeo de protección de datos, debida a la ausencia de un mecanismo centralizado en el RGPD para proveedores sin establecimiento en la Unión. Al carecer de filial en la UE, el mecanismo de ventanilla única no funciona, fragmentando la competencia y dificultando la ejecución coercitiva de sanciones contra entidades sin activos en territorio europeo. Si bien el caso *DeepSeek* expuso esta vulnerabilidad con una nitidez sin precedentes, la gobernanza europea ha comenzado a reaccionar para mitigarla. El despliegue de la Oficina de IA de la UE, prevista en el RIA, que centraliza la supervisión de modelos de propósito general de terceros países, y las recientes reformas de armonización procesal del RGPD pretenden colmar este vacío. No obstante, la articulación práctica entre estas nuevas herramientas de gobernanza continúa siendo uno de los problemas más complejos y urgentes del marco digital europeo.

B) Legislación española

Decisión administrativa automatizada, *por D. A. Cuesta Bárcena*

En Derecho Administrativo, el término «decisión» se reserva para el acto administrativo finalizador de un procedimiento. El art. 41.1 de la Ley 40/2015, de 1 de octubre, de Régimen Jurídico del Sector Público (LRJSP) define la «actuación administrativa automatizada» como «cualquier acto o actuación realizada íntegramente a través de medios electrónicos por una Administración Pública en el marco de un procedimiento administrativo y en la que no haya intervenido de forma directa un empleado público». Así, la «decisión automatizada» es el acto administrativo finalizador de un procedimiento dictado íntegramente a través de medios electrónicos y sin intervención directa de un empleado público.

El art. 13.1 del Reglamento de actuación y funcionamiento del sector público por medios electrónicos, aprobado por el Real Decreto 203/2021, de 30 de marzo (RAFME), habilita a todas las Administraciones públicas a automatizar sus procedimientos. En el ámbito estatal, ello debe autorizarse por resolución del titular del órgano administrativo competente por razón de la materia o del órgano ejecutivo competente del organismo o entidad de Derecho público y publicarse en la sede electrónica o sede electrónica asociada (art. 13.2 RAFME). Esta resolución tiene que indicar el órgano que debe ser considerado responsable a efectos de la impugnación de la actuación (art. 41.2 LRJSP), los recursos que procedan contra la misma, el órgano administrativo o judicial ante el que hubieran de presentarse y

el plazo para interponerlos y establecer medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos de las personas interesadas (art. 13.2 RAFME). A diferencia de la legislación administrativa de alguna comunidad autónoma, la estatal no limita los actos a que puede extenderse la automatización (v. gr. la Ley 26/2010, de 3 de agosto, de régimen jurídico y de procedimiento de las administraciones públicas de Cataluña, prevé que solo son susceptibles de actuación administrativa automatizada los actos que puedan adoptarse con una programación basada en criterios y parámetros objetivos [art. 44.2]). En todo caso, las sedes electrónicas correspondientes tienen que disponer, a disposición de los interesados, de la relación actualizada de las actuaciones administrativas automatizadas vinculadas a los servicios, procedimientos y trámites disponibles, acompañándose cada una de la descripción de su diseño y funcionamiento, los mecanismos de rendición de cuentas y transparencia y los datos utilizados en su configuración y aprendizaje (art. 11.1 i RAFME).

En la medida en que la automatización de actuaciones administrativas requiera el tratamiento de datos personales, estará sometida a la normativa reguladora de esta materia. En particular, el art. 22 del Reglamento (UE) 2016/679 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos, establece que todo interesado tiene derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado de datos personales, incluida la elaboración de perfiles, que produzca efectos jurídicos en él o le afecte significativamente de modo similar. Esta previsión está limitada por una serie de excepciones, entre las que destaca la no prohibición de decisiones automatizadas si el Derecho de la Unión Europea o de los Estados miembros las autoriza y establece medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado.

Asimismo, la automatización de actuaciones administrativas basadas en IA está sometida a la regulación de esta tecnología, encabezada por 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 IA. Esta norma no prohíbe el uso de IA por la Administración pública (art. 5) ni considera, con carácter general, de «alto riesgo» a los sistemas que la misma pueda emplear en la tramitación de sus procedimientos. Asimismo, el art. 23 de la Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación, establece que la Administración debe favorecer la puesta en marcha de mecanismos para que los algoritmos involucrados en la toma de decisiones tengan en cuenta criterios de minimización de sesgos, transparencia y rendición de cuentas; priorizar la transparencia en el diseño y la implementación de estos algoritmos y la capacidad de interpretación de las decisiones adoptadas por los mismos; y promover el uso de una IA ética, confiable y respetuosa con los derechos fundamentales.

Reserva de humanidad, *por J. Ponce Solé*

La reserva de humanidad nace en 2019 como un concepto doctrinal técnico-jurídico, hoy en día ya incorporado al ordenamiento jurídico. Postula la exclusión del uso de la IA en ciertos ámbitos de la actuación pública, donde ha sido especialmente desarrollado (si bien puede ser extendido al sector privado con las debidas adaptaciones por aplicación analógica).

Significa que la decisión final a adoptar, en ciertos casos, debe recaer necesaria y exclusivamente en un ser humano y que la automatización no puede ser total, si bien puede servir de asistencia al mismo, quien debe evitar, en todo caso, el sesgo cognitivo de automatización.

La reserva de humanidad puede ser comparable, salvando las distancias, a la reserva de funciones para funcionarios de carrera. Así como el legislador en España (art. 9 EBEP) reserva ciertas potestades a empleados públicos funcionarios, cuya relación estatutaria garantiza objetividad e imparcialidad frente a presiones externas, la reserva de humanidad garantiza que decisiones que afecten derechos de las personas sean tomadas por humanos dotados de sabiduría y ética humana.

1. Fundamentos técnicos, emocionales y cognitivos

La necesidad de una reserva de humanidad se apoya en tres limitaciones intrínsecas de los actuales sistemas de IA frente a los humanos.

En primer lugar, la ausencia de empatía. Las máquinas carecen de la capacidad de comprender el sufrimiento, la vulnerabilidad o los problemas de los seres humanos, un componente cognitivo y emocional esencial para adaptar el comportamiento ante situaciones humanas complejas, que enlaza con el principio de equidad recogido en el Código Civil español. Se ha sostenido que la IA es, por definición, "psicópata" en términos funcionales, ya que puede imitar emociones para manipular, pero no puede "sentir" ni identificarse con los infortunios ajenos. Ello, de acuerdo con la necesidad de combinar emociones, correctamente gobernadas, con la "fría racionalidad" para adoptar buenas decisiones (como señalan el neurocientífico Antonio Damasio o la filósofa Victoria Camps), implica que allí donde haya margen de apreciación existe un alto riesgo de mala decisión de las máquinas.

En segundo lugar, la IA es incapaz de realizar inferencias abductivas. Mientras que opera mediante la deducción (reglas simbólicas) o la inducción (estadística y patrones), carece de la capacidad de abducción, teorizada por el filósofo estadounidense Charles S. Peirce ya en el siglo XIX. La abducción permite a los humanos formular conjeturas creativas e integrar el sentido común

y la experiencia global del mundo ante situaciones inéditas, algo que la IA no puede realizar.

En tercer lugar, inexistencia de compromiso moral y de responsabilidad en el marco de un proceso debido de decisión motivada. La toma de decisiones por humanos conlleva un compromiso moral que la IA puede desplazar, por ejemplo, en el sector público. La reserva evita que los servidores públicos se "escuden" tras decisiones maquinales para eludir su responsabilidad personal y profesional. Además, el compromiso y la responsabilidad en la toma de la decisión debe ser fruto de un proceso debido de toma de decisiones, que en el sector público debe integrar todos los componentes del derecho y el principio de buena administración, incluyendo una fundamentación y una motivación adecuada.

2. Límites a la toma de decisiones discrecionales por máquinas

El núcleo del concepto en el ámbito público radica en la prohibición de la automatización total de las potestades que implican un margen de apreciación volitivo (conceptos jurídicos indeterminados valorativos y discrecionalidad *strictu sensu*). Y ello porque la discrecionalidad no es una elección libre entre opciones indiferentes, sino la obligación de ponderar con debida diligencia y equidad todos los elementos relevantes de un caso concreto para servir al interés general de acuerdo con las exigencias del principio y derecho a una buena administración.

El uso de IA simbólica o basada en reglas en estos ámbitos provocaría una *rigidificación* ilegal de la discrecionalidad (*fettering*, en la tradición jurídica del *Common Law*, que la prohíbe), al predeterminar soluciones sin atender a las circunstancias singulares del caso. Por su parte, en el caso de la IA conexionista, con uso de estadística (aprendizaje automático, generativa etc.), esta presenta el problema de las "cajas negras", que impide motivar de forma inteligible y humana el razonamiento que conduce a una decisión, vulnerando el principio de interdicción de la arbitrariedad (art. 9.3 Constitución española).

3. Manifestaciones normativas y principio de precaución

Aunque el concepto, como dijimos, parte de una construcción doctrinal en España extendida luego a otros países, como Italia, encuentra ya anclajes en el Derecho positivo vigente, explicitando o no la expresión reserva de humanidad.

Así, el RGPD (Art. 22) configura una reserva de humanidad general al otorgar, en principio y salvo las excepciones previstas, el derecho a no ser objeto de decisiones basadas únicamente en el tratamiento automatizado con efectos jurídicos o significativos. Asimismo, el Reglamento de IA de la UE prohíbe en su anexo III que la IA pueda substituir a las decisiones judiciales, de las que solo puede ser un apoyo, por ser estas intrínsecamente humanas, según se señala en

los Considerandos. Por su parte, el *Ombudsman* europeo en una actuación de oficio de finales de 2024 le ha indicado a la Comisión europea la imposibilidad de substituir el juicio humano discrecional por decisiones automatizadas, citando jurisprudencia del TJUE. Asimismo, la Carta de Derechos Digitales de España establece que la adopción de decisiones discrecionales debe quedar reservada a personas, salvo que la ley prevea garantías adecuadas. Leyes en Alemania y Cataluña ya restringen la actuación automatizada exclusivamente a potestades regladas. Por su parte, el art. 12 de la Ley gallega de IA de 2025 ha recogido explícitamente la expresión "reserva de humanidad".

La doctrina que se ha ocupado del tema propugna que, en aplicación extensiva del principio de precaución, ampliamente ya utilizado en el ámbito ambiental, se establezca una opción legal por defecto (*opt-out*) a favor del no uso de la automatización total en decisiones discrecionales, salvo ley específica que así lo recoja y pueda justificar que la misma respeta los límites señalados antes, de conformidad con la reserva legal de la Constitución (art. 18.4 Constitución española).

4. Reserva de humanidad y supervisión humana

Es crucial distinguir la reserva de humanidad de la supervisión humana (así, *human-in-the-loop*). Mientras la supervisión actúa una vez aceptado el uso de la IA para vigilar su funcionamiento y mitigar riesgos (especialmente en sistemas de "alto riesgo" según el Reglamento de IA de la UE, art. 14), la reserva de humanidad es una barrera previa que excluye directamente la tecnología de ciertos procesos por considerarla intrínsecamente inadecuada para adoptar ciertas decisiones.

Derechos digitales, *por D. A. Cuesta Bárcena*

El término «derechos digitales» hace referencia al conjunto de derechos que protegen al ciudadano y garantizan su libre desarrollo en el entorno digital. El «entorno digital» está formado por el conjunto de sistemas, aparatos, dispositivos, plataformas e infraestructuras que abren espacios de relación, comunicación, interrelación, comercio, negociación, entretenimiento y creación que permiten a las personas físicas o jurídicas, de forma bilateral o multilateral, establecer relaciones semejantes a las existentes en el mundo físico tradicional (Carta de Derechos Digitales).

El origen de los derechos digitales es diverso. Aunque algunos son de nuevo cuño, han sido reconocidos *ex novo* para dar respuesta a las nuevas realidades que plantea el entorno digital; otros constituyen, en realidad, la ampliación del contenido de derechos preexistentes en su extensión a este entorno o, incluso, meras proyecciones al mismo de derechos clásicos.

A pesar de que en España y en Europa se está avanzando en la sistematización de los derechos digitales, su carácter transversal impide su reunión en un único documento jurídico. Entre las distintas normas que reconocen esta nueva generación de derechos, pueden destacarse el Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos (RGPD) y, especialmente, la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales. Además, en los últimos años se han aprobado diversos textos de *soft law* que, si bien no tienen naturaleza normativa, destacan por su carácter avanzado y detallado, como la Carta de Derechos Digitales o la Declaración Europea sobre los Derechos y Principios Digitales.

La heterogeneidad en el origen de los derechos digitales y su dispersión en el ordenamiento jurídico, así como su número y diversidad, dificultan enunciarlos ordenadamente. Además, hay que tener en cuenta que se trata de un catálogo abierto, llamado a ser ampliado a medida que evoluciona la tecnología, especialmente la IA. En todo caso, parece haber consenso en que en el espíritu de los derechos digitales está garantizar la subordinación de la tecnología al individuo, situándolo en el centro de la transformación digital, preservar la dignidad humana y fomentar el avance conjunto de toda la sociedad, evitando discriminaciones por razones tecnológicas.

Atendiendo a su contenido y siguiendo la sistemática de la Carta de Derechos Digitales, pueden distinguirse cinco grandes categorías de derechos digitales: derechos de libertad, derechos de igualdad, derechos de participación y de conformación del espacio público, derechos del entorno laboral y empresarial y derechos en entornos específicos. A título ilustrativo, puede tomarse como referencia el catálogo de derechos digitales reconocidos en la Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales, base de la Carta. Entre los derechos de libertad, se incluyen el derecho a la protección de datos personales, a la seguridad digital (art. 82) y al testamento digital (art. 96). En relación con los derechos de igualdad, pueden destacarse el derecho de acceso universal a Internet (art. 81) y la protección de los menores y de sus datos en Internet (arts. 84 y 92). Los derechos de participación y de conformación del espacio público comprenden el derecho a la neutralidad de Internet (art. 80), a la educación digital (art. 83), a la libertad de expresión, a comunicar y recibir libremente información veraz y de rectificación en Internet (art. 85) y al olvido (arts. 93 y 94). En el entorno laboral y empresarial sobresalen el derecho a la desconexión digital (art. 88) y a la intimidad del trabajador en el uso de dispositivos digitales (art. 87), frente al uso de dispositivos de videovigilancia y de grabación de sonidos en el lugar de trabajo (art. 89) y ante la utilización de sistemas de geolocalización (art. 90). Finalmente, los derechos en entornos específicos son una categoría residual en la

que se recoge una variedad amplia de derechos digitales, entre los que merecen una especial referencia los relacionados con la IA, como el derecho a no ser objeto de decisiones individuales automatizadas, incluida la elaboración de perfiles, que produzcan efectos jurídicos o afecten significativamente de modo similar (art. 18, en conexión con el art. 22 RGPD).

Actuaciones procesales automatizadas, *por C. San Miguel Caso*

Las actuaciones automatizadas en el ámbito de la Administración de Justicia, contempladas por primera vez en el artículo 56 del Real Decreto-ley 6/2023, de 19 de diciembre, por el que se aprueban medidas urgentes para la ejecución del Plan de Recuperación, Transformación y Resiliencia en materia de servicio público de justicia, función pública, régimen local y mecenazgo, fueron diseñadas para la realización de tareas sencillas y repetitivas, en su mayoría de trámite, caracterizadas por no requerir interpretación jurídica alguna.

Desde una perspectiva conceptual, el citado precepto las define como aquellas actuaciones procesales que, sin necesidad de intervención humana directa, son producidas por un sistema de información adecuadamente programado. En consecuencia, la automatización de determinadas tareas de carácter procedimental se fundamenta en la elaboración, aplicación y ejecución de reglas previamente definidas, llevadas a cabo por sistemas de información programados para producir un determinado resultado de manera automática. De este modo, la ausencia de intervención humana directa en la ejecución de la actuación automatizable se configura como el elemento diferenciador respecto del modelo tradicional de tramitación seguido en las secciones y tribunales de nuestro país, en el que cada actuación procesal requiere de la intervención humana del encargado de su gestión y/o tramitación.

Respecto a la ausencia de intervención humana directa en este tipo de actuaciones, es oportuno recalcar que el cumplimiento de este extremo se debe a dos cuestiones fundamentales: en primer lugar, porque las tareas que pueden automatizarse responden a cuestiones, en su mayoría, procedimentales, que permitirán llevar a cabo una reorganización más adecuada y eficiente de las labores a desempeñar por el personal de la administración de justicia; y, en segundo lugar, porque dicha automatización se centra en actuaciones de escasa sustantividad en relación con el objeto del proceso. Razón por la cual se permite la automatización en aquellas tareas que se encuentran alejadas de requerir una interpretación jurídica para poder ser realizadas. En este último caso, nos encontraríamos ante las denominadas "actuaciones judiciales asistidas" que serán definidas con sustantividad propia en la siguiente entrada.

La regulación de las actuaciones automatizadas se llevó a cabo de forma conjunta con las **actuaciones proactivas**, siendo estas últimas un subgrupo de las primeras. Al igual que sucede con las automatizadas, las actuaciones proactivas tampoco requieren de intervención humana para su realización, ya que estas aprovechan la información incorporada en un expediente o procedimiento de una Administración Pública con un fin determinado para generar avisos o efectos directos a otros fines distintos, en el mismo o en otros expedientes, de la misma o de otra Administración Pública. En este sentido, como su propio nombre indica, este tipo de actuaciones se caracterizan por una auto-iniciación que requerirá, para que esto sea posible, la colaboración entre administraciones públicas, así como el establecimiento de los requisitos que sean necesarios para garantizar la compatibilidad entre las distintas administraciones. Ambas actuaciones se llevarán a cabo en el seno del Comité técnico estatal de la Administración judicial electrónica (CTEAJE), que deberá garantizar y favorecer sin ambages el cumplimiento de estas premisas para su posterior ejecución de forma real y efectiva.

Tanto las actuaciones automatizadas como las proactivas se conciben como herramientas de suma utilidad para lograr la pretendida eficiencia procesal a través de la digitalización o automatización de determinadas labores que, prescindiendo del componente humano, podrían realizarse de forma célere, albergando el mismo resultado. Sin embargo, en el propio Real Decreto-ley 6/2023 se contemplan una serie de elementos, que actúan como garantías y límites a la automatización, cuya concurrencia debe darse de forma simultánea y no alternativa. Al respecto, se exige que tanto las actuaciones automatizadas como las proactivas se puedan identificar, trazar y justificar, permitiéndose, así, conocer su origen y funcionamiento. Igualmente, debe garantizarse la posibilidad de que estas actuaciones puedan ser revertidas o deshabilitadas, dejando sin efecto las mismas. Este límite opera, en nuestra opinión, como una vía alternativa a la automatización, al permitir llevar a cabo la misma actuación de forma no automatizada.

Actuaciones judiciales asistidas, *por C. San Miguel Caso*

Las actuaciones judiciales asistidas tienen su origen conceptual y normativo en el artículo 57 del Real Decreto-ley 6/2023, de 19 de diciembre, por el que se aprueban medidas urgentes para la ejecución del Plan de Recuperación, Transformación y Resiliencia en materia de servicio público de justicia, función pública, régimen local y mecenazgo. Creadas para ser una herramienta de ayuda en la administración de justicia, las actuaciones asistidas consisten en la generación de borradores totales o parciales basados en datos, producidos por algoritmos, con la capacidad de convertirse en el fundamento o apoyo de una resolución judicial o procesal.

La diferencia principal que puede observarse entre estas actuaciones y las de carácter automatizado es que las primeras (las asistidas) precisan de la intervención humana tanto para su puesta en marcha como para su posterior validación y alcance en la decisión judicial, careciendo del carácter auto-iniciado que caracteriza a las actuaciones automatizadas. Por ello, la generación de borradores únicamente se llevará a cabo cuando sea requerido o solicitado por el usuario, entendiendo que podrán solicitar dicho borrador el juez o jueza, magistrado o magistrada, fiscal o letrado o letrada de la Administración de Justicia en el ámbito de sus respectivas competencias. Por otro lado, el valor que se le otorgue a ese borrador dependerá únicamente del usuario y de la validación que este realice sobre la propuesta que el sistema le ha arrojado. Es decir, por sí solo el documento no adquiere el valor de resolución judicial, salvo que el solicitante así lo considere tras un oportuno y concienciado proceso de revisión. En este sentido, es importante destacar que el juez o jueza, magistrado o magistrada, fiscal o letrado o letrada de la Administración de Justicia puede no otorgar ningún valor al documento generado por observar contradicciones, sesgos o alucinaciones del propio sistema, o considerar, por el contrario, que resulta adecuado, pudiendo llevar a cabo sobre él las modificaciones que estime oportunas.

Es importante recalcar que estas actuaciones, por su propia naturaleza, pueden llegar a comprometer, en el seno del proceso judicial, varios de los principios y garantías esenciales del mismo, siempre y cuando no se lleve a cabo un ejercicio de escrupulosa responsabilidad por parte del usuario del sistema. Esto implica, necesariamente, adoptar unas pautas enfocadas a establecer un método de supervisión tanto sobre el sistema de IA que genera el borrador como sobre la actuación posterior del usuario, que, necesariamente, debería demostrar que, al menos, la validación del borrador se ha llevado a cabo de una forma u otra, huyendo así de la mera conformidad basada en una delegación automática e irresponsable de funciones.

Por esta razón, el Consejo General del Poder Judicial (CGPJ), actuando como autoridad de vigilancia del mercado en aquellos casos en los que se apliquen sistemas de IA en la administración de justicia, ha elaborado la instrucción 2/2026 sobre la utilización de sistemas de IA en el ejercicio de la actividad jurisdiccional, proporcionando una enumeración de principios, límites y garantías que deben guiar la actuación de jueces/juezas y magistrados/magistradas en el ejercicio de su potestad jurisdiccional cuando estos se encuentren asistidos de sistemas inteligentes. Esta instrucción tiene un carácter sustancial respecto a las actuaciones asistidas, pero, ciertamente, su ámbito subjetivo se encuentra orientado exclusivamente a los titulares de la función jurisdiccional, excluyendo al resto de potenciales usuarios descritos en el Real Decreto-Ley 6/2023, como el Ministerio Fiscal o los Letrados de la administración de justicia, que, por el momento, carecen de cualquier previsión al respecto.

Sin embargo, sí se establecen desde una perspectiva unitaria, en el citado Real Decreto-Ley, todas aquellas garantías estructurales del sistema que deben apreciarse tanto en las actuaciones asistidas como en las automatizadas o proactivas referentes a especificaciones, programación, mantenimiento, supervisión y control de calidad y, en su caso, auditoría del sistema de información y de su código fuente, que recaerán bajo la responsabilidad de Comité técnico estatal de la Administración judicial electrónica (CTEAJE).

Tercera Parte.

LA IA JURÍDICA Y SUS EFECTOS

I. El Derecho en la era de la IA

Internet y Derecho, *por A. E. Pérez Luño*

Uno de los efectos más visibles de la tecnología en el Derecho contemporáneo es el derivado del desarrollo y utilización de la red. Internet se ha convertido desde los años noventa en el fenómeno central de las nuevas tecnologías de la información y la comunicación, y en el nuevo milenio aparece como un avance decisivo de los sistemas de información y comunicación a escala planetaria. Desde el propio domicilio, cualquier usuario puede acceder a centros de documentación remotos, realizar operaciones financieras y comerciales, disfrutar de una oferta inabarcable de ocio digital y comunicarse con un número indefinido de interlocutores, sin apenas límites de tiempo ni de espacio. La vieja imagen de la "aldea global" ha dado paso a una suerte de "hogar global", en el que cada terminal doméstica se integra en una red universal.

Ese panorama de posibilidades inéditas en la esfera del conocimiento, la actuación y la comunicación ha venido acompañado, sin embargo, de importantes motivos de preocupación. La red se ha utilizado como vehículo de contenidos ilícitos o manifiestamente nocivos, y no han faltado episodios que han sacudido a la opinión pública. Como otras grandes conquistas científicas y tecnológicas, Internet es una realidad ambivalente: sus logros resultan hoy irrenunciables y constituyen un signo del progreso de nuestro tiempo, pero eso no justifica una actitud pasiva frente a los riesgos de "abordaje" criminal que acompañan a la navegación por el ciberespacio.

En este sentido, Internet ha potenciado formas clásicas y nuevas de criminalidad, facilitando la difusión de sabotajes, virus y ataques a sistemas informáticos por un número imprevisible de piratas informáticos. Las llamadas "autopistas de la información" plantean riesgos relevantes para la protección de programas y otros activos inmateriales, mientras que la facilidad de intercambio de datos multiplica los peligros para la protección de la información personal. La

capacidad de difusión ilimitada de imágenes y contenidos convierte a la red en un vehículo especialmente poderoso para lesionar bienes jurídicos básicos: intimidad, imagen, dignidad y honor de las personas, libertad sexual, propiedad intelectual e industrial, mercado y consumidores, seguridad nacional u orden público.

El carácter internacional y deslocalizado de estas conductas complica su descubrimiento, prevención y castigo. Incluso cuando llegan a detectarse, surgen conflictos en torno a la jurisdicción competente y a la determinación de las responsabilidades en una cadena comunicativa en la que se entrecruzan distintos operadores (proveedores de red, de acceso, de servicios y de contenidos), a menudo radicados en Estados con legislaciones diversas. De ahí la paradójica situación de una eficacia global e ilimitada de los ataques contra bienes y derechos, frente a una capacidad de respuesta jurídica fragmentada por las fronteras nacionales. La regulación del flujo interno e internacional de datos se presenta así como uno de los grandes retos tanto de los ordenamientos internos como del Derecho internacional.

Las consecuencias de Internet para el Derecho reclaman, por último, un esfuerzo teórico por no desgajar la reflexión jurídica de los desafíos que plantea la práctica contemporánea. Teoría y práctica aparecen como eslabones de un mismo universo interconectado, marcado por la interdependencia. Frente a las tentaciones aislacionistas y autorreferenciales -las versiones más cerradas de un positivismo formalista o de modelos autopoiéticos del Derecho- cobran sentido marcos teóricos más omnicomprensivos, inspirados en tesis integradoras y sociológicas, que conciben el orden jurídico como un sistema abierto, interdependiente y heteropoiético, en diálogo constante con los entornos técnico, social y económico que lo condicionan.

Ciberciudadanía, *por A. E. Pérez Luño*

Uno de los aspectos más actuales del debate sobre la incidencia de Internet y de la IA en el ejercicio de las libertades es el referente a su impacto en el ámbito de la ciudadanía. En esta esfera se contraponen abiertamente los enfoques de quienes recelan de la Red, al ver en ella un riesgo de despolitización y consiguiente debilitamiento del *status* de ciudadanía activa en las sociedades democráticas; y la postura opuesta que confía en el reforzamiento de la vida cívica a través de su ejercicio por los nuevos cauces que dimanan de Internet y de la IA.

Las posturas críticas no vacilan en reconocer las posibilidades para una renovación política de la vida democrática, cimentadas por la inmensa capacidad informativa y comunicativa que entrañan Internet y la IA. No obstante, plantean algunos reparos de fondo de incuestionable calado. Entienden que la Red y la IA propician un tipo de información y comunicación política individualizada

y personalizada. Cada usuario se construye su propio "menú" de datos y documentación política. Ello puede conducir a una fragmentación, que dificulte la existencia de opciones y programas políticos colectivos y puede menoscabar la vertebración y la cohesión estructural de la experiencia democrática.

Los críticos entienden que la Red y la IA han generado un tipo de usuario-consumidor, han creado, por tanto, unos hábitos de uso que pueden extrapolarse a todos los ámbitos de su empleo. La búsqueda del provecho individual, que es inherente a todas las transacciones comerciales en la Red, puede proyectarse a las actividades políticas. De este modo, el usuario que, en su condición de ciudadano, debe asumir puntos de vista solidarios, puede verse fagocitado por el usuario-consumidor, que proyecta en todas sus actividades en la Red la obtención de beneficios inspirados en el egoísmo.

La visión optimista respecto al reforzamiento de los derechos cívicos a través de Internet y de la IA se ha tematizado bajo la rúbrica de la "ciberciudadanía". Entre las experiencias más estimulantes para la afirmación de la polaridad positiva de la Red, se inscribe el *Manifiesto por el ejercicio de una ciberciudadanía activa, responsable y comprometida.* Dicha declaración fue elaborada por el 1er Congreso ONLINE del Observatorio para la Ciber-Sociedad, celebrado en septiembre de 2002, donde fue aprobada mayoritariamente y en todos sus puntos. El Congreso reunió a 700 cibernautas de todo el mundo.

Los autores del Manifiesto entienden que el acceso a la cultura, el conocimiento y la información nunca estuvo tan al alcance de la humanidad como ahora. La invención y popularización de las Tecnologías de la Información y la Comunicación (TIC) tiene gran parte de responsabilidad sobre este hecho, que supone un cambio cualitativo radical en lo que a esta posibilidad de acceso se refiere. El propósito principal de dicha Declaración se cifra en reivindicar el ejercicio de una ciudadanía electrónica o *ciberciudadanía*, responsable y éticamente comprometida con una utilización de las TIC que trabaje para la consecución de una sociedad más solidaria, justa, libre y democrática. Esa ciudadanía debe estar cimentada en el derecho universal de acceso al ciberespacio y a su defensa y conservación como un ámbito social libre e igualitario. Un derecho que debe estar por encima de monopolios estatales, oligárquicos o empresariales. Para la consecución de ese objetivo se establecen varios puntos que, en una referencia compendiada, pueden englobarse en torno a tres postulados guía: 1º) Proclamación de la *libertad e igualdad del ciberespacio;* 2º) Fortalecimiento de la *cultura cívica;* 3º) Estrategias de *tutela* de la ciberciudadanía.

En defensa de la ciberciudadanía, sus partidarios son conscientes de la facilidad de hacer público todo tipo de información a través de la Red, por lo que abogan por promover la elaboración y defensa de los contenidos que puedan

servir de guía, referencia o información para reforzar la ciberciudadanía desde un ejercicio de responsabilidad ética, que tome en consideración las posibles consecuencias de la información publicada. Entienden que ya no es posible seguir poniéndole vallas al mundo, ni fronteras, ni aduanas, ni peajes monopolísticos u oligárquicos, porque estas vallas, fronteras, aduanas y peajes son testigos de un mundo caduco e injusto al que debemos renunciar para que la Humanidad sobreviva, y porque creen que el ciberespacio es el primer lugar donde esos obstáculos pueden ser, efectivamente, salvados.

Nuevas tecnologías, Derecho y libertades, *por A. E. Pérez Luño*

La era de la informática y de la IA ha contribuido a que se adquiera la convicción de que el *habitat* cívico del presente posee dimensiones planetarias, en la medida en que hoy, con el acceso a Internet, cada ciudadano puede establecer, sin salir de su domicilio, una conversación en tiempo real, sin límites en el espacio ni en las personas.

Los últimos años han acelerado con ritmo creciente los procesos de renovación tecnológica que tan profundamente inciden en las relaciones cívicas. La presencia de las redes de información y comunicación en los ámbitos jurídicos y políticos ha determinado que se adquiera consciencia de que nunca como hoy se había sentido tan intensamente la necesidad de concebir los valores y derechos de la persona como garantías universales. De esa exigencia de universalidad se infiere la reivindicación de que los derechos de la persona se tutelen sin discriminación alguna por razones de raza, de lengua, de sexo, de las religiones profesadas o de las convicciones ideológicas. Se siente hoy con mayor intensidad que en cualquier etapa histórica precedente la exigencia de que los derechos y las libertades no se vean comprometidos por el tránsito de las fronteras estatales; lo que implica tomar en serio el compromiso en pro de la ciudadanía cosmopolita.

Estos requerimientos imponen a la teoría jurídico-política una reflexión sobre las libertades que ya no puede transitar por los cómodos carriles preestablecidos por una larga historia doctrinal e institucional. En un mundo interdependiente, en el seno de sociedades interconectadas, la garantía de los derechos cívicos se halla en directa conexión, para bien o para mal, con los procesos que definen su instalación tecnológica. El estudio actual del Derecho y de los derechos humanos no puede omitir esa referencia contextual, ni puede abdicar del juicio crítico de sus implicaciones. Se trata de lograr que los desarrollos tecnológicos no menoscaben ni se alcancen a costa de las libertades cívicas. Por ello, las reflexiones interdisciplinarias tendentes a establecer un diálogo fluido entre el universo tecnológico, la experiencia jurídica y la esfera de los derechos de los ciudadanos se han hecho cada vez más perentorias.

Al asumir el desafío de nuestra época es posible trazar un balance sobre las actitudes de optimismo y pesimismo sobre el impacto de las nuevas tecnologías y la IA en el Derecho y los derechos. En relación con quienes se aferran a un mantenimiento, cerrado a cualquier transformación renovadora del Derecho y las libertades, opaco a las proyecciones tecnológicas, no se debe caer en la tentación de suponer que se ha logrado ya la realización plena de todas las potencialidades que encierran la experiencia jurídica y las libertades. Es necesario admitir que la siempre cambiante realidad exige la actualización coherente del Derecho y de las libertades y la adecuación cabal a los desafíos que el progreso científico-técnico trae aparejados. La actualización del Derecho y el reforzamiento de las libertades en las sociedades democráticas requiere la vigilancia estricta de los posibles vaciamientos de las instituciones jurídicas democráticas y un deber de pensar los ajustes que el Derecho y el sistema de libertades precisan para enfrentar los peligros que dimanan de los desafíos sociales, culturales y tecnológicos del presente. Al propio tiempo, respecto a la actitud de los que abogan por una inmediata aceptación de un nuevo diseño tecnológico del Derecho y de los derechos humanos, a través de la "panacea" de experiencias teledemocráticas, será preciso no admitir las falsas ilusiones que suelen tender un velo que distorsiona la realidad, al idealizar futuros inalcanzables y vedar el camino hacia soluciones sensatamente realizables.

Una de las cuestiones de mayor actualidad y relevancia en la que se hace patente la exigencia de la "consciencia tecnológica" de los juristas y politólogos es la evaluación del impacto de Internet y de la IA en los sistemas jurídicos actuales. Dicha conciencia tecnológica supone una actitud reflexiva crítica y responsable ante los nuevos problemas que, en las diversas esferas del acontecer social, suscita la tecnología, y ante los que ni el Derecho ni los derechos humanos pueden permanecer insensibles.

Constitucionalismo digital, *por M. A. Presno Linera*

En la doctrina histórico-constitucional se emplea la palabra "constitucionalismo" para aludir al movimiento destinado a limitar el poder del Estado con el fin de garantizar las libertades individuales; su fecha de nacimiento puede establecerse en la Inglaterra del siglo XVII. En las décadas siguientes ese término ha recibido diferentes adjetivaciones (conservador, progresista, racionalista, revolucionario…), siendo la última, o de las últimas, la "digital" para, con ella, aludir, en esencia, a la creciente presencia de lo digital en todos los ámbitos de las sociedades modernas, a su cada vez mayor impacto tanto en los principios que estructuran a los Estados -su carácter democrático, su naturaleza social y su condición de "Estados de Derecho"-, como en los derechos fundamentales, que, por decisión del constituyente, informan el conjunto de nuestra organización

jurídica y política, y, en definitiva, a la necesidad de hacer frente no solo, y como en el pasado, a los potenciales abusos procedentes de unos poderes públicos cada vez más tecnologizados sino, sobre todo, a los de unos poderes en buena medida privados y de alcance supranacional que, merced al enorme desarrollo que ha experimentado en los últimos años la IA y a su monopolización por parte de contadas empresas, han devenido, por utilizar la conocida expresión de Luigi Ferrajoli, en auténticos "poderes salvajes" que operan en un contexto de, en palabras de Shoshana Zuboff, "capitalismo de vigilancia".

Una posible respuesta del constitucionalismo frente a los abusos de estos poderes privados podría venir del recurso a la aplicación, en las relaciones que puedan mantener con ellos los sujetos particulares, de la llamada eficacia horizontal de los derechos fundamentales, de manera que cupiera invocar ante los órganos jurisdiccionales encargados de su garantía la lesión por parte de estos sujetos privados de los derechos y libertades que ya están constitucionalmente garantizados -la vida privada y familiar, la propia imagen, la protección de datos, el derecho a la no discriminación…-, de forma similar a como ha venido ocurriendo en el ámbito de las relaciones laborales o, como ocurrió, en el espacio propiamente tecnológico, con el reconocimiento del "derecho al olvido".

Hay quienes demandan un paso más: constitucionalizar o, cuando menos, incorporar a los tratados internacionales, nuevos derechos -los llamados "neuroderechos"- para, así, combatir, tanto frente a los poderes públicos como frente a los particulares, los peligros que los avances neurotecnológicos podrían implicar para la privacidad mental, la identidad personal, el libre albedrío, el acceso justo al aumento mental y la protección contra el sesgo. A propósito de esta reciente propuesta, mi opinión es que la incorporación de los "neuroderechos" a nuestros textos constitucionales no aportaría nada nuevo porque, en general, las garantías que supuestamente ofrecen ya están contempladas entre las facultades que amparan los derechos a la integridad personal, a la intimidad, a la no discriminación y a la libertad ideológica.

Y respecto a la primera propuesta, parece que su eficacia sería más que dudosa, pues, en primer lugar, para suponer una respuesta aceptable requeriría que la posición de asimetría que caracterizó históricamente a las relaciones privadas laborales o de consumo no supere ciertos límites, y resulta evidente que en el actual contexto digital la posición de las empresas tecnológicas es de abrumadora superioridad respecto a cualquier sujeto privado; en segundo lugar, sería también necesario que existiera una tutela estatal o, en nuestro caso, europea, capaz de imponerse jurídicamente al "capitalismo algorítmico", y el resultado que ofrece el Derecho público llamado a garantizar ese objetivo -el reciente Reglamento europeo de IA- es claramente insuficiente, pues no solo está orientado a tutelar la comercialización, con mínimas exigencias frente a las grandes tecnológicas, de

los sistemas de IA en detrimento de la protección de los derechos, sino que esas precarias cautelas también están en crisis ante las rebajas adicionales que suponen los cambios incluidos en el "ómnibus digital".

No parece exagerado concluir que el "constitucionalismo" se enfrenta al mayor desafío de su historia: a un Leviatán digital resultante de la interacción entre unos poderes públicos dotados de herramientas que hacen posible un control social impensable hasta ahora y unos gigantes tecnológicos privados capaces de facilitar a los Estados esas herramientas y al tiempo de liberarse en la práctica de casi cualquier atadura jurídica que interfiera en su mercado.

Neuroderechos, *por F. H. Llano Alonso*

Los neuroderechos designan un conjunto emergente de derechos humanos orientados a proteger la actividad cerebral y la vida mental de las personas frente a los riesgos que plantean la neurotecnología y la IA. Su aparición responde a un hecho relativamente nuevo: por primera vez en la historia, ciertas tecnologías no solo pueden leer, registrar e interpretar la actividad neuronal, sino también intervenir sobre ella. Las interfaces cerebro-ordenador (BCI, por sus siglas en inglés), por ejemplo, permiten que una persona con parálisis controle un dispositivo con el pensamiento; pero esas mismas herramientas, mal empleadas, abren la puerta a injerencias en aquello que hasta ahora considerábamos el último reducto de lo íntimo y lo humano: nuestros pensamientos, recuerdos y emociones.

La expresión se ha consolidado en el debate jurídico de la última década gracias, sobre todo, al impulso del neurocientífico español Rafael Yuste, director del *BRAIN Project* e impulsor de la *Neurorights Foundation*. Junto a Jared Genser y Stephanie Herrmann (2021), Yuste ha propuesto un catálogo de cinco neuroderechos que se ha convertido en la referencia básica de esta discusión: el derecho a la identidad personal, entendido como el control sobre la propia integridad física y mental; el derecho a la libertad de pensamiento y al libre albedrío para decidir cómo actuar; el derecho a la privacidad mental, que protege nuestros pensamientos frente a su divulgación; el derecho a un acceso equitativo a las técnicas de aumento de las capacidades mentales, de modo que sus beneficios no queden reservados a unos pocos; y el derecho a la protección frente a los sesgos algorítmicos, que evita que estas tecnologías reproduzcan prejuicios y discriminaciones.

Conviene no perder de vista las dos caras de la neurotecnología. En su vertiente terapéutica resulta prometedora: ayuda a tratar enfermedades neurológicas y devuelve autonomía a personas con discapacidad. En su reverso, sin embargo, puede emplearse con fines lesivos -desde la discriminación laboral por una lectura errónea de los datos cerebrales hasta usos coactivos en contextos

militares-, lo que explica la preocupación por dotar a estos bienes de una garantía jurídica sólida.

El reconocimiento de los neuroderechos ha abierto un debate doctrinal vivo y todavía sin cerrar. Para algunos autores estamos ante una auténtica cuarta generación de derechos humanos, ligada a los derechos digitales, que merece su propia formulación normativa; Chile fue, de hecho, pionero al reformar su Constitución en 2021 para proteger la integridad mental. Otros se muestran más cautos: advierten que multiplicar la lista de derechos puede debilitar su fuerza, o sostienen que bienes como la intimidad, la libertad o la dignidad ya están amparados por la Declaración Universal de los Derechos Humanos. Y existe una posición intermedia que entiende los neuroderechos no como derechos enteramente nuevos, sino como una concreción y actualización de garantías ya existentes, adaptadas a un contexto tecnológico inédito.

Más allá de cómo se resuelva esa controversia, los neuroderechos cumplen una función valiosa: nos obligan a preguntarnos qué parte de lo humano queremos preservar como inviolable ante el avance de la IA y la neurociencia. En ese sentido, constituyen menos un punto de llegada que una invitación a la responsabilidad -jurídica, ética y política- frente a unas tecnologías que ya están transformando nuestra manera de entender la identidad, la libertad y la mente.

Tecno-regulación, *por J. Crego Blanco*

La tecno-regulación, en ocasiones denominada *technological management*, se refiere a un modelo de regulación de la conducta humana a través de la tecnología dirigido a excluir la posibilidad de ciertos comportamientos modificando la esfera de lo fácticamente posible. El término suele emplearse para designar aquellos despliegues de tecnología regulatoria que no solamente reducen las posibilidades de una infracción, sino que la hacen imposible. En este sentido, promete una determinación perfecta de la conducta, excluyendo la posibilidad de actuar contra la norma subyacente al diseño tecnológico. A diferencia de otras formas de regulación que actúan directamente sobre los destinatarios de la norma, este tipo de regulación por diseño opera normalmente mediante el diseño de espacios, productos o procesos.

La tecno-regulación surge en el marco de reflexiones sobre el concepto de "regulación". Este concepto involucra diversos mecanismos por los que una autoridad pretende afectar deliberadamente el comportamiento de determinados sujetos. Al menos desde finales del siglo xx se ha identificado una pluralidad de mecanismos regulatorios, que afectan la conducta humana a través de diversos registros. Algunos modelos regulatorios hacen uso del registro moral, tratando de influir en la conducta a través de la calificación de ciertas conductas como buenas

o malas, debidas o prohibidas. Otros modelos emplean el registro prudencial, tratando de determinar el auto-interés del sujeto para que realice la conducta objetivo. La tecno-regulación actúa a través del registro de la practicabilidad o posibilidad, es decir, impidiendo que el sujeto lleve a cabo una determinada conducta. La diferencia radica en que la conducta no se considera simplemente prohibida, sino materialmente imposible. A diferencia del Derecho, la tecno-regulación puede regular la conducta sin necesidad de involucrar la agencia humana, o incluso sin que los sujetos tengan consciencia de que su conducta está siendo regulada.

Suelen emplearse diversos ejemplos relativos a la regulación de la conducción para clarificar el funcionamiento de la tecno-regulación. Actualmente, la conducta al volante se regula principalmente a través de normas de circulación que fijan ciertas velocidades máximas y confían en la presión moral y prudencial sobre quien conduce. Una alternativa tecno-regulatoria implicaría, por ejemplo, que el vehículo se diseñase de tal modo que recibiese información de la velocidad máxima en cierta vía y bloquease la posibilidad de acelerar por encima de dicha velocidad. El "código" o la "arquitectura" del vehículo determinaría la conducta de quien conduce. En este caso, se emplea una tecnología para alterar el registro de la posibilidad e impedir *de facto* que un vehículo acelere por encima de la velocidad establecida. En consecuencia, se consigue una eficacia total de la norma subyacente dirigida a determinar la conducta al volante.

La eficacia total de la tecno-regulación podría tener consecuencias positivas. Por ejemplo, el cumplimiento efectivo de la velocidad máxima podría reducir el número de accidentes. Además, se reducirían costes relativos a la publicitación de las normas, la promoción de determinadas conductas, la persecución de las infracciones o los potenciales litigios.

Independientemente de los posibles beneficios de la tecno-regulación, la literatura se ha centrado en clarificar los riesgos y perjuicios asociados a este modo de regular la conducta humana. Un primer problema deriva de los inevitables fallos de funcionamiento de la tecnología, cuyas consecuencias pueden ser catastróficas. Ante este problema resulta necesario desarrollar mecanismos de compensación o justicia correctiva para afrontar los daños ocasionados. En segundo lugar, la tecno-regulación puede disminuir el grado de agencia o autonomía de los sujetos, debido al registro a través del cual se regula su conducta. En tercer lugar, podría afectar a los propios órganos decisorios, cuya capacidad de considerar cómo regular la conducta podría disminuir. En un sentido similar, la pérdida de agencia podría afectar nuestra capacidad de implicación moral, es decir, la capacidad de reflexionar acerca de cuestiones morales. Por último, la tecno-regulación podría impactar negativamente en ciertos valores políticos y jurídicos, como el Estado de Derecho.

Derecho Cuántico, *por J. A. Nisa Ávila*

La noción "Derecho Cuántico" hace referencia a una propuesta teórica que adapta principios de la física cuántica al Derecho para desarrollar marcos normativos no deterministas y adaptativos basados en la indeterminación, la superposición de estados y la no localidad.

El Derecho Cuántico parte de la idea de que los sistemas contemporáneos de IA generan formas de decisión y representación cuya lógica interna no responde plenamente a esquemas causales clásicos basados en reglas explícitas, trazabilidad lineal y racionalidad completamente interpretable. Desde esta perspectiva, el Derecho Cuántico analiza el Derecho no únicamente como un sistema cerrado de normas estáticas, sino como una estructura dinámica de relaciones, probabilidades, contextos y estados potenciales de interpretación que interactúan constantemente con sistemas tecnológicos complejos.

El Derecho Cuántico funciona como una herramienta epistemológica para explicar por qué las categorías jurídicas tradicionales pueden resultar insuficientes ante arquitecturas neuronales avanzadas. Se parte de la base de que la inseguridad jurídica derivada de redes neuronales profundas no proviene únicamente de errores técnicos, sino de una transformación más profunda de la propia estructura del razonamiento automatizado.

II. Aplicación de la IA en el dominio jurídico

Informática jurídica, *por A. E. Pérez Luño*

La Informática jurídica tiene por objeto la aplicación de la tecnología de la información al Derecho. Es una disciplina bifronte en la que se entrecruzan una metodología tecnológica con su objeto jurídico que, a su vez, condiciona las propias posibilidades o modalidades de tal aplicación. La Informática jurídica estudia el tratamiento automatizado de: las fuentes conocimiento jurídico, a través de los sistemas de documentación legislativa, jurisprudencial y doctrinal (*informática jurídica documental*); las fuentes de producción jurídica, a través de la elaboración informática de los factores lógico-formales que concurren en el proceso legislativo y en la decisión judicial (*informática jurídica decisional*); y los procesos de organización de la infraestructura o medios instrumentales con los que se gestiona el Derecho (*informática jurídica de gestión*).

El pensamiento jurídico, desde finales del pasado siglo, no fue insensible a la urgencia de tomar en serio la tarea de construir una teoría del Derecho abierta,

y responsablemente comprometida con la respuesta a las nuevas necesidades y exigencias de los hombres que viven en la era de la informática. Esa nueva coyuntura reclama de los juristas, los filósofos y los teóricos del Derecho una "consciencia informática", para apelar a una actitud reflexiva, crítica y responsable ante los nuevos problemas que, en las diversas esferas del acontecer social, suscita la informática, y ante los que ni el Derecho, ni quienes lo aplican o lo estudian pueden permanecer insensibles.

Esa exigencia complica sobremanera la labor de los operadores jurídicos y los teóricos del Derecho, porque les obliga a ampliar el angosto horizonte de las autorreferencias normativas, con la apertura hacia los estímulos de la ciencia y la tecnología. Pero solo mostrando sensibilidad a esa exigencia, la teoría del Derecho será capaz de responder a los retos de las sociedades informatizadas que inician la andadura del siglo XXI; lo que es tanto como decir que solo en virtud de esa "consciencia informática" la teoría del Derecho tendrá pleno sentido.

Las repercusiones de la informática en el Derecho están siendo cada vez más extensas e intensas. Dar cuenta de todas ellas resulta una tarea prácticamente inviable, entre otras cosas, por el carácter abierto y dinámico que reviste esa proyección. Baste pensar que, en el horizonte tecnológico del presente, muchos de los problemas y de las soluciones jurídicas tradicionales aparecen irremediablemente caducos. Esa nueva situación impele al pensamiento jurídico a diseñar nuevos instrumentos de análisis y marcos conceptuales prontos para adaptarse a las exigencias de una sociedad en transformación.

En la actualidad, el contexto regulado por el Derecho es el de una sociedad donde la informática ha devenido el símbolo emblemático de nuestra cultura, hasta el punto de que para designar el marco de nuestra convivencia se alude reiteradamente a expresiones tales como la "sociedad de la información", o a la "sociedad informatizada".

El control electrónico de los documentos de identificación, el proceso informatizado de datos fiscales, educativos y médicos, el registro y gestión de las adquisiciones comerciales realizadas con tarjetas de crédito, así como de las reservas de viajes, representan algunas muestras bien conocidas de la omnipresente vigilancia informática de nuestra existencia habitual, y son fenómenos que afectan directamente a la experiencia jurídica.

Informática jurídica documental, *por A. E. Pérez Luño*

La informática jurídica documental tiene por objeto la automatización de los sistemas de información relativos a las fuentes de conocimiento jurídico: legislación, jurisprudencia y doctrina. En las circunstancias actuales el acceso de

los profesionales del Derecho a sistemas adecuados de documentación jurídica informatizada resulta del todo imprescindible. La inflación normativa es sin duda uno de los fenómenos característicos de las sociedades tecnológicamente avanzadas del presente. El flujo incesante de leyes y decisiones jurisprudenciales, cuyo exacto y puntual conocimiento es imprescindible para el funcionamiento correcto del sistema jurídico, hace materialmente imposible su discernimiento, interpretación y aplicación por los operadores jurídicos. La crisis de la información jurídica precipita en las tinieblas al ordenamiento jurídico. La transparencia del sistema normativo, que es presupuesto básico de la certeza del Derecho, se ve suplantada por su creciente opacidad e inescrutabilidad: el Derecho positivo deviene, por tanto, inaccesible para los propios especialistas. El valor de la seguridad jurídica se resquebraja por el menoscabo de la vigencia del ordenamiento jurídico, muchas de cuyas normas permanecen inoperantes por falta de información. La crisis de información y documentación en el Derecho de las sociedades tecnológicamente avanzadas, producida por la avalancha de legislación, jurisprudencia y doctrina, solo puede ser, a su vez, contrarrestada poniendo en juego medios proporcionados, o sea, a través del adecuado empleo de la tecnología informática y los sistemas de teledocumentación. Solo así el jurista, y en concreto el abogado, se halla en condiciones de restablecer un equilibrio entre el incesante flujo de datos jurídicos y su capacidad para asumirlos y aprovecharlos.

Informática jurídica decisional, *por A. E. Pérez Luño*

La informática jurídica metadocumental o decisional se halla integrada por los procedimientos dirigidos a la sustitución o reproducción de las actividades del jurista; a proporcionarle decisiones y dictámenes, es decir, a ofrecerle soluciones de problemas y no mera documentación sobre problemas. Actualmente uno de los sectores más dinámicos y en constante evolución de la informática jurídica metadocumental o decisional es el que se refiere a la aplicación al Derecho de la IA y los sistemas expertos. La IA alude al conjunto de actividades informáticas que si fueran realizadas por el hombre se considerarían producto de su inteligencia. La propia amplitud de estas operaciones, que abarcan desde la comprensión de lenguajes naturales, el reconocimiento de imágenes o sonidos, hasta una amplia y diversa gama de juegos y simulaciones, han determinado una necesidad de acotar y delimitar su ámbito. A ello también ha contribuido la contradicción que supone predicar de entidades ajenas al hombre el rasgo humano por excelencia, o sea, la inteligencia. No obstante, es un sector en constante evolución y sus aplicaciones se multiplican de forma constante.

En lo que hace referencia a los sistemas expertos debe señalarse que tales sistemas incorporan, de una manera práctica y operativa, el conocimiento que posee un experto en la materia de que se trate. Consisten en programas que

reproducen las actuaciones que ha previsto el experto que los diseña. Entre los sistemas expertos más notorios de nuestros días se encuentran los dirigidos al diseño artístico o arquitectónico, la localización de yacimientos minerales y el diagnóstico médico. También han proliferado en estos años una serie de proyectos y prototipos de sistemas expertos jurídicos en materias tales como liquidaciones tributarias, cálculo de indemnizaciones por accidentes laborales o de tráfico, predicción de las consecuencias jurídicas de impactos medioambientales, condiciones de adquisición de la nacionalidad y Derecho de familia, en concreto, matrimonio y divorcio. Estos sistemas pueden prestar un importante servicio al juez y al abogado al informarle sobre la normativa aplicable sobre determinados supuestos, así como sobre las consecuencias jurídicas derivadas de aplicar dicha normativa a situaciones tipo.

Debe tenerse presente que las máquinas pueden procesar informaciones y establecer inferencias lógicas, pero no pueden comprender la multiplicidad de circunstancias que concurren en las conductas humanas. Por ese motivo, en la actualidad no es posible, ni deseable, una suplantación plena del razonamiento jurídico del juez o del abogado por el cálculo informático del ordenador. Solo en aspectos de la experiencia jurídica rutinarios, estandarizados, formalizables, con variables predeterminadas cerradas, es posible recurrir a la IA o a sistemas expertos capaces de ofrecer soluciones operativas. Pero incluso en esos casos el juez o el abogado no pueden abdicar de la responsabilidad de su decisión o de su dictamen para delegarla en el ordenador.

Informática jurídica de gestión, *por A. E. Pérez Luño*

Uno de los sectores informáticos que registran un desarrollo más decidido en los últimos años es el que se refiere a la Informática de gestión, la Ofimática o la Burótica, denominaciones con las que se tiende a traducir la expresión inglesa *Office Automation*. Bajo dicho rótulo se inscriben todos los avances tendentes a la automatización de las tareas rutinarias que se llevan a cabo en cualquier oficina y, por tanto, en las oficinas o despachos jurídicos. Se trata de la realización a través de soportes informáticos o telemáticos de operaciones destinadas a recibir y trasmitir comunicaciones de cualquier tipo, de leer y escribir textos; de formar, organizar y actualizar archivos y registros; exigir y recibir pagos; estipular condiciones y controlar su cumplimiento. Los avances de la Ofimática permiten, en lo que atañe a la gestión de la justicia y la abogacía, automatizar todas aquellas operaciones estandarizadas y que obedecen a pautas regulares y constantes en la escritura, el registro, la trascripción, la contabilidad, la documentación, la comunicación y la certificación. Gracias a la gestión automatizada de la oficina judicial y del despacho profesional del abogado se tienden a lograr resultados más uniformes, imparciales, transparentes, rápidos y económicos. Los empleados de

la oficina jurídica, empezando por el juez y/o el abogado, podrán dedicarse así, de manera exclusiva, a labores que exijan una actividad creadora, o que precisen de la iniciativa personal o deban ser decididos "caso por caso". Es evidente que para el logro de todo ello se requieren programas aplicativos que respondan a las exigencias específicas de cada despacho y a un *software* desarrollado.

Legal Tech*, por M. Barrio Andrés*

Legal Tech es una combinación de los términos «servicios jurídicos» y «tecnología», y significa tecnología jurídica. No existe una definición universal. A mi juicio, podemos definirla del modo siguiente: «la adopción de tecnología y *software* innovadores para racionalizar y mejorar los servicios jurídicos; es decir, herramientas, aplicaciones y plataformas que han sido desarrolladas específicamente para el sector legal».

Existen otras definiciones más amplias y otras más acotadas. En sentido más amplio, se trata simplemente de la aplicación de la tecnología (sobre todo, *software*) a las profesiones jurídicas. Pero esta definición no aporta mucho más para esclarecer su objeto, y abarca una amplia gama de aplicaciones informáticas diferentes que solo tienen una única referencia común a los servicios jurídicos (p. ej., aplicaciones ofimáticas, gestión de expedientes, reconocimiento de voz y dictado digital, automatización de documentos, etc.). Es lo que también se ha denominado «*Tech Office*».

En sentido estricto, y como precisa Hartung, la *Legal Tech* es un *software* que impacta directamente en la prestación de servicios jurídicos, como sería la creación automatizada de contratos o la composición tipográfica, la mecanización del flujo de trabajo, la revisión de documentos, las herramientas de autogestión o las bases de datos inteligentes (como ROSS Intelligence de *IBM* o Jurimetría de *Aranzadi LA LEY*). Estas aplicaciones sirven de apoyo a la labor de los profesionales del Derecho, e incluso pueden sustituir en parte sus actividades. La tecnología a menudo produce mejores resultados y es más barata, por lo que ya no se justifica el uso de abogados para actividades mecánicas o burocráticas que se pueden realizar con un sistema informático. Estas últimas tecnologías (en la terminología de Bues, de *substantive law solutions*) no solo tienen el potencial de convertirse en disruptivas en un futuro lejano, sino que ya están perturbando el sector legal en la actualidad.

Nosotros no vamos a enredarnos aquí en divagaciones sistemáticas no demasiado significantes sobre la naturaleza de las cosas, y consideramos que lo interesante es cubrir todas las aplicaciones *Legal Tech* en sentido amplio. Es decir, no se limita a los abogados o las asesorías jurídicas de empresas, sino también

comprende a las Administraciones públicas, la Administración de Justicia o incluso las herramientas disponibles directamente para los ciudadanos.

Lo esencial es que la *Legal Tech* cambia la forma en la que se accede a los servicios jurídicos. De este modo, bajo tal concepto se incluyen la creación y uso de aplicaciones y servicios que, bien resuelven o automatizan la consulta sobre temas y procesos legales sin la intermediación física de un operador jurídico, bien solucionan o mecanizan el trabajo diario de los operadores jurídicos a través de herramientas que agilizan trámites burocráticos en los que habitualmente se invierten horas, o bien transforman las relaciones entre cliente y abogado a través de plataformas o servicios de consulta *on-line* como el abogado-robot.

Así las cosas, la *Legal Tech* persigue los siguientes objetivos:

- Incorporar herramientas tecnológicas, destacadamente el conjunto de tecnologías agrupadas bajo el supraconcepto de IA.

- Gestionar y automatizar ciertas tareas o funciones.

- Transformar las formas y canales mediante los cuales se ofrecen y prestan servicios legales.

A fin de estructurar las soluciones disponibles, la taxonomía de las empresas y servicios dedicados a la *Legal Tech* comprende:

a) Productos de asesoramiento legal automatizado.

b) Plataformas digitales de encuentro entre clientes y operadores jurídicos.

c) Externalización de procesos legales.

d) Automatización documental.

e) Herramientas de *e-discovery* y revisión documental.

f) Análisis predictivo de casos.

g) Asistentes legales inteligentes.

h) *e-Learning*.

***LawTech**, por J. I. Solar Cayón*

Algunos autores emplean el término *LawTech* para referirse específicamente a aquellas aplicaciones *LegalTech* que se hallan dirigidas directamente, no al profesional del Derecho, sino al consumidor final de los servicios legales. De modo que, conforme a quienes utilizan esta terminología, *LegalTech* sería el concepto general, más amplio, y *LawTech* un concepto específico englobado en aquel. No obstante, el hecho de que *LawTech* designe las aplicaciones de autoayuda legal para el público general hace que, por exclusión, *LegalTech* se emplee normalmente en este contexto para referirse a aquellas orientadas a mejorar la eficiencia en el trabajo de los profesionales jurídicos.

En definitiva, la noción de *LawTech* alude a aquellas aplicaciones tecnológicas que permiten que cualquier usuario pueda obtener una respuesta legal a necesidades jurídicas básicas sin necesidad de recurrir a un profesional del Derecho. Se incluyen aquí, por ejemplo, aplicaciones en línea para la generación de contratos y todo tipo de documentos legales, *chatbots* para el asesoramiento legal en materias básicas, plataformas de intermediación entre abogados y potenciales clientes con consultas en línea o sistemas de negociación automatizada de disputas, entre las más habituales. Aunque muchas de estas aplicaciones pueden ser desarrolladas por los propios bufetes para prestar determinados servicios legales automatizados a sus clientes, la mayoría son desarrolladas y explotadas comercialmente por otro tipo de entidades.

Uno de los efectos más disruptivos del desarrollo de la IA jurídica y la consiguiente comoditización de los servicios legales es la aparición de proveedores alternativos de este tipo de servicios a precios incomparablemente inferiores a los de los profesionales del Derecho. Algunos de estos proveedores alternativos -fundamentalmente compañías tecnológicas y de servicios que prestan servicios legales automatizados- han enfocado su modelo de negocio hacia el mercado corporativo, prestando a las grandes empresas e incluso a los propios bufetes servicios de *outsourcing* especializados en la realización de determinadas tareas jurídicas, pero la mayor parte se dirigen a los pequeños consumidores -particulares, emprendedores, pequeños empresarios y profesionales autónomos- poniendo a su alcance aplicaciones del tipo *do-it-yourself* a través de sus plataformas digitales. De este modo, el desarrollo del *LawTech* ha abierto la puerta de par en par a la era del *self-service* jurídico.

Si bien el desarrollo de este tipo de aplicaciones puede merecer una apreciación positiva desde un punto de vista social, en tanto puede contribuir a reducir la brecha de acceso a servicios legales básicos, particularmente en países donde aquellos no están al alcance de amplios segmentos de la población, conlleva también importantes riesgos para los consumidores. Particularmente teniendo

en cuenta el vacío regulatorio en el que se mueven este tipo de proveedores alternativos de servicios legales en muchos países en los que, si bien el régimen de provisión de estos servicios no está liberalizado, resulta no obstante difícil considerar que aquellos incurren en intrusismo profesional o práctica ilegal del Derecho, ya que lo único que hacen es poner a disposición de los usuarios -y bajo su cuenta y riesgo- aplicaciones para que ellos mismos generen sus propios productos legales (ya se trate de contratos o documentos legales, información jurídica ajustada al caso particular, acuerdos resultantes de una negociación automatizada entre las partes, etc.). Lo que evidencia la necesidad de un nuevo esquema de regulación de los servicios legales que tenga en cuenta estas fórmulas innovadoras de automatización y distribución de los servicios y sus proveedores alternativos.

RegTech*, por J. I. Solar Cayón*

RegTech es un término compuesto a partir de las palabras *Regulation* y *Technology*. Igual que ocurre con otros conceptos de naturaleza similar, como los antedichos *LegalTeh* o *LawTech*, se refiere a la aplicación de la tecnología en un determinado ámbito. En este caso, dicho ámbito es el relativo al cumplimiento normativo (*compliance*), ligado a la gestión de riesgos legales en los procesos operativos de una entidad. Se trata, por tanto, también de un concepto más específico que el de *LegalTech*, y comprendido en el mismo.

El cumplimiento normativo es una de las áreas de la actividad jurídica en las que primero se aplicó la IA y donde esta -en conjunción con otras tecnologías asociadas, como *big data* y *blockchain*- se ha revelado de mayor utilidad, hasta el punto de que actualmente el dominio *RegTech* constituye uno de los sectores más florecientes dentro de la industria *LegalTech*. Esta evolución, por otra parte, corre paralela a la creciente importancia de la función del *compliance* en las organizaciones. El cumplimiento normativo tiene por objeto el análisis de los procedimientos adoptados por empresas, administraciones y, en general, cualquier tipo de organización, para identificar los riesgos legales a los que se enfrentan en el desarrollo de su actividad y el establecimiento de mecanismos internos de prevención, gestión, control y reacción frente a los mismos. Tareas que han adquirido una dimensión central en la gestión corporativa debido a la profusión, complejidad y variabilidad de las exigencias legales (regulaciones sobre conflictos de intereses, blanqueo de capitales y prácticas fraudulentas, protección de datos, buen gobierno corporativo, criterios técnicos, etc.) y a las severas responsabilidades (administrativas, civiles e incluso penales) en las que pueden incurrir tanto sus directivos como la propia organización. La especificidad y complejidad de estas tareas ha propiciado la aparición de nuevos roles profesionales en las organizaciones, como el Director de riesgos legales (*Legal Risk Manager*)

o Director de cumplimiento normativo (*Compliance Officer*), que se mueve a caballo entre los dominios del Derecho, la gestión empresarial y la tecnología, en tanto esta se ha convertido en un elemento imprescindible para su trabajo.

La tecnología *RegTech* está orientada fundamentalmente a la realización de las siguientes funciones:

- Gestión de la información: gracias a su capacidad para procesar y analizar grandes volúmenes de datos, estas herramientas posibilitan el control y gestión de toda la información exigida por las múltiples y complejas regulaciones que inciden en el desarrollo de la actividad de la organización.

- Monitoreo de regulaciones y gestión del riesgo: otra de las tareas fundamentales de la tecnología *RegTech* es la monitorización continua de los cambios normativos, al objeto de identificar aquellos que afectan a la actividad de la organización y evitar los riesgos de incumplimiento legal. Asimismo, posibilitan el análisis de riesgos legales en tiempo real mediante el monitoreo de actividades, la detección de patrones anómalos y la emisión automatizada de alertas, lo que permite identificar riesgos potenciales y mitigarlos antes de que se materialicen.

- Automatización de procesos de *compliance* y de generación de documentación: la tecnología *RegTech* permite integrar las obligaciones regulatorias en los flujos operativos mediante reglas automatizadas, alertas y validaciones en tiempo real. Esta automatización de procesos elimina procedimientos manuales, minimiza errores y mejora la trazabilidad ante inspecciones. También pueden automatizar la elaboración y envío de informes periódicos a organismos supervisores.

Inteligencia artificial jurídica, *por J. I. Solar Cayón*

A lo largo de la última década hemos asistido a la conformación progresiva de un nuevo ámbito de investigación jurídica que viene configurado por la intersección entre la IA y el Derecho. Dentro del ya abundante *corpus* de literatura académica sobre "IA y Derecho" podemos diferenciar dos grandes campos en función del sentido de esa relación: de un lado, el estudio de la regulación jurídica de la IA; y, de otro, el estudio de la aplicación de la IA a las tareas jurídicas. En el primer caso hablamos del "Derecho de la IA", mientras que en el segundo nos adentramos en el dominio de la "IA jurídica". Desde esta perspectiva, entendemos por "sistema de IA jurídica" un sistema basado en IA que ha sido diseñado y/o

entrenado específicamente para automatizar tareas jurídicas o para asistir a los profesionales jurídicos en su realización.

La idea de automatizar los procesos de aplicación del Derecho es una aspiración que hunde sus raíces en los propios fundamentos de la cultura jurídica contemporánea. El proyecto codificador ilustrado se hallaba inextricablemente asociado a una concepción mecánica de la función judicial que entendía la tarea de aplicación de las normas jurídicas como una operación lógica, de carácter silogístico-deductivo. De ahí surge la imagen del juez como un mero autómata, que durante mucho tiempo prevaleció en el pensamiento jurídico positivista. No es extraño, por tanto, que desde el inicio de la computación se planteara la posibilidad de aplicar las nuevas capacidades de las máquinas a los procesos aplicativos del Derecho. Ya en 1949 L. Loevinger presentó la "jurimetría" como un nuevo campo de investigación orientado a la aplicación de los ordenadores al dominio jurídico. Pero será sobre todo con el desarrollo de la IA en las últimas décadas del siglo xx cuando aquella pretensión automatizadora comience a vislumbrarse como una realidad próxima.

La primera expresión de la IA jurídica la encontramos en los denominados sistemas expertos jurídicos, que aparecen en los años setenta. Se trata de aplicaciones que pretendían incorporar de una manera práctica y operativa el conocimiento de un experto en un área jurídica determinada, de modo que el usuario pudiera resolver su problema legal de manera inmediata sin necesidad de recurrir a un profesional. Su diseño se basaba en la codificación de una serie de reglas heurísticas del tipo "si…, entonces…" mediante las que se reproducían las soluciones previstas por el experto para las diversas situaciones posibles y que el sistema aplicaba automáticamente a cada caso particular. Estas herramientas llegaron a adquirir cierta relevancia en materias como liquidaciones tributarias, cálculo de indemnizaciones por accidentes laborales o de tráfico, predicción de las consecuencias jurídicas de impactos medioambientales, condiciones de adquisición de la nacionalidad o determinados aspectos del régimen matrimonial, entre otras áreas jurídicas. Pero su operatividad acabó siendo muy limitada. Por un lado, su utilidad se circunscribía a áreas jurídicas muy específicas en las que el significado de las normas pudiera ser determinado de antemano con gran precisión, de manera que resultara factible la modelización del razonamiento jurídico. Y, por otro, su diseño requería un laborioso proceso humano de determinación de las reglas de inferencia jurídica y de codificación de esas reglas en un lenguaje lógico-formal que la máquina pudiera entender. De hecho, este cuello de botella de la "representación" o "ingeniería" del conocimiento motivó un creciente desinterés por los sistemas expertos no solo en el Derecho, sino en todos los campos, lo que condujo a un "invierno" de la IA en los años noventa.

Esta situación solo pudo ser superada en la segunda década del siglo XXI gracias al desarrollo de un nuevo paradigma de IA basado en la combinación de *big data*, aprendizaje automático y procesamiento del lenguaje natural. En la actual IA basada en datos ya no es necesario que un experto humano codifique unas reglas predeterminadas. La conjunción de las capacidades de aprendizaje automático y *big data* ha posibilitado el diseño de sistemas capaces de procesar grandes volúmenes de datos, de aprehender patrones ocultos en los mismos y de generar modelos predictivos para aplicar esos patrones a nueva información, pudiendo de ese modo automatizar tareas de carácter cognitivo. Por otro lado, las técnicas de procesamiento del lenguaje natural han permitido que los ordenadores "entiendan" el lenguaje humano, lo que ha facilitado enormemente la aplicación de la IA a las tareas jurídicas. Todos estos avances han propiciado el auge de una creciente industria de tecnología legal que pone a disposición de los profesionales jurídicos -o, en algunos casos, de los propios consumidores de servicios legales- un amplio abanico de sistemas capaces de automatizar diferentes tipos de tareas: investigación jurídica, *compliance*, elaboración de contratos y otros documentos jurídicos, revisión documental, predicción de decisiones judiciales, negociación de disputas, asesoramiento legal... Si bien la abogacía fue pionera en el empleo de la IA jurídica, en los últimos años esta revolución tecnológica se ha expandido también a la Administración de Justicia, habiéndose implantado ya en diversos países distintos tipos de aplicaciones: herramientas para la evaluación de riesgos de reincidencia criminal, sistemas de codificación predictiva, sistemas de resolución negociada de disputas en línea e incluso sistemas de toma de decisiones judiciales.

Hoy, con la aparición de la IA generativa, estamos asistiendo a la conformación de una nueva generación de herramientas de IA jurídica basadas en modelos de lenguaje. Mediante técnicas de ajuste fino y de generación aumentada por recuperación (*Retrieved Augmented Generation* o RAG) los modelos de uso general pueden ser customizados para su aplicación específica en el dominio jurídico, dando lugar así a los denominados modelos de lenguaje legal, capaces de generar respuestas y contenidos más precisos, contextualizados y ajustados a las necesidades de los profesionales jurídicos que aquellos. La irrupción de este nuevo paradigma de IA está provocando un cambio en las estrategias de implementación de la IA jurídica en la abogacía: cada vez son más los despachos que, en lugar de utilizar distintos sistemas de IA predictiva o discriminativa para diferentes tareas específicas, están comenzando a emplear un modelo de lenguaje legal, susceptible de realizar múltiples tareas (búsqueda y análisis de las fuentes normativas y jurisprudenciales relevantes, respuesta a consultas legales, redacción de contratos y todo tipo de documentos jurídicos...). Y, en el ámbito judicial, el debate fundamental se centra hoy en su potencial utilización para generar todo tipo de resoluciones procesales y judiciales, incluyendo borradores de sentencias.

Sistemas expertos jurídicos, *por J. I. Solar Cayón*

Los sistemas expertos jurídicos son aplicaciones que, a partir de una base de conocimiento conformada por expertos en una determinada materia jurídica, y mediante la ejecución automática de una serie de reglas de inferencia, proporcionan respuestas a cuestiones legales específicas.

Los sistemas expertos aparecieron en la década de los 70 del siglo pasado y alcanzaron un desarrollo notable en algunas materias jurídicas muy específicas (liquidaciones tributarias, cálculo de indemnizaciones por accidentes laborales o de tráfico, condiciones de adquisición de la nacionalidad, determinados aspectos del régimen matrimonial...) en las que resultaba factible determinar objetivamente el significado de las normas jurídicas y codificarlo mediante reglas precisas y susceptibles de ser traducidas a un lenguaje informático. Sin embargo, el laborioso proceso manual de extracción y codificación computacional de esas reglas por parte de un experto, así como su limitado ámbito de aplicación, motivaron un creciente desinterés por este tipo de sistemas, que prácticamente desaparecieron a finales de siglo.

En la actualidad, la disminución de los costes tecnológicos, la aparición de un nuevo paradigma de computación basado en el análisis de datos mediante algoritmos de aprendizaje automático y la mayor accesibilidad de la información jurídica han propiciado un resurgimiento del interés por este tipo de sistemas para automatizar, sobre todo, tareas de asesoramiento legal, aunque también pueden resultar de utilidad en tareas de cumplimiento normativo (*compliance*).

Un sistema experto jurídico, como cualquier sistema experto, consta de tres elementos básicos:

a) Una base de conocimiento, que es la fuente de información sustantiva que alimenta el sistema. Si el sistema está basado exclusivamente en reglas, al estilo de los desarrollados en las últimas décadas del siglo xx, este conocimiento estará representado en un conjunto de frases, párrafos, reglas... predefinidos por el experto que capturen la lógica de la materia legal correspondiente. El diseño de este tipo de sistemas, dada su estructura lógica deductiva, se representa habitualmente como un árbol de decisiones cuyo tronco de preguntas centrales se va dividiendo, según sus variables, en ramas de posibles soluciones y respuestas. Si, por el contrario, el sistema está basado en los actuales métodos de aprendizaje automático y análisis de datos, la base de conocimiento será una base de datos (*data set*) integrada por un conjunto de materiales jurídicos (normas, sentencias judiciales,

contratos...) considerados relevantes para la resolución de los problemas legales en la materia.

b) Un motor de inferencia que despliega patrones de razonamiento y búsqueda a lo largo de la base de conocimiento para encontrar la respuesta a la consulta efectuada por el usuario del sistema. En el caso de los sistemas basados en reglas, la inferencia consiste en la ejecución de la regla predefinida prevista para la situación particular (si "A", entonces "B"). Por el contrario, aquellos basados en datos no poseen un menú de respuestas predeterminadas, sino que están programados para analizar el *dataset*, detectar patrones en la información y, en base a ese entrenamiento, crear un modelo que permite dar respuesta a las consultas del usuario del sistema mediante el contraste entre el nuevo caso y la información contenida en el *dataset*.

c) Una interfaz gráfica de usuario, esto es, un entorno virtual que permita la interacción del usuario con el sistema, de modo que aquel pueda ingresar la información necesaria para efectuar la consulta legal y recibir la respuesta. Esta interfaz puede ser de distintos tipos: de escritorio, táctiles, de realidad virtual y asistentes de voz. En este sentido, cada vez es más habitual que despachos de abogados, administraciones públicas e incluso tribunales pongan a disposición de los usuarios asistentes virtuales y *chatbots* para proporcionar asesoramiento legal o información jurídica básica de manera totalmente automatizada.

E-discovery (Descubrimiento electrónico)*, por M. A. Adame Martínez*

El término *e-discovery* -contracción de *electronic discovery*- designa el conjunto de procedimientos y técnicas mediante los cuales se identifican, preservan, recopilan, procesan, revisan y producen documentos y datos en formato electrónico en el contexto de un litigio, investigación o procedimiento regulatorio. Es una institución central del proceso civil y penal anglosajón —especialmente del estadounidense y del británico— que carece de equivalente exacto en el Derecho procesal continental, aunque sus técnicas están penetrando de forma creciente en el ámbito europeo.

El *e-discovery* abarca desde la identificación de fuentes de información potencialmente relevantes (correos electrónicos, documentos, mensajes, metadatos, registros de sistemas) hasta su producción ante la parte contraria o ante el tribunal. En cada una de estas fases puede intervenir la IA: los sistemas de *machine learning* permiten clasificar automáticamente grandes volúmenes de documentos por relevancia o confidencialidad, reduciendo drásticamente los

costes y plazos de la revisión manual. Esta aplicación específica recibe a veces el nombre de *Technology Assisted Review* (TAR) o codificación predictiva.

Aunque la tradición procesal continental no contempla el *discovery* como institución, el auge de la litigación transnacional y las normas sobre acceso a documentos en litigios de competencia (Directiva 2014/104/UE) o en acciones colectivas (Directiva 2020/1828/UE) han introducido mecanismos funcionalmente próximos. En el proceso penal, el Reglamento (UE) 2023/1543 (*e-Evidence*) establece el marco para el acceso transfronterizo a pruebas electrónicas entre autoridades de los Estados miembros y con prestadores de servicios. En el ámbito español, el Real Decreto-ley 6/2023, de 19 de diciembre, de medidas urgentes para la ejecución del Plan de Recuperación en materia de servicio público de justicia, introduce medidas de digitalización y eficiencia procesal que incrementan la relevancia de los estándares del *e-discovery* para los operadores jurídicos, en particular en lo relativo a la admisión, custodia y producción de evidencia electrónica.

Cuando los documentos sometidos a revisión contienen datos personales de terceros -lo que es frecuente en litigaciones laborales, mercantiles o regulatorias-, el proceso de *e-discovery* queda sujeto al régimen del RGPD y, en el ámbito nacional, a la Ley Orgánica 3/2018 (LOPDGDD). La base jurídica del tratamiento, la limitación de la finalidad y la minimización de datos son principios que condicionan tanto el alcance de la revisión como los requisitos de seguridad del sistema de IA empleado.

La automatización del proceso de revisión plantea cuestiones sobre el secreto profesional (qué documentos quedan excluidos por estar protegidos por el secreto del abogado), sobre la proporcionalidad de la obligación de producción y sobre la validez jurídica de los errores cometidos por el sistema de IA en la clasificación. La responsabilidad del abogado por una supervisión insuficiente del proceso automatizado es, en este contexto, una cuestión de creciente actualidad deontológica.

Codificación predictiva (*Predictive Coding* o *Technology-Assisted Review*), *por J. I. Solar Cayón*

La codificación predictiva (*Predictive Coding*) o revisión asistida por tecnología (*Technology-Assisted Review* o TAR) constituye tal vez el principal caso de éxito de aplicación de la IA en el sistema judicial. Se trata de una tecnología desarrollada originariamente en el marco del proceso judicial estadounidense para seleccionar y recuperar la información electrónica relevante en un determinado litigio, cuya utilización como un medio válido para que las partes satisfagan el

deber procesal de *discovery* fue admitida jurisprudencialmente en el caso *Da Silva Moore*, en 2012. Hoy no es solo una herramienta de uso habitual en los tribunales estadounidenses, hasta el punto de que algunos de ellos ya han sancionado a abogados por falta de competencia tecnológica en su manejo y comprensión, sino que ha sido aceptada también en otros países anglosajones, como Inglaterra-Gales, Australia, Irlanda o Canadá.

Para entender el éxito de esta tecnología es necesario hacer referencia al contexto procesal en el que surge. Una de las fases más importantes del procedimiento judicial en las jurisdicciones del *common law* es la de *discovery*, en la que cada una de las partes puede requerir a su oponente que revele toda la información obrante en su poder que sea relevante para la determinación de los hechos en los que se fundamentan la acción o la defensa, salvo aquella que esté protegida por algún privilegio procesal. Hoy, en su mayor parte, esa información se halla almacenada en dispositivos electrónicos, razón por la que habitualmente se habla de *e-discovery*. Precisamente, el incremento exponencial de la información electrónica disponible como consecuencia de la digitalización de todas nuestras actividades había generado ya desde comienzos de siglo complicaciones muy serias en el desarrollo de esta etapa procesal: por un lado, el aumento enorme de los costes que acarreaba la revisión manual de la información por parte de amplios equipos de abogados; y, por otro, la dilación del proceso y la multiplicación de disputas acerca del modo en que debía realizarse esa revisión (extensión de la búsqueda, determinación de los criterios de relevancia...).

En este contexto, el desarrollo de los sistemas de codificación predictiva constituyó una respuesta a la necesidad de agilizar y abaratar los costes del *e-discovery*. Estos sistemas utilizan algoritmos de aprendizaje automático supervisado que aprenden los criterios de relevancia jurídica en relación al objeto del litigio a partir del análisis de un subconjunto de los documentos a revisar que han sido previamente codificados o etiquetados (como relevantes o no relevantes) por un abogado experto y buen conocedor del caso. Es decir, el sistema necesita ser entrenado mediante ejemplos: el abogado de la parte obligada a realizar el *discovery* o, de manera conjunta, los abogados de las dos partes, seleccionan y proporcionan al sistema ejemplos positivos y negativos de lo que consideran relevante en relación al problema jurídico que se aborda en el proceso y, a partir de este *dataset*, los algoritmos de aprendizaje automático supervisado generan un modelo estadístico para predecir la probabilidad de relevancia del resto de documentos a revisar. Este entrenamiento es un proceso iterativo, en el que se suceden diversas rondas de revisión de los resultados del sistema, de manera que en cada una de ellas los expertos van detectando los fallos del sistema y proporcionándole nuevos ejemplos para afinar el modelo de relevancia hasta que alcance un nivel de exhaustividad y precisión en la búsqueda que se estime suficiente. El sistema no solo clasifica la información

como relevante o no relevante, sino que la prioriza, asignando a cada documento un grado concreto de probabilidad de relevancia, lo que permite desechar del proceso aquellos que no alcancen un determinado umbral. Todos estos aspectos relativos al entrenamiento del sistema y la utilización de sus resultados son fijados de antemano en un protocolo acordado por las partes, normalmente con la ayuda del juez o de un experto en esta tecnología, que suele comprender tres grandes apartados: obligaciones de transparencia de las partes, tecnologías y metodologías de búsqueda, y método de validación de los resultados.

La exitosa aplicación de la codificación predictiva ha impulsado reformas procesales importantes, basadas fundamentalmente en dos principios: el principio de proporcionalidad, que implica el deber de las partes de tomar en consideración esta tecnología para reducir el coste y la carga de trabajo del *e-discovery*; y el principio de cooperación, que enfatiza la necesidad de articular una pronta comunicación y colaboración entre las partes para acordar un protocolo en el que se definan y desarrollen conjuntamente, y de manera transparente, las metodologías de búsqueda de la información relevante en el litigio.

Sistemas de análisis predictivo*, por A. M. Morales Moreno*

Los sistemas de análisis predictivo son herramientas computacionales orientadas al tratamiento masivo de datos y a la extracción de información útil para la toma de decisiones. Desde un punto de vista técnico, estos sistemas dependen de manera directa de la disponibilidad de grandes volúmenes de datos y de su adecuada explotación. Esto explica que estén estrechamente vinculados con el *Big Data*, el *Data Mining* y el *Machine Learning*, en tanto que representan una aplicación funcional de estas técnicas cuyo objetivo principal es la generación de inferencias probabilísticas sobre acontecimientos futuros o sobre determinados resultados.

En definitiva, estos sistemas se encargan de recopilar y procesar datos, identificar patrones o correlaciones relevantes entre tales datos y, mediante modelos probabilísticos construidos a partir de estos datos, generar predicciones futuras que puedan ser utilizadas por operadores humanos o integradas en procesos automatizados de decisión.

Los sistemas de análisis predictivo constituyen una de las principales manifestaciones de la IA, tanto en general, como en el concreto ámbito del Derecho. En este último, su utilización se orienta tanto a la mejora en la prestación de los servicios jurídicos (el denominado *legaltech*) como a la propia aplicación del Derecho. Del primer grupo, forma parte el uso del análisis predictivo para optimizar la litigiosidad mediante la identificación de patrones en la interposición, evolución y resolución de procedimientos judiciales; así como para

examinar grandes volúmenes de sentencias con el fin de identificar tendencias jurisprudenciales, criterios interpretativos recurrentes o patrones decisorios en órganos jurisdiccionales. Todo ello, con la finalidad última de facilitar la sistematización del conocimiento judicial y anticipar cargas de trabajo.

Por otro lado, estos sistemas permiten evaluar riesgos en materia de fraude fiscal, blanqueo de capitales o incumplimiento normativo. Junto a ello, también destaca la posibilidad de investigar y predecir el crimen mediante la identificación de tipologías delictivas, de perfiles asociados a determinadas conductas ilícitas y de zonas de riesgo o *hot spots* (el conocido como *policing* o IA policial). En última instancia, una modalidad específica de estas herramientas, los sistemas de evaluación del riesgo, se utilizan como apoyo en la evaluación del riesgo de reincidencia delictiva y, en algunos casos, incluso como elementos auxiliares en la toma de decisiones judiciales sobre la determinación de la pena a imponer o la adopción de medidas procesales o cautelares (el *sentencing* o IA judicial).

Los sistemas de análisis predictivo, al igual que ocurre con otras manifestaciones de la IA, ofrecen numerosas ventajas que ya son una realidad, pero igual de evidentes comienzan a ser los riesgos que para los derechos fundamentales se derivan del desarrollo de esta tecnología.

En primer lugar, la calidad de sus resultados depende directamente de los datos con los que son entrenados, por lo que, frente a quienes garantizan una mayor objetividad en la toma de decisiones gracias a estos sistemas, la realidad muestra que los sesgos humanos también se trasladan a los datos que estos utilizan. En consecuencia, confiar ciegamente en sus resultados puede dar lugar a decisiones discriminatorias o a incrementar las desigualdades ya existentes en el ámbito laboral, en el acceso a productos financieros (como préstamos o seguros) y en decisiones de enorme calado como la de si un sujeto debe o no entrar en prisión.

En segundo lugar, ya hemos señalado cómo el funcionamiento de estos sistemas exige el tratamiento masivo de datos, lo que incrementa los riesgos para la privacidad y la protección de datos personales. Al mismo tiempo, la creciente autonomía de determinados modelos y su capacidad para generar resultados difícilmente previsibles refuerzan la necesidad de establecer mecanismos de control que garanticen un desarrollo y utilización compatibles con principios como la transparencia, la motivación de las decisiones o la rendición de cuentas.

Por último, los resultados arrojados por estos sistemas tan solo representan estimaciones probabilísticas, pero no certezas absolutas. Su utilización, por tanto, no debe sustituir la valoración humana, sino servir como instrumento de apoyo.

El extraordinario potencial de los sistemas de análisis predictivo para mejorar la eficiencia en la toma de decisiones debe conciliarse con las garantías propias de un Estado democrático de Derecho. En este sentido, el Reglamento de IA constituye la base legal sobre la que se asienta el desarrollo, comercialización y utilización de estos sistemas. Ahora bien, en la medida en que los sistemas de análisis predictivo no son sino una clase o modalidad de sistemas de IA, la norma no se refiere a ellos de forma concreta. En consecuencia, en función de la finalidad con la que se diseñen o introduzcan en el mercado, así como del ámbito en el que se desplieguen, podrán tener la consideración de prácticas prohibidas, de sistemas de IA de alto riesgo o de sistemas de IA de riesgo limitado, quedando sujetos a un tipo u otro de obligaciones de transparencia, supervisión humana, gestión de riesgos o protección de datos.

Sistema de evaluación de riesgo delictivo*, por M. I. González-Tapia*

Los sistemas de IA de evaluación del riesgo delictivo son aquellas herramientas algorítmicas que integran y tratan automáticamente múltiples datos para estimar la probabilidad de que una persona incurra en el futuro en un comportamiento criminal, normalmente expresada mediante una puntuación o clasificación del riesgo. Aunque tales sistemas pueden tener una vocación generalista, se orientan especialmente a la predicción del comportamiento violento.

Cuando hablamos de este tipo de evaluaciones, con o sin herramientas IA, debe distinguirse entre el riesgo de comisión inicial de una conducta delictiva por quien no ha sido todavía condenado, propio de la peligrosidad social o pre-delictual; y el riesgo de reincidencia o reiteración delictiva de quien ya ha delinquido y ha sido juzgado, propio de la peligrosidad criminal o post-delictual. Esta distinción es decisiva, pues el sistema penal español opera únicamente con respecto a esta última: en la imposición y ejecución de medidas cautelares, de las medidas de seguridad, en la suspensión de la ejecución y en la ejecución de la pena de prisión, para los permisos de salida, la progresión al tercer grado penitenciario o la libertad condicional. En el ámbito policial, en cambio, pueden emplearse sistemas predictivos más amplios, dirigidos a anticipar riesgos vinculados con personas, lugares, momentos o patrones delictivos.

Por su capacidad para integrar numerosas variables y detectar patrones complejos, estos sistemas están llamados a mejorar y apoyar, sin sustituir mediate decisiones automatizadas, los sistemas tradicionales del juicio clínico y los instrumentos actuariales de base estadística, que siempre han adolecido de serias limitaciones. Su utilidad dependerá de la calidad y actualización de los datos, de la validez del modelo y de la posibilidad de explicar y controlar la inferencia realizada.

Un ejemplo paradigmático es COMPAS, aunque no constituya propiamente un sistema de IA de aprendizaje automático, sino una versión informatizada de los tradicionales instrumentos actuariales. Su utilización ha suscitado importantes críticas, relevantes para futuros sistemas IA en sentido estricto, por la falta de transparencia, reproducción de sesgos presentes en los datos policiales y penitenciarios y el riesgo de convertir una puntuación estadística en una conclusión individual definitiva. Otro ejemplo, basado en aprendizaje automático, es la versión de RisCanvi utilizada en Cataluña para apoyar la valoración del riesgo de reincidencia violenta tras la excarcelación. También podría mencionarse el sistema VioGén2, si bien es una herramienta de valoración policial del riesgo de la víctima de violencia de género, pero también se usa por parte de los tribunales, servicios sociales y centros penitenciarios, funcionando materialmente como una valoración prospectiva del riesgo del autor. Aparte de estos instrumentos, no parece haber todavía un sistema de IA consolidado y de uso generalizado en la práctica judicial o penitenciaria.

La neuropredicción o IA neuropredictiva es una categoría específica que consiste en la predicción del riesgo mediante biomarcadores neurobiológicos, estructurales o funcionales, evidenciados por neuroimagen, genética u otras técnicas relacionadas con el funcionamiento cerebral y el comportamiento antisocial. La utilización de estos datos altamente sensibles no está consolidada en su uso forense, por sus problemas de validez, privacidad, ética y discriminación.

Los sistemas IA de evaluación del riesgo delictivo serán, en principio, sistemas de alto riesgo (art. 6 y Anexo III RIA), que habrán de cumplir las exigencias establecidas en los artículos 8 y ss. El art. 5.1. d) RIA, sin embargo, prohíbe, en coherencia con el principio del hecho, las evaluaciones de riesgo delictivo de una persona física basadas exclusivamente en un perfilado subjetivo, de su personalidad y de circunstancias individuales o sociales. Dicha prohibición no regirá cuando tal perfilado se inserte en una herramienta de evaluación que apoye una decisión humana y que ya se base en "hechos objetivos y verificables directamente relacionados con una actividad delictiva". Aunque el texto no es especialmente claro, parece que, en atención al Considerando 42 RIA, tales hechos no habrían de ser delictivos necesariamente, bastando con una indiciaria o sospecha de una actividad delictiva. Así pues, la prohibición se ciñe mínimamente a prohibir modelos que pretendan convertir perfiles personales en pronósticos suficientes, al estilo del Derecho Penal de autor. En este sentido, la neuropredicción quedaría prohibida cuando se pretenda predecir exclusivamente con ella el riesgo delictivo, pero no el uso de datos neurobiológicos dentro de una evaluación más amplia del riesgo, en los términos señalados.

Sistemas de negociación y resolución automatizada de disputas, *por J. I. Solar Cayón*

Los sistemas de resolución de disputas en línea fueron originariamente desarrollados por las grandes plataformas de comercio electrónico ante la necesidad de resolver los conflictos generados en dicha actividad mediante un cauce fácil, rápido y barato que pudiera ser utilizado por sus usuarios desde cualquier lugar del mundo y en cualquier momento. Estos sistemas han ido evolucionando rápidamente al compás de la innovación tecnológica. Y, hoy, la combinación de las técnicas de *big data*, minería de datos y aprendizaje profundo permite analizar rápida y eficientemente los datos generados en los millones de interacciones y transacciones entre compradores y vendedores, extrayendo patrones de conducta que una inteligencia humana sería incapaz de discernir, y diseñar *software*s que automatizan completamente los procesos de negociación y, en su caso, resolución de muchas de estas disputas. Estos *software*s codifican sistemas de negociación colaborativa basados en diversas reglas estratégicas para maximizar el interés común de las partes en sus propuestas de solución. Además, a medida que se incrementa el volumen de datos disponibles y de disputas resueltas, el sistema va aprendiendo a ajustar sus propuestas: identificando y asignando valores a los intereses de las partes, identificando potenciales áreas de acuerdo entre ellas a partir de sus respuestas y preferencias, otorgando prioridades a las distintas opciones de solución en función de las elecciones realizadas por las partes en disputas precedentes o del grado de satisfacción de otras personas con las soluciones propuestas en casos previos similares, etc. De este modo, grandes volúmenes de disputas pueden ser gestionados y resueltos automáticamente a un coste muy bajo. Como ejemplo, baste mencionar que la plataforma de adjudicación digital de disputas *Modria* -desarrollada conjuntamente por *eBay* y *PayPal*- resuelve de forma totalmente automatizada, solo entre usuarios de estas compañías, más de 60 millones de disputas cada año.

Actualmente estas herramientas de negociación automatizada están siendo utilizadas por muchas entidades privadas que han sido autorizadas para la provisión de servicios de resolución de disputas a través de plataformas digitales. Determinadas materias de Derecho de familia -por ejemplo, la negociación de acuerdos de divorcio-, las reclamaciones de deudas monetarias de limitada cuantía y las disputas en materia de consumo son algunos de los sectores legales más propicios para la operatividad de este tipo de sistemas algorítmicos, y en ellos se han centrado fundamentalmente los proyectos pioneros en este campo, aunque también se están extendiendo a otras áreas jurídicas, como las disputas contractuales en determinados ámbitos -por ejemplo, en materia de arrendamiento-, los accidentes de tráfico o las relaciones laborales.

Pero también existen proyectos exitosos de inclusión de herramientas digitales de negociación automatizada en el seno de la Administración de Justicia. Las experiencias más innovadoras en este ámbito vienen proporcionadas por algunos de los tribunales en línea implantados en los últimos años en algunos países. El primero que integró en su seno mecanismos de negociación automatizada para la resolución de disputas en línea fue el *British Columbia Civil Resolution Tribunal*, que desde 2016 resuelve conflictos sobre reclamaciones monetarias, accidentes de tráfico, condominio y en materia de asociaciones y cooperativas. Antes de que un caso pueda llegar a la fase de adjudicación judicial, las partes han de emprender un proceso de negociación en línea a través de una herramienta de negociación automatizada que estructura su interacción mediante el empleo de plantillas y menús desplegables para explorar la posibilidad de alcanzar un acuerdo (con la asistencia, si es necesario, de un facilitador humano). Y procedimientos similares están siendo implantados en el Reino Unido dentro de un plan general de reforma de la Administración de Justicia civil. En el marco de su innovadora estructura de tribunales digitales, en la actualidad ya se hallan en funcionamiento diversas plataformas para la resolución en línea de reclamaciones en distintas materias en las cuales, antes de que el caso tenga que ser resuelto por un juez, las partes son remitidas a un proceso de negociación en línea en el que actualmente se combina el empleo de aplicaciones de negociación automatizada con la asistencia de un facilitador. En una fase posterior, cuando se hayan recopilado datos suficientes de los casos resueltos cuyo análisis permita diseñar sistemas de resolución automática de disputas, está previsto el empleo de sistemas de IA que generen de forma plenamente automatizada propuestas de solución. Si bien estos proyectos son los más significativos, también en EEUU, Holanda, Canadá, Reino Unido, India, Brasil y China existen tribunales que han incorporado herramientas para la resolución amistosa de disputas en línea con diferentes grados de automatización en el procedimiento de negociación.

Sistemas de lectura automática y análisis de contratos, *por J. I. Solar Cayón*

Una de las primeras aplicaciones de la IA jurídica en la pasada década fue el desarrollo de sistemas de lectura automática y análisis de contratos. Se trata de avanzados programas de *software*, basados fundamentalmente en el procesamiento del lenguaje natural y el aprendizaje automático, que pueden ser instalados en el equipo del usuario o a los que se puede acceder a través de plataformas digitales de diversas compañías que ofrecen este servicio en la nube. Estos programas identifican diversos tipos de contratos (independientemente del formato, estructurado o no estructurado, en el que se encuentren), los revisan, captan su estructura, extraen la información esencial de los mismos y sintetizan su contenido, les comparan para identificar elementos comunes y divergencias, detectan anomalías (cláusulas redactadas en un lenguaje inusual, cláusulas

habituales que no existen en determinados contratos, cláusulas adicionales no habituales o incluso páginas perdidas) y les organizan y clasifican conforme a las categorías deseadas por el usuario (según tipos de contratos y de cláusulas, franjas de importes, fechas, partes contratantes, garantías…). Además, son capaces de aprender los objetivos del análisis a partir de los criterios de búsqueda manejados por el usuario y detectar así sus áreas de interés, sugiriendo contenidos relevantes basados en esos intereses. A partir de la información extraída, estos sistemas pueden generar informes automatizados con los parámetros requeridos; presentar los resultados gráficamente, de una manera sencilla y claramente aprehensible, mediante técnicas de visualización de datos, de manera que el profesional pueda "ver" inmediatamente dónde es preciso centrar su atención; o exportar la información obtenida a formatos estructurados de Word, Excel o directamente a la base de datos o el sistema de gestión de contratos de la firma jurídica o la empresa que lo utiliza, de manera que aquella información pueda ser inmediatamente reutilizada.

La funcionalidad básica de estos sistemas es la identificación, extracción, clasificación y presentación de la información. Se trata, pues, fundamentalmente, y desde un punto de vista técnico, de sistemas de recuperación de la información. Pero, a partir de esta información, la interactividad de la herramienta permite que el usuario establezca una serie de reglas o instrucciones al objeto de que el programa realice automáticamente otra serie de tareas adicionales, como por ejemplo enviar automáticamente determinados tipos de contratos a ciertas personas o alertar con una antelación predefinida del advenimiento de fechas relevantes para el seguimiento de la ejecución de los contratos (notificaciones, prórrogas automáticas, expiración…).

La irrupción de la IA generativa ha supuesto también en este campo un salto cualitativo importante, aportando nuevas funcionalidades a estos sistemas, convertidos hoy en asistentes virtuales capaces de entender el contexto legal, extraer con mayor precisión obligaciones e informaciones complejas, proporcionar respuestas interactivas al usuario, evaluar riesgos en tiempo real, y generar de forma automática resúmenes estructurados de miles de contratos o informes de evaluación de riesgos. De ahí que se hayan constituido en herramientas indispensables de *contract management* que las grandes empresas y bufetes están integrando en sus procesos internos de trabajo para racionalizar la gestión de los contratos a lo largo de todo su ciclo vital, en tanto esta es una cuestión esencial para el control de los riesgos de todo tipo (jurídicos, financieros y económicos, operacionales, etc.) y la reducción de costes.

Estas herramientas son especialmente útiles en las tareas de auditoría legal o *due diligence*, que requieren un intensivo trabajo de revisión de grandes cantidades de contratos en un período de tiempo muy limitado. Los procesos de

auditoría legal o *due diligence* consisten en la revisión o examen de la situación jurídica de una empresa u organización. Son frecuentes cuando se pretenden llevar a cabo fusiones o adquisiciones de empresas, financiaciones, contratos de distribución y suministro importantes u otras operaciones de cierta envergadura, al objeto de conocer fielmente el estado jurídico de la empresa objeto del proceso y evitar posibles problemas posteriores a la realización de la operación. Por tanto, uno de los aspectos fundamentales de esta auditoría es la revisión de todos los aspectos contractuales de la empresa, ya que es preciso tener una idea clara de sus obligaciones civiles, mercantiles y financieras. Se trata de un trabajo rutinario que tradicionalmente era realizado por extensos equipos de abogados. La rapidez y la precisión se constituyen en parámetros fundamentales a tener en cuenta para la adecuada realización de esta tarea. Y la IA ha demostrado que no solo es incomparablemente más rápida sino también más precisa que la inteligencia humana en la lectura y el análisis de contratos y otros documentos legales, cuya revisión manual está expuesta a la incidencia de una serie de factores subjetivos que pueden afectar a la corrección de los resultados (inexperiencia de los revisores, divergencias de criterio entre ellos, premura de los plazos, cansancio, etc.).

Esta área del análisis de contratos constituye actualmente uno de los sectores de negocio más pujantes y concurridos en el mercado *legaltech*, habiéndose generado un gran número de compañías tecnológicas que han centrado su actividad en este sector de la actividad jurídica.

Diligencia debida automatizada (*Automated due diligence*), *por M. A. Adame Martínez*

La diligencia debida (*due diligence*) es el proceso de investigación y evaluación sistemática que precede a una operación corporativa significativa -adquisición de empresas, fusión, financiación estructurada, inversión en capital riesgo- con el fin de identificar riesgos jurídicos, financieros, fiscales y regulatorios. La diligencia debida automatizada es la ejecución de ese proceso -total o parcialmente- mediante herramientas de IA, en particular sistemas de lectura y análisis automático de contratos.

Aunque guarda relación con los sistemas de lectura automática y revisión de contratos, la diligencia debida automatizada presenta rasgos propios que justifican su tratamiento autónomo: opera en el marco específico de una operación corporativa, con plazos comprimidos, volúmenes documentales elevados y consecuencias patrimoniales de primera magnitud para las partes. Precisamente esa compresión temporal es la que agrava el riesgo de dependencia acrítica del *output* del sistema, al reducir el margen disponible para la supervisión humana.

Los sistemas de IA aplicados a este proceso permiten: ingerir y clasificar automáticamente grandes volúmenes de documentos contractuales; extraer cláusulas relevantes (limitaciones de responsabilidad, compromisos de no competencia, cláusulas de cambio de control, derechos de rescisión anticipada); identificar desviaciones respecto de plantillas o estándares de mercado; detectar inconsistencias entre documentos; y generar resúmenes e informes estructurados. En los últimos años han alcanzado un nivel de madurez comercial significativo diversas plataformas especializadas en *contract intelligence*, entre las que se cuentan soluciones integradas en entornos de gestión documental legal y asistentes jurídicos basados en modelos de lenguaje de gran escala.

En materia de responsabilidad profesional: el abogado o el asesor que delega en un sistema de IA la identificación de riesgos jurídicos sin supervisión adecuada asume el riesgo de responder por los errores o lagunas del sistema. La diligencia debida automatizada no exime al profesional de su deber de competencia técnica; desplaza el foco de la responsabilidad hacia el diseño del proceso de supervisión.

En materia contractual: los sistemas de IA identifican con mayor fiabilidad cláusulas tipificadas y frecuentes; son más propensos a error en estipulaciones atípicas, en documentos redactados en lenguas distintas de aquellas sobre las que fueron principalmente entrenados, o en estructuras contractuales complejas y multicapa. El resultado debe tratarse siempre como un primer nivel de análisis sujeto a validación humana.

En materia de confidencialidad: la carga de documentación contractual sensible en plataformas de IA de terceros plantea cuestiones sobre el secreto profesional, las obligaciones contractuales de confidencialidad y el régimen de tratamiento de datos aplicable al proveedor del sistema.

Modelo de lenguaje legal, *por J. I. Solar Cayón*

El término "modelo de lenguaje legal" hace referencia a un modelo de lenguaje, grande o pequeño, que ha sido entrenado específicamente para ser aplicado al dominio jurídico.

Con la irrupción de GPT a finales de 2022 y, poco después, de otros modelos de lenguaje de gran tamaño, pronto se suscitó el interés por la IA generativa en el dominio jurídico, particularmente en la abogacía, por su potencial para asistir a los profesionales en diversos tipos de tareas: investigación legal, análisis y razonamiento jurídico, comparación de documentos, generación de resúmenes, revisión de contratos e identificación de cláusulas críticas, y, sobre todo, redacción de toda clase de documentos legales (informes jurídicos, contratos, demandas, escritos de alegaciones, etc.). Esta capacidad de los modelos de lenguaje para

llevar a cabo múltiples tareas y su accesibilidad ha abierto una nueva etapa en el desarrollo de la IA en el ámbito jurídico, poniéndola al alcance de despachos que no podían afrontar los enormes costes que requería el diseño, entrenamiento y mantenimiento de los diversos tipos de sistemas de IA discriminativa. Sin embargo, la utilización de modelos de IA de uso general en el dominio jurídico presenta graves riesgos, en tanto sus resultados pueden resultar imprecisos, descontextualizados o falsos (alucinaciones), debido -entre otras razones- a que han sido pre-entrenados con cantidades masivas de datos sin etiquetar procedentes de todo tipo de fuentes en internet. En este sentido, son ya abundantes los casos de abogados que han tenido que hacer frente a sanciones judiciales y a responsabilidades disciplinarias por vulneración de deberes deontológicos básicos al haber presentado ante los tribunales escritos generados por GPT y otros modelos de uso general que contenían groseras alucinaciones: referencias normativas inexistentes o incorrectas, precedentes judiciales inventados o correspondientes a otras jurisdicciones, citas textuales que no se corresponden con el contenido de la norma o del precedente citado, etc.

Para mitigar tales riesgos estos modelos de IA de uso general pueden ser sometidos a un proceso de "ajuste fino" mediante un entrenamiento adicional enfocado específicamente en la realización de tareas jurídicas, con conjuntos de datos fiables e interacción con profesionales expertos. En la actualidad, la técnica más utilizada para aplicar de manera efectiva los modelos de uso general al dominio jurídico viene constituida por la "generación aumentada por recuperación" (*Retrieval Augmented Generation* o RAG), que consiste en conectar el modelo entrenado a una base de datos indexada en tiempo real -puede tratarse, por ejemplo, de bases de datos de legislación y jurisprudencia, o de la base de datos del propio despacho u organización-, de modo que el sistema recupere de esa base específica la información relevante para generar respuestas más precisas y detalladas. Esta es la razón que explica por qué han sido precisamente las compañías propietarias de bases de datos jurídicas las primeras en aprovechar el potencial de la IA generativa: aquellas pronto incorporaron a sus plataformas de investigación jurídica las capacidades de GPT, aprovechando sus ingentes bases de información y sus avanzadas herramientas de búsqueda y recuperación de la información para customizar ese modelo general mediante la técnica RAG. Ello les ha permitido construir sus propios modelos de lenguaje legal basado en GPT y desarrollar a partir de ellos una serie de aplicaciones para sus clientes.

Pero, aparte de estas opciones comerciales, el desarrollo de las técnicas de customización y entrenamiento de los modelos de lenguaje está permitiendo que incluso los propios despachos desarrollen sus propias aplicaciones de IA generativa. Posibilidad que resulta cada vez más accesible gracias al desarrollo de los modelos de lenguaje pequeños. Mediante la combinación de un modelo de este tipo con un sistema RAG que proporciona acceso a una base de datos del despacho,

este puede construir su propio modelo de lenguaje legal. Estos modelos, que se hallan enfocados a la realización de tareas específicas, presentan importantes ventajas con respecto a los modelos de lenguaje de gran tamaño. En primer lugar, pueden ser "afinados" mucho más rápidamente que estos, utilizando *datasets* de entrenamiento muy precisos y ajustados a cada tarea, lo que reduce el riesgo de alucinaciones. Por otra parte, su tamaño compacto hace que requieran un poder de computación considerablemente menor, de modo que pueden ser implementados en servidores locales. Esto, además de reducir muy significativamente sus costes de desarrollo, hace que no sea necesario trasladar la información a la nube, garantizándose así que los datos confidenciales permanezcan bajo el control del despacho o la organización. En definitiva, la flexibilidad de estos modelos para adaptarse a las necesidades específicas de cada despacho y el hecho de que puedan alojarse en un servidor local está contribuyendo a "democratizar" el uso de la IA generativa en los despachos, que no tienen que afrontar los elevados costes y los obstáculos técnicos asociados a la implementación de los modelos de lenguaje de gran tamaño, al alcance únicamente de las grandes firmas jurídicas.

Justicia digital, *por J. Nieva-Fenoll*

La expresión "justicia digital" posee dos acepciones. La primera designa a una estructura de tribunales que funciona de acuerdo con los últimos avances de la tecnología. La segunda, en cambio, hace referencia a un enjuiciamiento autónomo proveído por las aplicaciones que utilizan IA generativa. Conviene, por tanto, separar ambas nociones.

En cuanto a la primera, se trata de describir cómo funciona una administración de justicia en la que la tecnología ya opera, con mayor o menor intensidad, en todas -o al menos en la mayoría de- las fases de la tramitación y enjuiciamiento. El más elemental de esos aportes es la automatización de los actos procesales reiterativos, que son los que se refieren a la tramitación. A día de hoy, poseyendo la enorme mayoría de ciudadanos un teléfono móvil, puede que una dirección de correo electrónico o incluso algún perfil en una red social, todos los actos de comunicación del tribunal con la ciudadanía y los profesionales de la justicia son mucho más sencillos que antaño y deben hacerse prioritariamente a distancia, hallándose notificada la persona en el acto y permanentemente informada del curso de las actuaciones del proceso que le afecta.

Además, todo el proceso debe constar en una web de acceso intuitivo para los usuarios, de manera que las partes y sus abogados puedan comunicarse entre sí y con el tribunal, con plena fehaciencia y eficacia, lo que facilita la constancia, contenido y coste cero de esas comunicaciones, cosa importantísima para los intentos de utilización de medios alternativos de resolución de conflictos.

Por otra parte, unos tribunales digitalizados deben funcionar asistidos por herramientas de IA generativa y, por descontado, con audiencias telemáticas, salvo que se haga imprescindible la presencialidad en casos que la ley debe acotar. Las citadas herramientas de IA generativa permiten facilitar sobremanera la redacción de resoluciones judiciales, acelerando su elaboración e incluso mejorando eventualmente su redacción en los múltiples casos en que la motivación es repetitiva. Por su parte, las audiencias telemáticas evitan desplazamientos y deben ser la norma cuando no se trata de actos en los que es imprescindible la inmediación presencial por poder aportar algo, desde el punto de vista científico, ese contacto personal. Estas situaciones, no obstante, en absoluto incluyen todos los interrogatorios y en ningún caso deben generalizarse porque son excepcionales, incluso en los procesos penales.

La segunda acepción de "justicia digital", como se ha dicho, supone un enjuiciamiento autónomo con la asistencia de la IA generativa, o con la tecnología que en el futuro permita mejorar su rendimiento.

Para su funcionamiento se precisa, al menos por ahora, que los escritos de parte se elaboren orientados al dato, de manera que la herramienta de IA pueda trabajar con más precisión sobre esos escritos, dejando atrás toda la retórica del pasado, que solo tenía sentido cuando se trataba de persuadir a un juez humano. De ese modo, las solicitudes de la parte consistirán en trabajar con una aplicación en la que se deberán ir dando los distintos datos de la reclamación, que la herramienta, con todo el ordenamiento jurídico en su base de datos, procesará automáticamente dando una respuesta, tras una audiencia en las mismas condiciones a la parte contraria. Además, las pruebas, que en su enorme mayoría serán documentales -documentos multimedia inclusive- y periciales, también serán procesadas en busca de los datos que permitan resolver el conflicto. Con respecto a los interrogatorios que se sigan celebrando, pese a su carácter acientífico a la luz de la Psicología del testimonio, la herramienta ofrecerá asistencia a los jueces en forma de sugerencias que permitan valorar la credibilidad, ordenando la información recibida del interrogado, siguiendo modelos parecidos al SVA-CBCA.

Una herramienta semejante no se puede utilizar en la actualidad, porque sustituye casi por completo a los jueces en casos en que la resolución no es repetitiva y automatizable. Con todo, el futuro está abierto y no se puede descartar ese uso, que puede ser tan sensible para la preservación de la democracia que precisa la elaboración de la herramienta por parte de las autoridades de cada Estado -o grupo de Estados-, siempre con constante e irrenunciable control y supervisión institucional. Es peligrosísimo confiar la elaboración de una herramienta de semejante envergadura a una empresa privada, porque puede condicionar el contenido del ordenamiento jurídico.

Tribunal en línea, *por J. I. Solar Cayón*

Uno de los rasgos más disruptivos de la IA es su potencial para rediseñar los procesos legales, entre ellos los procedimientos de administración de justicia. En esta dirección, en la actualidad, los proyectos más ambiciosos vienen constituidos por la implantación de tribunales en línea para la resolución de determinados tipos de disputas. De este modo, el tribunal deja de ser un lugar para convertirse en un servicio; y el espacio físico de la sala de justicia es sustituido por una plataforma digital.

Esta virtualización del tribunal abre vías innovadoras en las formas de administrar justicia. Por un lado, posibilita la interacción asíncrona en línea entre los participantes en el proceso, que ya no precisan concurrir simultáneamente en un espacio y en un momento determinados: los documentos, pruebas y alegaciones pueden ser presentados y alojados por las partes en la plataforma digital en cualquier momento. Y, por otro, el tribunal puede proveer, a través de una serie de tecnologías integradas en la plataforma, servicios adicionales a la adjudicación judicial, convirtiéndose en una especie de tribunal "extendido" o "ampliado" que asume nuevas funciones en el tratamiento y resolución de los conflictos. Se promueve así el acceso a la justicia en un doble sentido. De una parte, el tribunal en línea permite un acceso más fácil, rápido y barato, desde cualquier lugar y en cualquier momento, a los servicios de la Administración de Justicia. Y, de otra, abre paso a una concepción más amplia de justicia que no se limita a la adjudicación judicial de las disputas, sino que comprende el acceso a una serie de herramientas que empoderan jurídicamente a los ciudadanos.

La iniciativa pionera fue la implantación, en 2016, del *British Columbia Civil Resolution Tribunal* en Canadá, que conoce disputas relativas a reclamaciones de deudas hasta cierta cantidad, accidentes de tráfico, condominio y en materia de asociaciones y cooperativas. Inspirándose en esa experiencia, Reino Unido ha instaurado como elemento central de su reforma de la justicia civil el *Online Solutions Court*, un tribunal digital en el que progresivamente se van poniendo en marcha diversos servicios: *Divorce Online*, para procesos de divorcio; *Online Civil Money Claims*, para reclamaciones de deudas hasta determinada cuantía; *Damages Claims Portal*, para determinadas reclamaciones de indemnización por daños; *Family Public Law and Adoption*, en Derecho de familia; y una plataforma para la resolución de disputas en materia de *Possession*. Al igual que su modelo canadiense, este tribunal está diseñado conforme a una arquitectura tecnológica de tres capas que refleja las diversas funciones comprendidas en esta nueva forma de administrar justicia:

a) La capa de acceso a la plataforma digital tiene como objetivo la prevención del conflicto mediante el asesoramiento legal al potencial

demandante. Este nivel se halla completamente automatizado. A través de una interfaz simple, una aplicación de IA guía al usuario de un modo interactivo, mediante una serie de preguntas estructuradas, hacia la información legal relevante en el caso, ayudándole a entender el Derecho aplicable, a categorizar jurídicamente sus pretensiones, a conocer las opciones y recursos a su disposición, y conectándole con los servicios legales, sociales, asistenciales... pertinentes para la resolución del problema. Si no es posible la auto-resolución del problema mediante estas vías, se pasa a la siguiente capa.

b) El objetivo de la segunda capa es la resolución amistosa de la disputa. En este nivel entran en juego herramientas de negociación automatizada en línea para que las partes puedan alcanzar un acuerdo por sí solas, si bien, en caso de ser necesario, puede intervenir un "facilitador" a través de internet o telefónicamente. En una fase posterior, cuando se hayan recopilado datos suficientes de los casos resueltos cuyo análisis permita diseñar sistemas de resolución automática de disputas, está previsto el empleo de sistemas de IA que generen de forma plenamente automatizada propuestas de solución.

c) La tercera capa es la correspondiente a la resolución judicial de la disputa: si esta no ha podido evitarse ni resolverse amistosamente, se somete al conocimiento de un juez que trabaja en línea. Para ello se establece un proceso estructurado, sin necesidad de concurrencia presencial ni síncrona, que se lleva a cabo a través de internet, con el apoyo de medios telefónicos y de video-conferencia si fuera necesario. Este proceso se basa en un sistema de *continuous online hearing*, de modo que las partes pueden presentar documentos, aportar pruebas y realizar sus argumentaciones, con la intervención activa del juez para guiarles en la explicación y comprensión de sus respectivas posiciones, durante un período razonable de tiempo. Una vez concluido ese plazo, el juez puede emitir su decisión sin necesidad de una vista, si bien, en caso de considerarla necesaria, puede llevarla a cabo de manera no presencial.

Codificación computacional del Derecho (*Law as Code*), *por J. I. Solar Cayón*

La aplicación de la IA al dominio jurídico no solo alcanza a procesos de interpretación y aplicación del Derecho, sino que pretende extenderse también al proceso legislativo para mejorar la calidad de las leyes y su implementación. En este contexto se sitúa la iniciativa *Better Rules*, impulsada desde 2018 por el gobierno neozelandés y que ha inspirado proyectos similares en países como

Dinamarca, Francia, Australia y Canadá. *Better Rules* supone la adopción de un nuevo enfoque estratégico de la tarea legislativa que trata de integrar las nuevas tecnologías -particularmente, la IA- en las operaciones de creación de las normas jurídicas, con el objetivo de mejorar tanto la calidad de los textos legales como el modo en el que pueden hacerse ejecutables a través de medios computacionales.

Para ello promueve, desde la fase inicial de formulación de cualquier proyecto legislativo, la participación conjunta de diseñadores de políticas públicas, legisladores, juristas, administraciones, diseñadores de servicios, expertos en IA y desarrolladores de *software*, para articular de la manera más consistente y eficiente las normas jurídicas y los servicios digitales a través de los que aquellas políticas públicas se implementarán en la práctica. Una vez configurado este equipo multidisciplinar, la metodología *Better Rules* se desenvuelve conforme al siguiente *iter*:

a) Modelado digital de la legislación proyectada.

b) Una vez completado el modelado digital, puede ser utilizado para realizar simulaciones y testar los efectos de la política legislativa en distintos escenarios, al objeto de detectar y corregir potenciales defectos en la formulación del proyecto legislativo (inconsistencias, lagunas, zonas de ambigüedad, efectos indeseados...).

c) Por último, en aquellos casos en que se considere conveniente, el texto legislativo promulgado en lenguaje natural (o parte de él) puede ser también formulado y puesto a disposición de administraciones públicas, proveedores de servicios, usuarios, etc., en código informático susceptible de ser ejecutado por sistemas digitales.

Este último paso nos introduce en la idea de la codificación computacional del Derecho. Es decir, en el proyecto de transformar el Derecho escrito -formulado en un lenguaje natural- en un Derecho computable -formulado en código ejecutable por máquinas, al estilo de un *smart contract*-. El objetivo de este proyecto codificador es la elaboración y promulgación de las normas jurídicas directamente como código informático abierto y reutilizable. Se trataría de que el propio legislador elabore dos versiones igualmente válidas de las normas jurídicas: la formulada en lenguaje natural y su gemelo digital, que podría ponerse a disposición del público a través de APIs y de repositorios digitales oficiales, de manera que estas normas en formatos ejecutables computacionalmente puedan ser implementadas directamente por los usuarios en sus sistemas digitales sin necesidad de ser traducidas desde una fuente original en lenguaje natural. Así, ese gemelo digital podría ser directamente integrado en los sistemas digitales de las administraciones para prestar servicios públicos o automatizar la toma de

decisiones jurídicas, o en los sistemas de los entes regulados para automatizar los procesos de *compliance* y asegurar el cumplimiento de las normas jurídicas en sus procesos operativos. Incluso, esos gemelos digitales podrían ser utilizados por aquellas empresas que quieran desarrollar aplicaciones y servicios basados en tales normas.

De este modo, invirtiéndose el famoso *dictum* de L. Lessig ("*Code is Law*"), se produciría el salto a un nuevo paradigma en el que el Derecho se articularía oficialmente como código informático ("*Law is Code*"). El país que hasta ahora ha ido más lejos en la realización de esta idea es Dinamarca. Ante la creciente complejidad que conlleva la gestión de casos en el extenso esquema del Estado de bienestar de aquel país, en enero de 2018 su Parlamento adoptó un acuerdo para generar «legislación de fácil digitalización», con el objetivo de lograr un marco legal más simple y claro, que resulte más fácil de entender y de trasladar a soluciones digitales seguras y fáciles de usar. En virtud de dicho acuerdo, todas las leyes nuevas deben ser, como regla general, «digitales por defecto», lo que significa que deben cumplir una serie de condiciones que permitan que el gobierno pueda digitalizar total o parcialmente su administración y emplear las nuevas tecnologías para tomar decisiones e implementar mejores servicios públicos.

Smart contract*, por R. Campione*

Un *smart contract* es un protocolo de transacción informatizado que ejecuta los términos de un contrato. Los objetivos generales del diseño de 'contratos inteligentes' se dirigen a cumplir las condiciones contractuales habituales (como las condiciones de pago, los derechos de retención, la confidencialidad, incluso la ejecución), a reducir al mínimo las excepciones, tanto intencionales como accidentales, y a minimizar la necesidad de intermediarios supeditados a la confianza de las partes. Entre los objetivos económicos perseguidos por estas herramientas se incluyen la reducción de las pérdidas por fraude, de los costes de arbitraje y ejecución y de otros costes de transacción.

De este modo, el informático y jurista Nick Szabo bautizaba la noción de *smart contract* en 1994; una nueva tecnología con un funcionamiento históricamente ligado al *blockchain*. De su evolución se puede observar que los *smart contracts* han sido pensados para ser empleados en operaciones financieras como instrumentos ejecutivos de pago y no para regular situaciones jurídicas subjetivas. Se trata, por tanto, de protocolos de transacción que pueden referirse tanto a la fase de celebración y ejecución de un contrato, como únicamente a la fase de ejecución, así como incluso a la ejecución de protocolos que no tienen nada que ver con el contrato. Ello implica que, amén de las cláusulas escritas

contenidas en el contrato, una parte consistirá en un protocolo de actuación que estará en formato electrónico o transcrito a un código de programación.

La idea fundamental de los *smart contracts* es que se pueden integrar en el *hardware* y el *software* con los que se elaboran diversos tipos de cláusulas contractuales (como embargos, fianzas, delimitación de derechos de propiedad), de tal manera que el incumplimiento del contrato resulte costoso para la parte incumplidora.

El *smart contract* prevé todas las posibilidades que puedan surgir durante su ejecución y resuelve de forma automática cualquier problema: precisamente esta sería la oportunidad principal para los operadores, que, de este modo, se ahorrarían los costes de transacción, la intervención de intermediarios y prevendrían el contencioso.

Sin embargo, a pesar de no constituir un problema legal su carácter autoejecutable, cabe destacar un par de cuestiones problemáticas que esta herramienta plantea para la teoría jurídica.

Por un lado, las ventajas mencionadas, al menos de momento, no requerirían una redacción legal específica, tal como ha ocurrido, en cambio, en países de nuestro entorno (por ejemplo, Italia, donde el art. 8ter.2 de la Ley 12/2019 sobre la simplificación para las empresas y la administración pública, define expresamente el *smart contract* como un programa informático que opera a través de tecnologías basadas en registros distribuidos y cuya ejecución vincula de manera automática dos o más partes sobre la base de efectos preestablecidos por las mismas). Si acaso, a la espera de la legislación de otros países y de la jurisprudencia en la materia, podrían resultar necesarias herramientas de armonización normativa a nivel europeo o convenios internacionales. En este punto, doctrina autorizada ha puesto de relieve que, si bien las operaciones económicas y su forma jurídica exigen certeza, se trata de técnicas de negociación, de redacción y ejecución de contratos confiadas a la autonomía de los particulares. Por tanto, estos *smart contracts* habrían tenido relevancia sin necesidad de una definición legislativa, aunque el texto normativo cumple una mayor función de legitimación. Incluso el elemento de producción automática de efectos contemplado por estos modelos contractuales ya estaría en uso en muchos contratos en el ámbito de la *lex mercatoria*, donde, cuando es jurídicamente posible, se garantiza a priori la fase ejecutoria mediante mecanismos de garantía, sobre todo financieros, que entran en funcionamiento en el mismo momento en que se produce el incumplimiento, siguiendo patrones en realidad antiguos, como los del pacto comisorio.

Por otro lado, se trata de programas informáticos que, al autoejecutarse en el momento en que se produce un acontecimiento desencadenante, operan de forma

inmodificable, actuando instrucciones predefinidas con efectos inexorablemente automáticos, conforme a la lógica del *if this, then that*, típica del lenguaje informático. Precisamente esta lógica conferiría al modelo una rigidez contractual que generaría problemas en el plano jurídico: piénsese en la eventualidad de las circunstancias sobrevenidas en el marco de la relación contractual; en la ejecución irreversible de contratos declarados nulos; en la inhibición de las técnicas de integración de la regulación de intereses predefinidos por las partes o en la imposibilidad de medidas preventivas de autotutela contractual. Desde este punto de vista, el esquema conceptual del *smart contract* desafiaría la asentada y básica distinción kelseniana entre el principio de causalidad y el principio de imputación.

III. Efectos de la IA jurídica

Deber de competencia tecnológica, *por J. I. Solar Cayón*

La utilización de la IA en la práctica profesional del Derecho incide de forma directa en algunos de los principios deontológicos básicos de la abogacía. El primer deber deontológico es el de competencia profesional y en los últimos años se ha ido extendiendo la consideración de que dicho deber no comprende solo el conocimiento del Derecho sino también la posesión de ciertas competencias tecnológicas imprescindibles para su práctica, incluyendo algunas relativas al adecuado manejo de sistemas basados en IA.

Dado el liderazgo de los Estados Unidos en el desarrollo de la IA jurídica no es sorprendente que la *American Bar Association* fuera pionera en esta demanda. En 2012, cuando apenas habían comenzado a introducirse los primeros sistemas de IA en la litigación, introdujo una modificación en los Comentarios a sus *Model Rules of Professional Conduct* para incorporar como una dimensión inherente al deber genérico de competencia profesional el de "mantenerse al tanto de los cambios en el Derecho y su práctica, incluyendo los beneficios y riesgos asociados a la tecnología relevante". De manera más específica, en la *Resolution 112 and Report* (agosto de 2019) ha dispuesto que los abogados deben tener "una comprensión básica de cómo funcionan las herramientas basadas en IA" generalmente utilizadas en su profesión, lo que significa "entender las capacidades y limitaciones de la herramienta, y los riesgos y beneficios de sus resultados". Siguiendo la estela de esta asociación de ámbito federal, hoy son ya más de 40 las asociaciones estatales de la abogacía que han incluido en sus reglas deontológicas este deber de "competencia tecnológica", exigiendo en algunos casos su acreditación para poder acceder al ejercicio de la profesión. Algunas herramientas de IA están ya tan integradas en la práctica ordinaria de la abogacía estadounidense que difícilmente podría ser considerado competente

un profesional que no comprendiera su funcionamiento ni supiera utilizarlas. Es el caso de los sistemas de codificación predictiva, de uso común en los tribunales estadounidenses para la selección de los documentos relevantes en litigios en los que es necesario revisar grandes volúmenes de información electrónica. De hecho, diversos tribunales han sancionado ya a abogados por falta de competencia en su manejo y en la comprensión de sus protocolos de uso.

Hoy, la asunción de este deber de competencia tecnológica se ha expandido internacionalmente. En 2022 el Consejo de la Abogacía Europea, en su *Comentario a la Carta de Principios Esenciales de la Abogacía Europea*, lo incorporó dentro del principio de competencia profesional en términos muy similares a los contemplados por la ABA. Se dispone que el abogado tiene la obligación de adquirir "la formación de posgrado (formación profesional continua) apropiada para mantenerse al día de los cambios en el Derecho y su práctica, incluyendo los cambios en el contexto tecnológico y económico relevante en el que trabaja", lo que implica que debe "conocer los beneficios y los riesgos del uso de las tecnologías relevantes en su práctica profesional". De modo similar, en 2024, la *International Bar Association* también introdujo en los *International Principles on Conduct for the Legal Profession* un Comentario sobre el deber de competencia profesional al objeto de incluir en el mismo la competencia tecnológica, especificando que esta "abarca desde el uso eficaz de las comunicaciones electrónicas hasta la interacción con la IA".

En la actualidad, el debate sobre el alcance de este deber se centra fundamentalmente en el manejo de las aplicaciones de IA generativa, dada su rápida expansión y la facilidad de acceso a las mismas por parte de los profesionales jurídicos. En esta dirección, el 29 de julio de 2024, el Comité sobre Ética y Responsabilidad Profesional de la *American Bar Association* emitió la *Formal Opinion 512 "Generative Artificial Intelligence Tools"*. El comité indica que, para que la IA generativa pueda ser utilizada de manera competente, el abogado no necesita ser un experto, pero sí debe tener un entendimiento razonable de las capacidades y limitaciones de la herramienta específica que va a utilizar. A efectos de la evaluación de dicha competencia se aplica, por tanto, un estándar de "esfuerzo razonable" y un "enfoque específico de los hechos". En el ámbito europeo, el Consejo de la Abogacía Europea ha publicado en octubre de 2025 una *Guía sobre el uso de IA generativa por la abogacía*. En ella se indica que el deber de competencia tecnológica en relación al uso de esta tecnología se traduce en la obligación del profesional de verificar sus resultados antes de utilizarlos; entender sus capacidades y limitaciones; comprender los diferentes contextos de uso de esta tecnología, así como las implicaciones y riesgos de su utilización; seguir programas de formación relevantes para garantizar un uso apropiado de la misma; y sujetarse a las directrices y orientaciones específicas de su colegio o asociación, en caso de que existan.

Comoditización de los servicios legales, *por J. I. Solar Cayón*

Uno de los efectos más evidentes de la automatización de diversas tareas legales como consecuencia del desarrollo de la IA jurídica es la comoditización de determinados servicios legales. La identificación de patrones recurrentes en la realización de aquellas tareas gracias a la combinación de *big data* y aprendizaje automático, y el aprovechamiento de las capacidades de la computación en la nube, posibilitan la aplicación de la tecnología para estandarizar procesos y proporcionar servicios legales (redacción de documentos legales, asesoramiento jurídico, servicios de *e-discovery*, tareas de investigación jurídica…) más baratos y de suficiente calidad. De este modo, un creciente abanico de servicios legales están siendo automatizados y transformados en productos genéricos (en el sentido de indiferenciados e intercambiables, al menos desde la perspectiva del consumidor) que pueden ser suministrados a gran escala, sin que existan diferencias apreciables en su calidad, por una amplia variedad de proveedores a un precio notablemente inferior al de los servicios legales tradicionales o convencionales. Además, este nuevo tipo de servicios encuentran su vehículo principal de distribución a través de plataformas digitales e incluso de *apps* diseñadas para todo tipo de dispositivos electrónicos, pudiendo llegar así simultáneamente a millones de potenciales clientes con un coste marginal prácticamente nulo, en tanto que los costes son fundamentalmente costes de conocimiento.

Esta comoditización de los servicios legales no está reñida, sin embargo, con su personalización, en cuanto el producto resultante de las herramientas de IA jurídica es una respuesta o solución ajustada a las circunstancias específicas de cada caso y a las necesidades del usuario de la aplicación. A diferencia de lo que ocurría en los sistemas de producción en masa originados en revoluciones tecnológicas anteriores, uno de los rasgos más característicos de la IA es su potencial de *mass-customization* o "personalización masiva", esto es, la capacidad de producir bienes y servicios que pueden satisfacer las demandas y necesidades específicas de cada cliente manteniendo una eficiencia análoga a la de aquellos sistemas de producción en masa. Ya los sistemas de IA jurídica predictiva o discriminativa estaban diseñados para permitir a sus usuarios participar en la configuración del servicio, y contribuir así a la customización de su propia solución, mediante un proceso de interacción con el *software* a través de una interfaz. En algunos casos este usuario es el profesional del Derecho, quien participa en el proceso de diseño de la aplicación, suministra fuentes de información, o entrena y supervisa el sistema, al objeto de acomodarlo a sus particulares necesidades y propósitos. Así ocurre, por ejemplo, en los procesos de diseño de sistemas expertos, de entrenamiento de los algoritmos de aprendizaje supervisado empleados en los sistemas de codificación predictiva o de configuración de las herramientas de análisis, revisión y gestión de contratos. Pero en otros casos puede ser directamente el consumidor final de los servicios legales quien se convierta en copartícipe del diseño final o de la

generación del producto requerido, como ocurre en los sistemas de elaboración automática de documentos legales o de resolución automática de disputas en línea, en los que a través de un cuadro de diálogo el propio cliente proporciona a la herramienta una información que esta utiliza para configurar una solución jurídica ajustada a las circunstancias particulares del caso. Hoy, con el desarrollo de la IA generativa, la participación del consumidor en la configuración de la solución específica para su problema legal es plena y directa a través del diseño de los *prompts* de entrada.

Precisamente una de las consecuencias más disruptivas de este proceso de comoditización de los servicios legales es su **desintermediación**: esto es, la eliminación del profesional jurídico de la cadena de prestación de determinados servicios ante la posibilidad del consumidor de acceder directamente a los mismos a través de la tecnología. El desarrollo de la IA jurídica y la distribución de los productos generados por la misma a través de plataformas digitales y *apps* de todo tipo ha hecho que determinados servicios legales puedan ser convertidos en servicios o productos de *software* que cualquiera puede adquirir, descargar y usar -a precios muy bajos o de manera completamente gratuita- para resolver problemas jurídicos sin necesidad de contratar un abogado. De este modo, la IA jurídica ha propiciado el inicio de una nueva era: la era del *self-service* jurídico.

Proveedor alternativo de servicios legales, *por J. I. Solar Cayón*

Otro de los efectos disruptivos del desarrollo de la IA jurídica, seguramente el de mayor calado, es la aparición de proveedores alternativos de servicios legales (ALSPs, por sus siglas en inglés) que han quebrado *de facto* el monopolio de la abogacía. La comoditización de algunos servicios legales básicos (redacción de documentos jurídicos, asesoramiento legal en materias básicas, revisión documental...) hace que puedan ser suministrados por una amplia variedad de proveedores sin diferencias apreciables de calidad, o, al menos, con garantías de calidad suficiente. Proveedores que ya no son necesariamente los despachos de abogados, en tanto la automatización posibilita que, independientemente de la regulación de los servicios legales existentes en cada país, estos puedan ser prestados tecnológicamente por agentes alternativos.

Este grupo de proveedores alternativos de servicios legales está conformado por un conglomerado muy heterogéneo de actores cuya característica común es la flexibilidad de sus modelos de negocio y una alta eficiencia basada en el análisis de los procesos y la explotación de las nuevas aplicaciones tecnológicas. Sin duda, el principal motor de este profundo cambio ha sido la industria *LegalTech*: muchas de las compañías tecnológicas que han desarrollado herramientas de IA jurídica se han convertido en proveedoras de servicios legales mediante la explotación

comercial de sus aplicaciones a través de plataformas digitales, constituyéndose así en una competencia directa para la abogacía. Algunas de esas compañías han enfocado su modelo de negocio hacia el mercado corporativo, prestando a las grandes empresas e incluso a los propios bufetes servicios de *outsourcing* especializados en determinadas tareas (investigación jurídica, *e-discovery*, revisión documental...), pero la mayor parte se dirigen a los pequeños consumidores, poniendo a su alcance aplicaciones del tipo *do-it-yourself* y otro tipo de servicios legales anejos.

Un ejemplo paradigmático de este fenómeno y de la dinámica de ocupación de espacios cada vez más amplios en la prestación de servicios legales por parte de estos proveedores alternativos nos lo proporciona la empresa estadounidense *LegalZoom*, pionera en el desarrollo de aplicaciones de IA para la generación de contratos y otros documentos legales. El modelo inicial de negocio de la compañía consistía en la explotación comercial de esas aplicaciones a través de su plataforma digital: cualquier internauta podía acceder a ellas, a precios módicos, para generar sus propios documentos mediante un sencillo proceso interactivo y estructurado de preguntas y respuestas entre un *chatbot* y el propio usuario. En muy poco tiempo se hizo con una clientela de millones de particulares y pequeños empresarios, y el siguiente paso fue la implementación de un sistema de consulta telefónica mediante el que los usuarios de las aplicaciones podían consultar sus posibles dudas legales con cualquiera de los miles de abogados independientes que se adscribieron a la plataforma para obtener unos ingresos extra. Actualmente, aprovechando el potencial de ese *pool* de abogados, oferta también planes de asistencia jurídica de diverso alcance en función de la cuantía de la tarifa plana que pague el usuario. ¿El resultado? *LegalZoom*, sin ser una firma jurídica ni tener contratados abogados, se ha convertido hoy en el principal proveedor de servicios legales en Estados Unidos, con un volumen de negocio muy superior al de los principales bufetes.

El nuevo escenario también ha brindado la oportunidad de convertirse en proveedores alternativos de servicios legales a otro tipo de compañías con un fuerte componente de innovación tecnológica que ya prestaban servicios auxiliares a la abogacía. El caso más claro es el de las compañías suministradoras de información jurídica, como *Thomson Reuters* y *Lexis Nexis*, que con la adquisición de *startups* desarrolladoras de sistemas de IA jurídica se han convertido hoy en plataformas que ofrecen un amplio abanico de servicios de *outsourcing* (análisis predictivo, *e-discovery*, generación automática de documentos jurídicos, análisis y gestión de contratos, *compliance*) a los propios despachos y a departamentos jurídicos de empresas, pasando así de ser meras proveedoras de sistemas de búsqueda y análisis de la información para los profesionales jurídicos a convertirse también en competidores. Una situación similar es la de aquellas compañías especializadas en proporcionar a despachos y departamentos jurídicos profesionales para la

realización de proyectos complejos que normalmente requieren la utilización de sofisticadas tecnologías, las cuales, aprovechando su conocimiento del mercado y su potencial tecnológico prestan también servicios de *outsourcing* legal. También compañías como las *Big Four* de la auditoría están aprovechando los nuevos nichos de negocio corporativo propiciados por la aplicación de la tecnología a los procesos jurídicos y, para reforzar sus estrategias en ese ámbito, han adquirido compañías proveedoras de servicios legales y firmas jurídicas de nuevo cuño caracterizadas por sus procesos innovadores.

En definitiva, hoy puede afirmarse sin ambages que la prestación de servicios legales ha dejado de ser, *de facto*, un monopolio exclusivo de los despachos de abogados.

Legal outsourcing*, por J. I. Solar Cayón*

El *legal outsourcing* (externalización del trabajo jurídico) es uno de los efectos más relevantes provocados por la automatización de determinadas tareas jurídicas y la consecuente desagregación o segmentación de la actividad legal para reasignar cada una de las tareas que la integran al agente, humano o tecnológico, más eficiente en su realización. En este contexto de rediseño de los procesos legales, favorecido también por los efectos de la crisis económica de 2008, fueron fundamentalmente los departamentos jurídicos de las grandes empresas quienes comenzaron a recurrir a esta práctica del *outsourcing*, que consiste en externalizar parte de las tareas jurídicas que ellos mismos realizaban o que confiaban a costosos bufetes, derivándolas hacia compañías tecnológicas o de servicios especializadas en la realización a muy bajo coste de grandes volúmenes de determinadas tareas jurídicas mediante la aplicación de sistemas de IA jurídica. Hoy esta práctica se ha extendido también a las grandes firmas jurídicas, que cada vez derivan una mayor parte de sus tareas "automatizables" a estos *legal process outsourcers* (*LPOs*), permitiendo así que sus profesionales jurídicos se enfoquen en tareas más estratégicas y de mayor valor añadido.

Bajo la denominación de ***legal process outsourcers*** encontramos un amplio abanico de compañías de servicios tecnológicos, pero también consultoras y compañías especializadas en la gestión de procesos o en la gestión de la información que, apoyándose en el *expertise* de su personal (juristas, tecnólogos, analistas de procesos, gestores de proyectos…) y, sobre todo, en la automatización de los procesos mediante la aplicación de herramientas de IA jurídica, son capaces de realizar tareas hasta hace poco reservadas exclusivamente a profesionales de la abogacía a un coste mínimo. La eficiencia de estos proveedores alternativos de servicios legales radica, pues, en su especialización en la realización de tareas jurídicas específicas con el apoyo de las más sofisticadas herramientas tecnológicas

y, en un buen número de ellos, también en la deslocalización de sus centros de trabajo en países en vías de desarrollo donde pueden disponer de una mano de obra cualificada y barata (***offshoring***). India es actualmente el principal foco de la industria global del *outsourcing* jurídico en el ámbito anglosajón, seguida de países como Filipinas, Sudáfrica y Singapur. Esta ubicación proporciona, además, una ventaja adicional a los despachos y departamentos jurídicos que recurren a estos centros de *offshoring*: la diferencia horaria permite un trabajo prácticamente ininterrumpido a lo largo de las 24 horas del día, de manera que tareas que son derivadas a esos centros al finalizar la jornada laboral en el Reino Unido o Estados Unidos pueden estar ya resueltas al comienzo de la siguiente jornada, dando continuidad así a los procesos de trabajo.

En cuanto a las tareas que son objeto de *outsourcing*, es apreciable una evolución significativa. Esta práctica comenzó con la externalización de grandes volúmenes de trabajo legal estandarizado, rutinario y de bajo valor, como las tareas de *e-discovery* y de revisión documental, que tradicionalmente han sido asignadas a los estratos inferiores de las plantillas jurídicas: paralegales (en el ámbito anglosajón), abogados contratados temporalmente y abogados junior. Pero, aunque tales tareas siguen constituyendo una parte muy importante del trabajo derivado a los LPOs, al igual de lo que ha sucedido con otros procesos derivados de la automatización, la entrada de estas compañías en los estratos inferiores del mercado les ha permitido establecer sus propias relaciones con las grandes corporaciones y firmas jurídicas, posicionándose así estratégicamente para expandir la oferta de servicios prestados a las mismas. Servicios que actualmente abarcan también actividades como la elaboración de documentos jurídicos, la supervisión y gestión de la ejecución de los contratos, la investigación jurídica, *due diligence*, *compliance*, análisis de datos y predicción de resultados, análisis de riesgo legal, análisis de procesos jurídicos, *legal project management* e incluso sistemas de resolución automática de disputas en línea para grandes compañías y administraciones públicas. Los LPOs han ido de este modo ascendiendo rápidamente por la escala de valor de los servicios legales, posicionándose hoy como actores muy relevantes dentro de la industria legal global.

Alternative Business Structure*, por J. I. Solar Cayón*

La conjunción de la IA jurídica, la computación en la nube y las TICs ha propiciado la aparición de proveedores alternativos de servicios legales, en tanto ha posibilitado que empresas tecnológicas y otro tipo de compañías estén prestando servicios que hasta hace poco solo podían ser proporcionados por despachos de abogados. De este modo se ha ido produciendo una progresiva liberalización *de facto* del mercado de servicios legales. Y esta realidad está siendo reconocida también *de iure* cada vez en más países.

La aceptación de estos proveedores alternativos ha sido relativamente fácil allí donde el mercado de servicios legales ya se hallaba liberalizado, como Inglaterra-Gales y Australia, donde aquellos pueden ser prestados por estructuras corporativas alternativas a los bufetes que se hallan sujetas a un régimen más flexible en cuanto a su propiedad y funcionamiento. Estas *Alternative Business Estructures* (Inglaterra-Gales) o ***Incorporated Legal Practices*** (Australia) son entidades prestadoras de servicios legales y, frecuentemente, también de otros tipos de servicios profesionales, cuyo principal rasgo diferencial respecto de las firmas jurídicas tradicionales es que tanto su propiedad como su dirección y gestión pueden hallarse mayoritaria o completamente en manos de personas que no son abogados. Bajo tales estructuras encontraron fácilmente cobijo legal todo tipo de proveedores alternativos de servicios legales: despachos multiprofesionales y compañías de servicios tradicionalmente relacionados con aquellos -por ejemplo, empresas de consultoría-, compañías tecnológicas que prestan servicios legales automatizados, compañías que proporcionaban temporalmente a bufetes y departamentos jurídicos de empresas profesionales para desarrollar determinados proyectos y ahora ofrecen sus servicios directamente al consumidor final, e incluso empresas dedicadas a los más diversos objetos sociales, sin relación alguna con el Derecho, en cuya carta de servicios se incluyen ahora también los de carácter jurídico -la primera licencia ABS que se otorgó en Inglaterra fue a una popular cadena de supermercados-. Pero también han sido muchos los bufetes que han adoptado estas nuevas formas jurídicas para aprovechar determinadas ventajas de su régimen jurídico, como la posibilidad de captar financiación externa o la facilitación de la cooperación entre abogados y otros profesionales.

En aquellos países en los que pervive el régimen tradicional de monopolio de la abogacía en la prestación de servicios legales, la imparable progresión de los nuevos proveedores tecnológicos se ha convertido en un factor impulsor de la liberalización del mercado. Así está sucediendo en los EEUU, donde Estados como Arizona, Utah, Tennessee o Colorado han aprobado regulaciones que permiten el desarrollo de modelos innovadores de prestación de servicios legales por parte de firmas cuya propiedad y dirección no se halle en manos de abogados. No es sorprendente que las primeras compañías que han constituido firmas jurídicas acogiéndose a las nuevas fórmulas de *Alternative Business Structures* en estos Estados han sido empresas que proporcionaban servicios de tecnología legal a bufetes y departamentos jurídicos de empresas o que ya prestaban servicios legales automatizados directamente al público a través de aplicaciones de IA. Incluso en aquellos Estados que no han modificado el régimen tradicional, este tipo de compañías están buscando fórmulas jurídicas para convertirse en proveedoras directas de servicios legales. Es el caso de *Norm Ai*, una *startup* que presta servicios *LegalTech* a algunas de las compañías de servicios financieros más importantes del país y que en noviembre de 2025 anunció la creación de una firma jurídica

independiente, *Norm Law LLP*, con sede en Nueva York. Para no vulnerar la regla de este Estado que prohíbe que no abogados posean firmas jurídicas o presten servicios legales, la titularidad y la gestión de *Norm Law* será independiente de la de *Norm Ai*. La firma jurídica utilizará los sistemas de IA de *Norm Ai* para prestar no solo servicios de tecnología legal sino también, directamente, servicios legales a las compañías financieras con las que ya trabajaba su matriz tecnológica y a potenciales nuevos clientes, y con los beneficios generados pagará a *Norm Ai* por el uso de sus herramientas. Pese a la diferencia en las fórmulas jurídicas, en el fondo, la realidad es que *Norm Law* es la firma jurídica de *Norm Ai*, usará su tecnología y estará bajo su control.

En el marco de estas tendencias liberalizadoras promovidas por el desarrollo de la IA jurídica cabe destacar también, en el contexto europeo, que en octubre de 2021 entró en vigor en Alemania la *Ley para promover ofertas orientadas a los consumidores en el mercado de servicios legales*, significativamente conocida como *Ley Legal Tech*, que no solo "regulariza" sino que también -como indica su propio título- "promueve" la prestación de ciertos tipos de servicios legales directamente al consumidor por parte de compañías tecnológicas.

Mercado jurídico virtual (*Online Legal Marketplace*)*, por J. I. Solar Cayón*

Una proporción cada vez más elevada de servicios legales se presta en el espacio virtual de internet. Y, en este ecosistema digital, juegan un papel muy relevante los mercados jurídicos virtuales. El término *online legal marketplace* hace referencia a una plataforma digital de intermediación entre demandantes de servicios legales y la red de abogados *freelance* asociados a aquella. Las compañías propietarias de estas plataformas se convierten así en gestoras de un espacio o punto de encuentro virtual entre abogados y potenciales clientes, lo que les permite instaurar sus propias reglas de juego. Este fenómeno supone la traslación al ámbito jurídico de los sistemas de *crowdworking* característicos de la economía colaborativa (*on-demand economy*), que han adquirido un desarrollo muy notable en otros sectores profesionales. De ahí que la emergencia de este modelo de negocio en el ámbito legal haya llevado a hablar de la "uberización" de los servicios jurídicos. El potencial disruptivo de estas plataformas de intermediación reside fundamentalmente en el hecho de que las compañías propietarias de las mismas, en función de cuál sea su modelo de negocio, y aprovechando la ventaja competitiva que les otorga su no sujeción a las normas que regulan la actuación de las firmas jurídicas, imponen una serie de reglas o condiciones para la participación -a abogados y/o a usuarios- que están alterando la forma y las condiciones de trabajo de los profesionales del Derecho, sus relaciones con los clientes, los sistemas de fijación de precios y las reglas de la competencia.

Las fórmulas de negocio de estas compañías son muy diversas y, consecuentemente, también lo son las reglas de juego de sus plataformas. Una de las más importantes a nivel global basa su modelo de negocio en la publicidad de los propios abogados, interesados en lograr visibilidad en una plataforma que atrae a millones de potenciales demandantes de servicios legales gracias a un impresionante sistema de IA que es capaz de recabar la información disponible de todos los profesionales en cualquier tipo de fuente abierta (datos de colegios de abogados, registros de los tribunales, información de agencias reguladoras, revistas, páginas webs, noticias, etc.) y, mediante un sofisticado algoritmo, evaluar esa información y establecer un ranking de abogados por áreas de práctica jurídica. Otras -la mayoría- basan su modelo de negocio en las comisiones que cobran a los profesionales por cada cliente que obtengan, e intentan atraer a los potenciales clientes mediante agresivas estrategias de fijación de precios. En esta dirección, es bastante frecuente la utilización de diversos procedimientos de subasta: en unos casos es el propio usuario quien fija el precio máximo que está dispuesto a pagar por el servicio que demanda, en otros simplemente expone sus necesidades legales y a partir de ahí recibe las ofertas de varios abogados que un algoritmo selecciona como los más apropiados para llevar a cabo el trabajo requerido. Otras plataformas juegan con procedimientos de subasta entre el abogado y la compañía propietaria de la plataforma: así, en el caso de una de las más importantes de nuestro país, es cada abogado quien determina el precio que está dispuesto a pagar a la compañía por cada contacto con un potencial cliente, mientras que en otra el profesional paga una cuota mensual y, en función de su cuantía, dispone de tokens para conseguir más o menos casos -y de mayor o menor relevancia-.

El proceso de selección de los abogados más apropiados para cada caso viene determinado por algoritmos secretos que suelen establecer unos ratings en función de múltiples factores, como pueden ser las necesidades de los potenciales clientes, las diversas características de los profesionales, el precio que -según el modelo de negocio- aquellos o estos estén dispuestos a pagar, las opiniones de clientes anteriores sobre cada profesional u otro tipo de sistemas de reputación, etc. De modo que mediante esta "algocracia", que gobierna de manera invisible las relaciones que se entablan en la plataforma, estas compañías pueden influir de manera más o menos intensa en la configuración de las condiciones de prestación de los servicios jurídicos. Más allá de los interrogantes deontológicos que plantean algunos aspectos de su funcionamiento, parece evidente que estos mercados virtuales tienen una gran capacidad para redefinir las reglas del mercado tradicional y remodelar aspectos esenciales del trabajo profesional, como la fijación de los precios de los servicios jurídicos, la elección de abogado por parte de los usuarios, las prácticas de competencia entre los profesionales, la forma y las condiciones de trabajo de estos e incluso el propio desarrollo de la relación entre abogado y cliente.

Litigación basada en datos (*Data-driven lawyering*), *por A. J. Sánchez Sáez*

La litigación basada en datos representa el cambio de paradigma desde el "Derecho de biblioteca" hacia el "Derecho de computación". Este concepto describe el uso de herramientas de análisis predictivo y minería de datos para optimizar la defensa de los intereses de las partes en un proceso. Ya no hablamos solo de conocer la ley, sino de prever el comportamiento del sistema judicial mediante el procesamiento de *big data* jurisprudencial.

Para un abogado administrativo, esto implica utilizar sistemas de análisis predictivo para evaluar la viabilidad de un recurso contencioso. Al analizar miles de sentencias previas mediante un modelo de lenguaje de gran tamaño (LLM) específico para el dominio legal, el profesional puede identificar patrones de decisión en magistrados concretos. Esto no debe confundirse con la simple búsqueda de precedentes; es una analítica que utiliza vectores y pesos para calcular probabilidades de éxito. Esta práctica obliga a la Administración a elevar su estándar de defensa, pues se enfrenta a un administrado que conoce, con base estadística, las debilidades de la interpretación administrativa habitual. Es, en definitiva, la singularidad legal aplicada a la estrategia de defensa diaria.

La litigación basada en datos está redefiniendo no solo el ejercicio de la abogacía, sino la propia estructura de la tutela judicial efectiva en nuestro Estado de Derecho. Desde la estricta óptica del Derecho Administrativo, debemos entender que ya no nos encontramos ante una simple fase de digitalización de expedientes o el uso de bases de datos de jurisprudencia; estamos ante un cambio de paradigma profundo donde el lenguaje del Derecho comienza a hibridarse con la estadística inferencial y el análisis de grandes volúmenes de datos. La litigación tradicional se ha cimentado históricamente en la hermenéutica jurídica, es decir, en la interpretación de la norma y su aplicación al caso concreto mediante el silogismo jurídico, donde el abogado busca el precedente para construir un argumento persuasivo de corte lógico-lingüístico. Sin embargo, el *data-driven lawyering* introduce el análisis predictivo como eje central, desplazando la pregunta tradicional sobre qué dice la ley hacia una nueva interrogante sobre qué probabilidades estadísticas existen de que un tribunal específico falle a favor en un procedimiento o proceso determinado. Este paso de la intuición del letrado a la certeza probabilística resulta crucial al impugnar actos administrativos de masa, donde la propia Administración ya utiliza algoritmos para denegar subvenciones, imponer sanciones de tráfico o gestionar criterios de inspección tributaria.

Para comprender esta disciplina en toda su extensión es necesario manejar herramientas que antes resultaban ajenas a los planes de estudio, como la jurimetría, que permite realizar un análisis quirúrgico del juzgador mediante *software* que procesa miles de sentencias previas para identificar patrones en la

toma de decisiones de un magistrado concreto. Esto no solo nos revela su tasa de estimación, sino incluso los argumentos doctrinales que más valora o los tiempos medios de resolución, permitiendo diseñar una estrategia procesal mucho más refinada y menos azarosa. A esto se suma el uso del *e-discovery* y el procesamiento de lenguaje natural (NLP) para el análisis documental masivo, algo vital en litigios contra el sector público donde el volumen de documentación puede ser inabarcable para un equipo humano. Además, el uso de modelos matemáticos permite ahora una cuantificación de daños y riesgos mucho más exacta, reduciendo la histórica incertidumbre que suele rodear a la responsabilidad patrimonial de la Administración Pública.

Sin embargo, esta evolución plantea un desafío directo al principio constitucional de igualdad de las partes en el proceso. Si la Administración utiliza potentes sistemas de *Data Analytics* para fiscalizar al ciudadano, el abogado moderno debe estar capacitado para realizar una contra-auditoría algorítmica. Ya no basta con invocar la ley; la defensa debe atacar la validez del dato y la lógica del algoritmo subyacente si estos generan resultados arbitrarios. Pero no todo es eficiencia; como juristas debemos estar alertas ante el riesgo del determinismo jurídico, que podría petrificar el Derecho si solo litigamos basándonos en lo que la estadística dice que ocurrirá, olvidando que el progreso jurídico nace a menudo de sentencias disruptivas que rompen con el pasado. Asimismo, la brecha económica en el acceso a estas herramientas costosas podría generar una justicia de dos velocidades, vulnerando el derecho fundamental a la defensa. En definitiva, la litigación basada en datos no sustituye al criterio jurídico, sino que lo dota de una base científica necesaria para fiscalizar con éxito el poder de un Estado cada vez más tecnificado y automatizado.

Inversión de la carga de la prueba algorítmica, *por A. J. Sánchez Sáez*

El principio general de nuestro ordenamiento, derivado de la legislación decimonónica y el artículo 1902 del Código Civil, exigía que el dañado probara el defecto, el daño y la relación de causalidad. Sin embargo, cuando nos enfrentamos a sistemas de IA (SIA), este esquema tradicional se desmorona ante lo que la doctrina denomina una "prueba diabólica" para el ciudadano, provocada por la opacidad del efecto "caja negra", la complejidad técnica y la autonomía de estos sistemas, que dificultan identificar al responsable y probar los requisitos de una demanda. En el Derecho Administrativo, donde la Administración ya utiliza algoritmos para denegar ayudas o gestionar servicios, surge una asimetría informativa insalvable que obliga a aplicar la teoría de la facilidad probatoria: quien posee la tecnología y los datos —la Administración o el fabricante— es quien debe probar que no hubo error.

Esta doctrina no es ya una mera aspiración teórica, sino que se encuentra codificada en un complejo entramado normativo. La pieza angular es la Propuesta de Directiva sobre responsabilidad en materia de IA (PD-RIA), de septiembre de 2022, que faculta a los tribunales para ordenar la exhibición de pruebas sobre sistemas de alto riesgo. Si el demandado incumple esta orden, el artículo 4 de dicha propuesta activa una presunción refutable de culpabilidad. Además, introduce la innovadora presunción de la relación de causalidad: si el demandante demuestra la culpa del operador y es razonablemente probable que esta influyera en la salida del sistema, el nexo causal se presume legalmente. Esta facilitación es vital para equilibrar el proceso, especialmente cuando el particular no puede acceder al algoritmo propietario o a la lógica interna de la decisión.

En el marco del derecho a la reparación por productos defectuosos, contamos con la reciente Directiva (UE) 2024/2853, de 23 de octubre, que debe ser traspuesta al ordenamiento español antes de diciembre de 2026 para reformar los artículos 128 a 149 del Texto Refundido de la Ley General para la Defensa de los Consumidores y Usuarios (TRLGDCU). Esta norma refuerza la responsabilidad objetiva, estableciendo que el fabricante responde por el riesgo que introduce al mercado, independientemente de su negligencia. Lo más relevante es que el artículo 9 de esta Directiva introduce una flexibilización de la carga de la prueba en tres escenarios clave: cuando el demandado no exhiba documentación técnica, cuando el daño sea compatible con un mal funcionamiento manifiesto y, fundamentalmente, cuando la complejidad técnica del sistema haga excesivamente difícil para la víctima probar el defecto o la causalidad.

En el ámbito estrictamente administrativo, nuestra Ley 40/2015, de Régimen Jurídico del Sector Público, establece en su artículo 32.1 una responsabilidad patrimonial objetiva que no requiere acreditar culpa, sino solo que el daño sea consecuencia del funcionamiento del servicio público. Si un robot quirúrgico en un hospital público lesiona a un paciente, o un sistema de diagnóstico omite un tumor, la Administración no puede ampararse en la evitabilidad o en la autonomía del sistema. Al contrario, teorías como la del daño desproporcionado o la culpa virtual fuerzan a la Administración a probar que la IA actuó correctamente, pues es ella quien crea el riesgo al integrar estas tecnologías en la prestación de servicios.

Incluso bajo el TRLGDCU actual, los artículos 135 y siguientes ya establecen que el productor es responsable de los daños causados por defectos con independencia de su mala fe, liberando al consumidor de la carga de probar errores en la cadena de diseño. Esta protección se complementa con el Reglamento (UE) 2023/988 de seguridad de los productos, que impone obligaciones preventivas *ex ante*. Como juristas, debemos entender que esta inversión no es solo una regla procesal, sino una salvaguarda de la tutela judicial efectiva frente a un Estado tecnificado; si la Administración decide delegar decisiones en máquinas

inteligentes, debe asumir la carga de demostrar que estas operan dentro de los estándares de seguridad que la ciudadanía tiene derecho a esperar.

Policía predictiva, *por C. San Miguel Caso*

La policía predictiva hace referencia a la utilización de algoritmos y modelos de IA que, a partir del análisis de datos históricos y patrones de conducta, pretenden delimitar zonas de riesgo con un mayor índice de criminalidad, predecir la comisión de futuros delitos, y, en su caso, identificar a sus posibles autores. Para determinar el origen conceptual de este fenómeno debemos remontarnos a la década de 1990, concretamente en Estados Unidos, donde se puso en marcha un novedoso sistema de predicción del delito denominado COMPSTAT que fue inicialmente utilizado por el departamento de policía de Nueva York, extendiéndose posteriormente a otras ciudades. Con este nuevo sistema se introduce un cambio de paradigma en la labor policial enfocado en una mayor eficiencia de la gestión basada en datos. Sin embargo, este fenómeno evolucionó significativamente en la década de los 2000, extendiéndose a otros países y experimentando un auge notable que aún no ha cesado y que estuvo motivado, en sus orígenes, por la irrupción de las nuevas tecnologías en la propia comisión de los hechos delictivos, encontrándonos ante nuevos tipos delictivos que, necesariamente, requerían de técnicas más certeras orientadas a una mayor eficiencia policial.

Terminológicamente, debemos diferenciar esta figura de las investigaciones prospectivas, pues difieren significativamente en cuanto a su naturaleza. Mientras que la policía predictiva se basa en estadísticas y predicciones con base en datos de carácter general, las investigaciones prospectivas (contrarias a nuestro ordenamiento jurídico) toman como referencia la observación directa de conductas individuales sobre sujetos concretos, atentando significativamente contra sus derechos fundamentales.

La policía predictiva tiene, además, como objetivo principal lograr una mejora significativa de la gestión policial, optimizando todos los recursos disponibles y coadyuvando, así, al fomento de la seguridad pública a través de la reducción de los conflictos y de la prevención como elemento disuasorio de la criminalidad. En consecuencia, la predicción, la reducción y la protección son los tres ejes centrales y vertebradores de esta figura. Teniendo como finalidad la potenciación de la prevención, no debe pasar desapercibido que la introducción de la IA en la prevención e investigación de los delitos implica una modificación sustancial en los protocolos habituales de actuación, pasando de un enfoque reactivo, que únicamente se produce una vez que han acaecido los hechos, a un enfoque de tipo preventivo, que conducirá a adoptar, en las distintas unidades de policía, una perspectiva de mayor alcance, más general y menos focal. No obstante,

la predicción de futuros delitos presenta unos escollos particulares -sobre todo en delitos esporádicos o de carácter emocional- que impedirán prescindir del enfoque reactivo. Se trata, por lo tanto, de invertir la balanza a través de una nueva metodología inteligente y tecnológica en donde la prevención se imponga, paulatinamente, a la perspectiva tradicionalmente reactiva empleada para toda clase de delitos.

En relación con los sistemas predictivos empleados por la policía, estos son múltiples y muy variados dependiendo del objetivo que pretendan conseguir. Sin embargo, es posible establecer una clasificación, diferenciando los *placed-based*, aquellos que permiten determinar las zonas geográficas más conflictivas identificando los puntos calientes de la delincuencia, de los *person-based,* sistemas que, por el contrario, toman como referencia a la persona, y no al lugar, para predecir la probabilidad o riesgo de reincidencia.

Es importante tener en cuenta que estos sistemas se basan en datos históricos que precisan de unas exigencias adicionales para reflejar la realidad delictiva y no proporcionar una distorsión de la misma. Por esta razón, los datos empleados deben ser representativos de la realidad que pretenden reflejar y, a su vez, se debe exigir una actualización constante de los mismos con una alta asiduidad pues, de lo contrario, podría producirse la paradoja de que estos sistemas llevaran a resultados erróneos, invirtiendo los objetivos y finalidades previamente descritos. La actualización constante de los datos y lograr una representatividad adecuada de los mismos son los dos criterios irrenunciables en cualquier sistema predictivo, pues, de lo contrario, se potenciaría la estigmatización de zonas y personas, creando situaciones de desigualdad y dibujando un panorama delictivo ajeno a la realidad.

Insurtech*, por I. Lledó Benito*

El término "*insurtech*" deriva del inglés *insurance* (seguro) y *technology* (tecnología), y se asocia a las empresas aseguradoras que aplican las tecnologías y herramientas más avanzadas -IA, *Big Data*, el Internet de las cosas- para optimizar y garantizar unas pólizas más rentables y más competitivas en costes para los consumidores, mediante la personalización de las coberturas que ofrecen en las pólizas.

Ahora bien, como ha afirmado Hidalgo Cerezo, con el vocablo «*insurtech*» nos referimos, por tanto, a una interacción entre las nuevas tecnologías y el sector asegurador, pero pueden incluirse multitud de acepciones y actividades: optimización de procesos, generación y administración de documentación automática, nuevas formas de contratación o tramitación de siniestros, incluidos *chatbots*, contratación en línea, OCR, reconocimiento de imágenes por IA o,

incluso, el aseguramiento de nuevos ciberriesgos ("*Insurtech*: ¿evolución o revolución de la actividad aseguradora?", en *Retos inminentes de derecho digital*, Edit. Dykinson 2024, pág. 240).

Las empresas especialistas en estos servicios ofrecen a las compañías de seguros una información real y con ahorro de costes en eficiencia y transparencia, a la hora de concretar reclamaciones, riesgos, daños, ya que el ahorro es significativo en personal (peritos, técnicos, analistas) con resultados digitales, sustituyendo esta intervención humana con resultados más económicos para el cliente consumidor.

La figura tradicional del corredor de seguros se ha ido sustituyendo por estas aplicaciones, que ofrecen mayor flexibilización de los contratos y unas coberturas de la póliza diseñadas *ad personam*. Este servicio se ha ido implementando en las compañías aseguradoras, ya que se viabiliza la posibilidad "ya real" de gestionar "digitalmente" las pólizas mediante un análisis de *Big Data*. Supone una verdadera revolución tecnológica en el mundo de las empresas de seguros. Y así se ha dicho, reproducimos literalmente, que estas empresas de *insurtech* desarrollan tecnología que permite realizar una valoración de riesgo de los usuarios que se ajustan a las características concretas de los ciudadanos (vide Life 5, *insurtech* especializada en seguros de vida, que ha revolucionado las pólizas convencionales).

Las empresas *insurtech* son un presente de revolución tecnológica en el ámbito de la contratación de seguros, donde la IA y el *Big Data* ha obligado a las compañías del sector a la aplicación de estas herramientas, que conllevan un ahorro de costes, una precisión en la cobertura de los seguros, en la gestión personalizada de los contratos y en la prima efectivamente por pagar. La digitalización de los servicios de contratación personalizada facilitará la agilización de la oferta y una mayor rentabilidad para el consumidor.

IV. Nuevos perfiles profesionales

Analista de procesos legales (*Legal Process Analyst*), *por J. I. Solar Cayón*

El analista de procesos legales es la persona responsable de diseñar y optimizar los flujos de trabajo en el despacho o en el departamento jurídico de una empresa, coordinando los equipos jurídicos y tecnológicos de la organización, e incluso, en su caso, los proveedores externos -puede tratarse de un *Legal Process Outsourcer*, de un proveedor alternativo de servicios legales o de un despacho, especialmente en el caso de departamentos jurídicos de empresas-, al objeto de mejorar la eficiencia de aquellos procesos y, en definitiva, la calidad del producto final.

Debido a la automatización de tareas propiciada por el desarrollo de la IA jurídica y a la consecuente desagregación de la actividad jurídica, esta figura se ha convertido en una pieza esencial en la organización de los principales despachos y de los departamentos jurídicos de grandes compañías. El analista de procesos legales analiza los flujos de actividad, descompone el trabajo jurídico en tareas susceptibles de ser realizadas de manera independiente e identifica en cada caso el tipo de agente más adecuado para su realización más eficiente, ya se trate de diversos tipos de profesionales -de la propia organización o proporcionados por terceros-, de diversos tipos de herramientas tecnológicas o de diversos tipos de compañías -firmas jurídicas, *Legal Process Outsourcers* u otro tipo de proveedores alternativos de servicios jurídicos-. Es, por tanto, el responsable de la arquitectura de los procesos de trabajo, diseñando estos de modo que la organización pueda optimizar los potenciales beneficios de la tecnología y de las nuevas dinámicas organizativas que esta genera.

Este profesional precisa una amplia experiencia y un conocimiento profundo de la naturaleza y los objetivos de las diversas áreas del trabajo jurídico, así como de los modos más eficientes de llevarlo a cabo. Asimismo, ha de poseer una visión global del modelo de negocio del bufete o de la actividad global del departamento jurídico, ocupando una posición clave y transversal en la organización: tiene que trabajar en estrecho contacto con la dirección para entender los objetivos de la organización, la estructura de sus procesos y los resultados deseados, pero también es fundamental el contacto con los equipos internos -legal y tecnológico- y con los proveedores externos que intervienen en los procesos de trabajo, así como con los clientes que finalmente reciben o utilizan los productos y servicios que constituyen el resultado de esos procesos.

Director de tecnología legal (*Chief Legal Technology Officer*), *por J. I. Solar Cayón*

La incorporación de la tecnología, y particularmente de la IA, en las firmas jurídicas y en los departamentos jurídicos de las empresas ha dado lugar a la aparición de nuevos perfiles profesionales de carácter híbrido, surgidos en el área de intersección entre el Derecho, la ingeniería y la gestión empresarial. En un contexto eminentemente tecnológico, donde los sistemas de IA realizan una parte cada vez mayor de las tareas jurídicas, la práctica profesional del Derecho, especialmente en los grandes despachos, ya no es únicamente un asunto de abogados, sino que requiere equipos interdisciplinares y líderes capaces de comprender todas las dimensiones -jurídica, tecnológica y empresarial- de la industria legal.

Tal vez el ejemplo más prominente de la emergencia de este nuevo tipo de perfiles nos lo proporciona la figura del Director de tecnología legal (*Chief Legal Technology Officer*), o la muy similar del **Director de innovación legal** (***Chief Legal Innovation Officer***). Se trata de un rol que apareció en los Estados Unidos a mediados de la década de 2010 como consecuencia del creciente interés de las firmas jurídicas por integrar en sus procesos operativos y de provisión de servicios legales las nuevas herramientas de IA jurídica. El Director de tecnología o de innovación legal es el responsable de diseñar y dirigir la estrategia de implementación de la tecnología legal en un despacho o en un departamento jurídico de una empresa, de optimizar a través de ella los procesos de provisión de los servicios legales y de promover una cultura innovadora en la organización. No se trata simplemente de un experto en tecnología legal, capaz de monitorizar el mercado y evaluar las distintas herramientas IA disponibles para su integración efectiva en la firma, sino de un profesional de alto perfil directivo que ha de poseer también un conocimiento profundo de la práctica jurídica y de la estrategia de negocio del despacho, al objeto de desarrollar una estrategia tecnológica y un esquema de gobernanza y uso de la IA alineadas con la cultura de la organización y ajustadas a sus necesidades específicas. Dada la creciente importancia del elemento tecnológico en las firmas jurídicas, esta figura juega un papel esencial en el diseño de la forma en la que los servicios legales son prestados a los clientes, que lógicamente ha de ser coherente con el modelo de negocio del despacho.

Director de IA aplicada (*Director of Applied AI, Head of Machine Learning Practice, Head of Legal Technology Adoption*)*, por J. I. Solar Cayón*

Actualmente, con el desarrollo de la IA generativa, algunos despachos anglosajones están creando nuevos roles próximos al Director de tecnología o de innovación legal para dirigir el proceso de implementación efectiva de aplicaciones basadas en modelos de lenguaje legal en la práctica diaria de la organización, tanto en tareas de gestión como en tareas estrictamente jurídicas. Se trata de figuras con denominaciones heterogéneas y funciones también variables en el organigrama de cada organización (a modo de ejemplo, *Director of Applied AI* en *Reed Smith*, *Head of Machine Learning Practice* en *Baker McKenzie*, *Head of Legal Technology Adoption* en *HFW*). Fundamentalmente, actuando bajo la supervisión del *Chief Technology Officer*, lideran el desarrollo de los modelos de lenguaje y el diseño de aplicaciones de IA generativa en el despacho, dirigiendo un equipo multidisciplinar de ingenieros de IA, científicos de datos, ingenieros de *prompts*, abogados...

Científico de datos jurídicos (*Legal Data Scientist*), *por J. I. Solar Cayón*

Hoy, gracias a la sofisticación de las técnicas de *big data* y aprendizaje automático, que han propiciado el desarrollo de complejos modelos predictivos a partir de la gestión, análisis e interpretación del inmenso caudal de datos registrados y disponibles, la explotación del valor de la información ha adquirido una nueva dimensión, convirtiéndose en un elemento estratégico de la actividad de los bufetes que incide directamente en la calidad de sus servicios jurídicos. En este contexto surge el rol del científico de datos jurídicos.

La función básica de esta figura consiste en transformar los ingentes volúmenes de datos disponibles en conocimiento útil, que ayude tanto a los responsables de los despachos y departamentos jurídicos de empresas como a los profesionales del Derecho a tomar las mejores decisiones. No se trata simplemente de un **analista de datos jurídicos** o ***Legal Data Analyst*** -esto es, de alguien que recolecta datos, los analiza y les da un sentido-, un rol que la industria legal ha venido incorporando ya desde hace tiempo, sino de un nuevo tipo de profesional que, a diferencia de aquel, necesita poseer conocimientos en estadística, programación e IA para desarrollar métodos analíticos novedosos y modelos de aprendizaje automático que permitan descubrir nuevas perspectivas y extraer información más valiosa de los datos. Su campo de juego es, por tanto, básicamente la minería de datos, una disciplina a caballo entre la estadística y las ciencias de la computación cuyo objetivo es extraer información de grandes volúmenes de datos y transformarla en una estructura comprensible para su uso posterior. Fundamentalmente, buscará identificar correlaciones, tendencias, patrones y enfoques en la información jurídica (entendiendo por tal tanto la contenida en los documentos jurídicos como la relativa a la conducta de los distintos agentes que intervienen en los procesos jurídicos) para generar modelos predictivos que faciliten la toma de decisiones.

Pero, para ello, ha de comprender también los problemas jurídicos, al objeto de saber formular las preguntas adecuadas, de seleccionar los indicadores relevantes o de aplicar los criterios de filtrado oportunos para desentrañar el sentido de la información. Se trata, por tanto, de un perfil profesional híbrido y muy específico, que precisa tener formación jurídica pero también conocimientos importantes de estadística y computación.

Ingeniero de conocimientos jurídicos (*Legal Knowledge Engineer*), *por J. I. Solar Cayón*

El término *Legal Knowledge Engineer* fue acuñado por Richard Susskind en su libro *The End of Lawyers?*, 2008, y ha sido ampliamente adoptado en la

industria legal con el desarrollo de las herramientas de IA jurídica. El ingeniero de conocimientos jurídicos juega un papel fundamental en el diseño de esas herramientas, trasladando el *expertise* jurídico a un formato con el que puedan trabajar los desarrolladores de *software*, de modo que aquel conocimiento práctico sea convertido en aplicaciones capaces de realizar determinadas tareas jurídicas. El proceso de comoditización del Derecho requiere de profesionales que organicen, analicen y representen grandes volúmenes de materiales jurídicos y que desarrollen procedimientos estandarizados al objeto de que el conocimiento experto pueda ser incorporado a los sistemas digitales. Como indica Susskind, este rol requiere un alto grado de conocimientos jurídicos y es mucho más exigente intelectualmente que el papel tradicional del abogado de asesoramiento sobre un tema específico, ya que requiere el desarrollo de modelos de conocimiento más genéricos, que puedan implementarse y aplicarse para solucionar problemas en un amplio rango de circunstancias. Pero, además, este profesional necesita entender los aspectos técnicos del proceso de diseño y desarrollo de las aplicaciones de IA para dialogar con el experto en tecnología.

Con la irrupción de la IA generativa, hoy su función se centra fundamentalmente en la elaboración y gestión de las bases de conocimiento que posibilitan el impulso de los flujos de trabajo mediante sistemas de IA. En este sentido, algunas de sus tareas fundamentales son:

a) desarrollar y gestionar activos de conocimiento, bibliotecas de contexto y materiales de referencia que sirvan como entradas para flujos de trabajo impulsados por IA;

b) crear y mantener *datasets* de evaluación, casos de prueba y estándares de calidad para herramientas impulsadas por IA;

c) desarrollar e implementar protocolos de garantía de calidad y marcos de evaluación para los resultados de la IA;

d) traducir los requerimientos de la práctica legal en repositorios de conocimiento estructurado, modelos de *prompts* y especificaciones de contexto que optimicen las entradas del sistema para conseguir resultados de alta calidad.

El ingeniero de conocimientos jurídicos constituye, pues, el interfaz entre el jurista y el tecnólogo. En ocasiones son expertos en tecnología que, tras desarrollar su trabajo durante años en el seno de una firma jurídica, llegan a familiarizarse con los procesos legales y a adquirir un conocimiento jurídico suficiente. Pero lo más frecuente es que sean abogados capacitados tecnológicamente -y frecuentemente con experiencia en gestión de proyectos jurídicos- que ven la

oportunidad de mejorar los procesos jurídicos y los servicios al cliente mediante la utilización inteligente de la tecnología. En cualquier caso, el ejercicio de su rol se halla caracterizado por la creatividad e innovación, la búsqueda de la eficiencia, la mejora de los procesos de trabajo jurídico y el compromiso con la prestación de un mejor servicio al cliente mediante el diseño y la utilización de sistemas tecnológicos.

Ingeniero de *prompts* jurídicos (*Legal Prompt Engineer*)*, por J. I. Solar Cayón*

Con el desarrollo de los modelos de lenguaje legal, algunas firmas jurídicas, departamentos jurídicos de empresas y compañías de tecnología legal han introducido el rol de ingeniero de *prompts* jurídicos. Su función es la de diseñar, testar y optimizar *prompts* para las diversas tareas (investigación jurídica, análisis de casos, redacción de documentos legales…), al objeto de que los resultados del modelo sean precisos, ajustados a los principios de cumplimiento normativo y éticos. Entre sus responsabilidades principales están la creación de bibliotecas de *prompts* especializadas, la implementación de cadenas de *prompts (prompt chaining)* para descomponer tareas legales complejas en pasos estructurados y controlables, y el entrenamiento de los abogados en las técnicas de *prompting*.

En cuanto al perfil de este profesional, es fundamentalmente jurídico, en tanto debe poseer competencias en tareas de investigación jurídica, redacción de documentos legales y litigación, así como conocer los principios éticos y deberes deontológicos de la abogacía. No obstante, debe tener también conocimientos en *ingeniería de prompts* que le capaciten para generar *prompts* estructurados y que aporten riqueza de contexto.

Responsable de gestión de proyectos legales (*Legal Project Manager*)*, por J. I. Solar Cayón*

La gestión de proyectos legales o *legal project managment* consiste básicamente en la aplicación de las técnicas de gestión de proyectos a la gestión de los procesos de prestación de servicios legales. Este enfoque surge principalmente en grandes firmas jurídicas de Estados Unidos, Reino Unido, Canadá y Australia al inicio de la década de 2010, como consecuencia de la crisis económica, para racionalizar los procesos de trabajo y reducir costes, especialmente ante la demanda de sus clientes de servicios jurídicos a precios cerrados. Pero, como consecuencia del desarrollo de la tecnología legal, este rol jurídico ha evolucionado y ha adquirido un perfil específico, combinando habilidades de gestión de proyectos, competencias tecnológicas y métodos de mejora de procesos para incrementar la eficiencia y la productividad de la organización. En la actualidad la gestión de proyectos legales es cada vez más utilizada tanto por los bufetes como por

las asesorías jurídicas de empresa como un elemento estratégico para mejorar la ejecución y coordinación de los diversos procesos que integran un determinado plan de trabajo, en el que a menudo están implicados diversos equipos de profesionales y departamentos de la organización, y para controlar los costes globales del proyecto. De ahí la creciente relevancia de la figura del responsable de la gestión de proyectos jurídicos o director de proyectos jurídicos (*Legal Project Manager*), responsable de diseñar las herramientas y los métodos para alcanzar tales objetivos. Se trata de un experto en planificación, organización y asignación de recursos, planificación presupuestaria y control, que, al mismo tiempo, ha de conocer bien el negocio y la cultura jurídica: tiene que entender lo que quiere el cliente, por lo que la comunicación con este resulta esencial, y ser un coadyuvante del abogado, liberándole de las tareas de gestión para permitirle centrarse en su trabajo estrictamente jurídico.

Además, el recurso cada vez más frecuente al *outsourcing* ha revalorizado aún más la importancia de este rol, en la medida en que la derivación de una parte cada vez mayor y más relevante de la actividad jurídica a proveedores externos a la organización exige una gestión de los tiempos, de los riesgos y de los costes de los procesos de trabajo mucho más rigurosa que si estos fueran ejecutados íntegramente *in-house*. Una vez que el analista de procesos legales ha terminado su diseño de los flujos de trabajo, descomponiendo la actividad jurídica en tareas que pueden ser realizadas de manera independiente por diversos equipos de trabajo internos y diversos tipos de proveedores externos, es preciso asegurar la correcta ejecución y coordinación de dichos procesos. Y esta es la responsabilidad del director de proyectos legales, que ha de gestionar la logística global de la cadena de suministro, asegurándose de que cada una de las tareas es asignada al proveedor adecuado, verificando que todas ellas son realizadas a tiempo y ajustándose al presupuesto disponible, controlando su calidad, supervisando su resultado y el ensamblaje de los distintos componentes, y, en definitiva, garantizando que el producto final es un servicio correctamente prestado.

Índice Analítico

A

Acceso algorítmico al Derecho extranjero: 200.

Actuaciones judiciales asistidas: 225.

Actuaciones proactivas: 225.

Actuaciones procesales automatizadas: 224.

Agente con herramientas: 90.

Agente de IA: 29.

Agentic AI: 29.

AI Alignment: 121.

Ajuste fino: 85.

Alfabetización digital: 190.

Algocracia: 129.

Algoritmo: 47.

Algoritmo evolutivo: 49.

Algoritmo genético: 51.

Alineamiento de la IA: 121.

Alineamiento engañoso: 123.

Alineamiento posterior al entrenamiento: 89.

Alternative Business Structure: 278.

Alucinación: 103.

Analista de datos jurídicos: 290.

Analista de procesos legales: 287.

Antropodatismo: 134.

Aprendizaje automático: 55.

Aprendizaje autosupervisado: 59.

Aprendizaje no supervisado: 58.

Aprendizaje por refuerzo: 60.

Aprendizaje profundo: 61.

Aprendizaje supervisado: 56.

Argument Mining: 64.

Armas autónomas: 138.

Arquitectura de transformadores: 83.

Artificial General Intelligence: 25.

Asimetría cognitiva algorítmica: 212.

Auditoría algorítmica: 152.

Autocodificador: 69.

Autoencoder: 69.

Automated due diligence: 262.

Autonomy: 117.

Autoridad garante del cumplimiento del Derecho: 174.

Autoridad notificante: 181.

B

Beneficence: 114.

Bias: 159.

Big Data: 31.

Black Box: 65.

Blockchain: 33.

Brechas digitales: 155.

C

Capa: 66.

Chief Legal Innovation Officer: 289.

Chief Legal Technology Officer: 288.

Ciberciudadanía: 232.

Ciberseguridad: 45.

Científico de datos jurídicos: 290.

Codificación computacional del Derecho: 268.

Codificación predictiva: 253.

Código fuente: 52.

Comoditización de los servicios legales: 274.

Compliance-by-Design: 146.

Computación biocibernética: 44.

Cómputo molecular: 45.

Confianza mutua algorítmica: 209.

Constitucionalismo digital: 235.

Context window: 97.

Contrato de IA transfrontrizo: 214.

Convolutional Neuronal Network: 63.

Cuantización: 100.

Cuidados e inteligencia artificial: 140.

Cumplimiento por diseño: 146.

D

Data-driven lawyering: 282.

Data mining: 32.

Datos: 70.

Datos de entrada: 75.

Datos de entrenamiento: 73.

Datos de prueba: 73.

Datos de salida: 75.

Datos de validación: 73.

Datos personales: 72.

Datos sintéticos: 76.

Deber de competencia tecnológica: 272.

Debido proceso tecnológico: 165.

Deceptive Alignment: 123.

Decisión administrativa automatizada: 218.

Decisión algorítmica: 148.

Deepfake: 184.

Deep Learning: 61.

Derecho a explicación: 193.

Derecho Cuántico: 240.

Derechos digitales: 222.
Descubrimiento electrónico: 252.
Desintermediación: 275.
Destilación de modelos: 99.
Diligencia debida automatizada: 262.
Director de IA aplicada: 289.
Director de innovación legal: 289.
Director de tecnología legal: 288.
Director of Applied AI: 289.
Discriminación algorítmica: 157.

E

E-discovery: 252.
Embeddings: 99.
Espacio controlado de pruebas: 186.
Espacio latente: 99.
Ethics-by-Design: 126.
Ética de la IA: 111.
Ética por diseño: 126.
Etiquetado: 194.
Evaluación de la conformidad: 179.
Explicability: 118.

F

Fine tuning: 85.
Foundation Model: 82.
Function calling: 91.

G

Garantía del cumplimiento del Derecho: 174.
Gemelo digital: 39.
Generación condicionada: 92.
Generative Adversarial Network: 64.
Gobernanza de la IA: 109.
Guardarraíles: 125.
Gubernamentalidad algorítmica: 129.

H

Head of Legal Technology Adoption: 289.
Head of Machine Learning Practice: 289.
Human-in-the-loop: 128.
Human-on-the-loop: 128.
Human-out-of-the-loop: 128.

I

Identidad digital transfronteriza: 199.
Identificación biométrica: 176.
Incorporated Legal Practices: 279.
Incrustaciones vectoriales: 99.
Inferencia: 92.
Informática jurídica: 240.

Informática jurídica decisional: 242.

Informática jurídica de gestión: 243.

Informática jurídica documental: 241.

Ingeniería de *prompts*: 101.

Ingeniero de conocimientos jurídicos: 290.

Ingeniero de *prompts* jurídicos: 292.

Inputs: 75.

Insurtech: 286.

Inteligencia artificial estrecha: 24.

Inteligencia artificial general (AGI): 25.

Inteligencia artificial jurídica: 248.

Inteligencia artificial simbólica: 27.

Inteligencia artificial y doctrina social de la Iglesia: 143.

Inteligencia Biológica Sintética: 42.

Internet y Derecho: 231.

Interpretación algorítmica transfronteriza: 206.

Inversión de la carga de la prueba algorítmica: 283.

J

Justice / *Fairness*: 115.

Justicia algorítmica: 120.

Justicia digital: 265.

L

Labelling: 194.

Large Language Model: 82.

Law as Code: 268.

LawTech: 246.

Legal Data Analyst: 290.

Legal Data Scientist: 290.

Legal Knowledge Engineer: 290.

Legal outsourcing: 277.

Legal Process Analyst: 287.

Legal process outsourcers: 277.

Legal Project Manager: 292.

Legal Prompt Engineer: 292.

Legal Tech: 244.

Lethal Autonomous Weapon Systems: 138.

Lex fori digital: 197.

Libertad informática: 145.

Litigación basada en datos: 282.

Llamada a funciones: 91.

M

Machine Learning: 55.

Máquina pensante: 23.

Marcado de agua: 194.

Medios sintéticos: 105.

Memoria conversacional: 95.

Mercado jurídico virtual: 280.

Metadato: 78.

Metaverso: 37.

Minería de argumentos: 64.

Minería de datos: 32.

Modelo de IA de uso general: 84.

Modelo de IA específico de dominio: 84.

Modelo de lenguaje de gran tamaño: 82.

Modelo de lenguaje legal: 263.

Modelo de lenguaje pequeño: 82.

Modelo fundacional: 82.

Modelo generativo: 80.

Modelo multimodal: 82.

Modelo preentrenado: 84.

Muestreo: 93.

Multimodalidad generativa: 92.

N

Narrow AI: 24.

Natural Language Processing: 64.

Neuroderechos: 237.

Neurotecnologías: 40.

Nodo: 67.

Non-maleficence: 112.

Nuevas tecnologías, Derecho y libertades: 234.

Nuevas tecnologías y doctrina social de la Iglesia: 141.

O

Offshoring: 278.

Online Legal Marketplace: 280.

Ontología: 29.

Organismo de evaluación de la conformidad: 180.

Organismo notificado: 182.

Orquestación: 100.

Outputs: 75.

P

Perfilado: 188.

Pericial algorítmica transfronteriza: 204.

Personalidad algorítmica: 150.

Pesos: 68.

Policía predictiva: 285.

Posthumanismo: 133.

Post-training Alignment: 89.

Predictive Coding: 253.

Preentrenamiento: 85.

Principio de autonomía: 117.

Principio de beneficencia: 114.

Principio de explicabilidad: 118.

Principio de justicia: 115.

Principio de no malevolencia: 112.

Privacidad algorítmica transfronteriza: 216.

Procedencia del contenido: 106.

Procesamiento del lenguaje natural: 64.

Prompt: 101.

Propiedad intelectual e IA: 164.

Proveedor: 171.

Proveedor alternativo de servicios legales: 275.

Provenance: 106.

R

Razonamiento algorítmico transfronterizo: 202.

Razonamiento en cadena como fenómeno técnico: 104.

Realidad virtual: 36.

Reconocimiento facial: 175.

Redes generativas antagónicas: 64.

Redes neuronales: 62.

Redes neuronales convolucionales: 63.

RegTech: 247.

Reserva de humanidad: 220.

Responsabilidad algorítmica: 151.

Responsable de gestión de proyectos legales: 292.

Responsable del despliegue: 173.

Retrieval-Augmented Generation: 87.

S

Salida estructurada: 105.

Sampling: 93.

Sandbox: 186.

Servidor: 80.

Sesgo: 159.

Sesgo de automatización: 160.

Sesgo de confirmación: 162.

Sesgo positivista: 136.

Sistema de categorización biométrica: 177.

Sistema de evaluación de riesgo delictivo: 257.

Sistema de IA: 167.

Sistema de IA de alto riesgo: 169.

Sistema de IA generativa: 80.

Sistema de identificación biométrica: 176.

Sistema de reconocimiento de emociones: 178.

Sistema multimodal: 82.

Sistemas de análisis predictivo: 255.

Sistemas de lectura automática y análisis de contratos: 260.

Sistemas de negociación y resolución automatizada de disputas: 259.

Sistemas expertos jurídicos: 251.

Small Language Model: 82.

Smart contract: 270.

Superinteligencia Artificial (ASI): 25.

Supervisión humana: 183.

Synthetic Biological Intelligence: 42.

Synthetic data: 76.

Synthetic media: 105.

T

Technological due process: 165.
Technology-Assisted Review: 253.
Tecno-regulación: 238.
Temperatura: 94.
Token: 96.
Token de contexto: 98.
Tokenización: 96.
Tool Use: 89.
Top-K: 94.
Top-P: 94.
Traducción algorítmica: 207.
Transhumanismo: 131.
Transparencia algorítmica: 153.
Tribunal en línea: 267.
Tutela judicial algorítmica transfronteriza: 196.

U

Ultrasuplantación: 184.
Uso de herramientas: 89.

V

Vectores de posición: 69.
Ventana de contexto: 97.
Verificación biométrica: 177.

W

Watermarking: 194.
Weak AI: 24.
Web scraping: 78.
White Box: 66.

La irrupción de la inteligencia artificial en el mundo del Derecho ha traído consigo un vocabulario nuevo que el jurista ya no puede permitirse ignorar. Este *Glosario básico de inteligencia artificial jurídica*, dirigido por Fernando H. Llano Alonso y José Ignacio Solar Cayón, reúne a cerca de medio centenar de especialistas de universidades e instituciones españolas, europeas y americanas para ofrecer una guía rigurosa y accesible de los conceptos esenciales de esta nueva realidad. La obra se articula en tres partes: los fundamentos técnicos de la IA —del algoritmo al modelo de lenguaje de gran tamaño—, su ecosistema ético y jurídico —con especial atención al Reglamento europeo de IA y a la normativa española— y las aplicaciones de la IA al dominio jurídico, de la *Legal Tech* a la codificación computacional del Derecho. Cada entrada conjuga la explicación técnica con el análisis jurídico, en la convicción de que solo puede regularse con rigor aquello que se comprende. Un instrumento indispensable para estudiantes, académicos y profesionales del Derecho.

Esta publicación es parte del Proyecto de I+D+i "Inteligencia artificial jurídica y Estado de Derecho" [PID2022–139773OB-I00], financiado por MICIU/AEI/10.13039/501100011033 y por FEDER, UE, y del Proyecto de I+D+i DENORIA [PID2023-146621OB-C21].

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Cofinanciado por la Unión Europea

AGENCIA ESTATAL DE INVESTIGACIÓN

ISBN 979-13-88025-85-3