Mirando hacia el futuro. Cambios sociohistóricos vinculados a la virtualización

Edición a cargo de
Olivia Velarde Hermida y Manuel Martín Serrano

CIS
Centro de Investigaciones Sociológicas

Mirando hacia el futuro. Cambios sociohistóricos vinculados a la virtualización / edición a cargo de Olivia Velarde Hermida y Manuel Martín Serrano. – Madrid: Centro de Investigaciones Sociológicas, 2022
(Academia; 51)

1.Activismo Político 2. Movimientos de Protesta
316.42

Las normas editoriales y las instrucciones para los autores pueden consultarse en:
www.cis.es/publicaciones/AC/

Colección, ACADEMIA 51

Catálogo de Publicaciones de la Administración General del Estado
http://publicacionesoficiales.boe.es

Primera edición, septiembre 2022

Impreso y hecho en España
Printed and made in Spain

NIPO (papel): 092-22-009-8 — NIPO (electrónico): 092-22-010-0
ISBN (papel): 978-84-7476-883-1 — ISBN (electrónico): 978-84-7476-884-8
Depósito legal: M-21992-2022

Fotocomposición e impresión: Editorial MIC

Para la impresión de este libro se ha utilizado papel con certificación FSC, ECF y PEFC.
Esta publicación cumple los criterios medioambientales de contratación pública.

5. La gobernanza de los sistemas artificiales inteligentes

Pablo Noriega[1] y Pompeu Casanovas[2]

Did I request thee, Maker, from my clay
To mould Me man? Did I solicit thee
From darkness to promote me?
John Milton, *Paradise Lost* (X. 743-45)[3]

5.1. Introducción

En mayo de 2014 se reunió en la Universidad de Cambridge un grupo de científicos destacados, entre los que estaba el premio Nobel de Física Stephen Hawking, para debatir sobre el futuro de la humanidad. Al terminar el evento, Hawking hizo unas declaraciones a la prensa que tuvieron repercusión mundial: «La Inteligencia Artificial es quizá el mayor invento de la humanidad –dijo– y puede ser el último» (CNBC, 2014).

A este sombrío pronóstico se sumó días después un debate temático dentro del Foro Mundial de Davos, cuyo tema central fue el futuro del trabajo y cómo la robótica y la inteligencia artificial (IA) pueden afectarlo (WEF, 2014). El optimismo catastrófico de esa reunión se resume en una frase, posiblemente apócrifa: «en el futuro habrá dos únicos trabajos en los que los humanos serán imprescindibles: programador de IA e instructor de *zoomba*».

En esta línea de pronósticos, Kurzweil, uno de los vicepresidentes de Google e inventor de uno de los primeros lectores automáticos para invidentes, afirmó convencido que «en el año de la singularidad», los ordenadores serían tan poderosos como para recibir y utilizar todo el contenido de la mente de un humano y que, reinterpretando la historia de Frankenstein, este se podría transferir a un ente silicónico que para todo efecto práctico lo haría inmortal.[4]

---

[1] Instituto de Investigación en Inteligencia Artificial. Consejo Superior de Investigaciones Científicas. Campus UAB, Bellaterra (pablo@iiia.csic.es).

[2] La Trobe University, Melbourne, Australia. Instituto de Derecho y Tecnología, UAB, España (P.CasanovasRomeu@latrobe.edu.au, pompeu.casanovas@uab.es).

[3] De la primera edición de M. Shelly, *Frankenstein; or, The Modern Prometheus* (Lackington, Hughes, Harding, Mavor y Jones, 1818).

[4] El año de la singularidad se refiere al punto de la curva exponencial en que la «ley de la aceleración de rendimientos» tecnológicos hará que los ordenadores puedan controlar la evolución (Kurzweil, 2005). El término y la idea provienen de J. Von Neumann (Ulam, 1958, p. 5).

Si bien las tres anécdotas anteriores tienen un punto de dramatismo que apela a la fantasía popular, el éxito evidente de las aplicaciones de las técnicas de aprendizaje automático y ciencia de datos –en temas como el juego de GO, el reconocimiento de rostros y las campañas políticas– no solo han llamado la atención del gran público, sino que han propiciado una inversión sustancial en el desarrollo de tecnologías y productos asociados con la IA. Esta actividad ha tenido dos efectos notables. Por una parte, se ha conformado una visión de oportunidad urgente a la que tanto los gobiernos como las empresas y, en cierta medida, también las universidades, han respondido con la publicación de manifiestos estratégicos, la formulación de políticas y el patrocinio a veces caótico de una gran variedad de iniciativas y proyectos. Por otra parte, los medios de comunicación y las redes sociales han dado voz a visiones apocalípticas y a fantasías distópicas que suelen tener poco fundamento.

En realidad, esta situación refleja una paradoja que ya se ha dado en más de una ocasión y que en estos años se vive intensamente en torno a la IA. Collingridge la formuló de la siguiente forma: «Cuando aparecen tecnologías suficientemente disruptivas es muy difícil anticipar sus efectos; pero una vez que estas se entienden cabalmente, puede ser demasiado tarde para contrarrestar los problemas que generan» (Collingridge, 1980, p. 11).

La paradoja de Collingridge resuena en la advertencia de Hawking que citamos anteriormente e invita a analizar en qué medida la IA es disruptiva (Christensen, 1996) –o, si se prefiere, «perturbadora»– y explorar cómo anticipar y atender sus consecuencias indeseables. Y eso es lo que nos proponemos hacer en este ensayo. Para ello, primeramente, delimitaremos el campo de atención; luego discutiremos los aspectos que consideramos más disruptivos de la IA; y, finalmente, apuntaremos algunas líneas de intervención, especialmente en ética y en lo que denominaremos «gobernanza jurídica» (*legal governance*) para la IA.

5.2. La Inteligencia Artificial como tecnología disruptiva

En estos últimos años, tanto los gobiernos como las empresas y la prensa en general han prestado gran atención a la IA. Ese interés está motivado por la percepción generalizada del rol que la IA desempeña tanto en numerosas aplicaciones que dependen directamente de algún artefacto de IA (como traductores automáticos, asistentes personales como Siri o Alexa, sistemas de reconocimiento de rostros, aplicaciones de *micromarketing,* robótica…), como en aquellas plataformas en línea cuya existencia sería imposible sin IA (como Google, Amazon, Facebook o Uber).

Conviene, sin embargo, aclarar que esa exitosa imbricación no es accidental, es el resultado de la confluencia reciente de dos fuerzas directrices. La primera es consecuencia de la madurez de la disciplina al cabo de sesenta años de desarrollo, madurez que se manifiesta: i) por la disponibilidad de artefactos de IA de propósito general (motores de razonamiento automático, reconocimiento de patrones, aprendizaje de máquina, comprensión de lenguaje natu-

ral, etc.); ii) por la asimilación de aquellos en las tecnologías de la información convencionales; y iii) por la consolidación de una masa crítica de especialistas, profesionales y empresas, que aprovechan tales artefactos y continúan desarrollando ciencia, tecnología y prácticas en torno a la IA. La segunda fuerza directriz es la existencia de un sustrato fértil para esa disciplina madura. Tal sustrato es el resultado de la combinación de una cada vez más poderosa infraestructura de procesamiento de datos (capacidad de proceso, conectividad, almacenamiento) que, finalmente, permite explotar cabalmente numerosas tecnologías de IA y de la adopción masiva de Internet, que da vida a interacciones en línea de toda índole y con un fuerte e inevitable contenido digital.

Más allá de ese reconocimiento generalizado, hay un amplio consenso sobre el valor estratégico de la IA.[5] Los documentos internacionales estratégicos identifican como sus efectos más significativos: i) los impactos económicos de la IA (alteración del mercado laboral, dimensiones del negocio de IA, consecuencias socio-políticas de la las grandes empresas basadas en IA) y los impactos sociales (la virtualización de las relaciones y prácticas sociales); ii) la inadecuación de instrumentos jurídicos convencionales; iii) la aparición de nuevas prácticas comerciales, de publicidad y propaganda; y iv) la utilización de los rastros de la actividad personal en línea. En ese marco, realzan los beneficios potenciales asociados a la IA, así como algunas estrategias para alcanzarlos. Sin embargo, reflejan también la paradoja de Collingridge al manifestar claramente una preocupación por algunos efectos económicos y sociales negativos, así como las implicaciones legales y éticas que conlleva el uso de la IA.[6] Si el éxito de la IA se puede explicar por su madurez y un sustrato fértil de aplicación y los efectos disruptivos son innegables, ¿en dónde radica la inquietud y cómo contender con ella? La respuesta obvia es que la inquietud (y su consecuente atención) tiene dos fuentes: el factor humano y los artefactos que la propia disciplina produce. Pero esta es una descripción que, aplicable a toda tecnología disruptiva, tiene que ajustarse a las especificidades de IA.

¿Qué distingue a la IA de otras tecnologías disruptivas? ¿En dónde radican sus riesgos y promesas específicos? El rasgo distintivo de la IA es que crea teoría y artefactos que corresponden a procesos de inteligencia y, por lo tanto, crea artefactos que exhiben conductas que se pueden calificar de «racionales» (Simon, 1996, 2007). La inquietud profunda radica en que la IA se ocupa de facultades característicamente humanas asociadas a la inteligencia y que sus artefactos pueden trivializar, malinterpretar, competir y sustituir la racionali-

---

[5] Véase el documento del High-Level Expert Group on Artificial Intelligence (AI HLEG): https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai, titulado *Policy and Investment Recommendations for Trustworthy AI*, presentado durante la primera European AI Alliance Assembly en junio de 2019 (véase: https://wayback.archive-it.org/12090/20210104133629/https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence). Este documento sigue al *Informe sobre Trustworthy Artificial Intelligence*, presentado en abril (AI HLEG, 2019a y 2019b).

[6] Véase también, en este mismo libro, el capítulo anterior sobre la doble implosión del mercado jurídico y lo que está sucediendo con la irrupción de las empresas de Lawtech y Fintech.

dad natural. Postulamos que –para bien y para mal– el componente clave de la disruptividad de la IA es la autonomía en sistemas artificiales inteligentes. Por ello, los efectos positivos y negativos previsibles en la paradoja de Collingridge que hay que anticipar radican justamente en esa autonomía adscribible a los sistemas artificiales y, consecuentemente, en su gobernanza.

5.3. La «autonomía» en sistemas de IA

En términos muy laxos, la autonomía es cierta capacidad de decidir independientemente de terceros. En teoría económica, derecho o filosofía, un «principal» (una persona física o moral) delega en otra persona la capacidad de decidir y llevar a cabo acciones bajo ciertas condiciones. En inteligencia artificial, hay una delegación análoga en un «agente autónomo» pero que es una entidad computacional. La noción más general de sistema inteligente autónomo abarca sistemas informáticos que tienen la propiedad de autonomía, pero cuya inteligencia emerge de la combinación de sus propios componentes. La diferencia importante respecto a la autonomía convencional es que para el agente o el sistema autónomo artificiales siempre hay un diseño y una ingeniería (un artefacto) que implementa de una u otra forma esa autonomía delegada.

En este contexto, proponemos distinguir cinco niveles en los que la delegación de autonomía en los sistemas inteligentes artificiales es progresivamente más grande y, consecuentemente, su gobernanza más compleja. Ante la imposibilidad de un tratamiento detallado de este análisis, nos limitaremos a hacer una descripción de las características de cada nivel, ilustrarla con ejemplos concretos y apuntar los aspectos de los sistemas autónomos de ese nivel que pueden ser inquietantes.[7]

5.3.1. Nivel 1

El sistema autónomo recibe una delegación instrumental en la que, una vez definido un proceso, el sistema de IA toma de manera autónoma unas decisiones que se aplican en una parte bien delimitada del proceso y dentro de un universo predefinido de situaciones. La complejidad y la responsabilidad delegada pueden ser muy diversas. Hay algunos bastante inocuos, como por ejemplo los programas que interpretan las instrucciones por voz en los sistemas de atención telefónica al cliente, o los *chatbots* de atención de usuarios que establecen un diálogo que «ayuda» a precisar las reclamaciones. También caen en este nivel los robots (como Roomba) que hacen tareas elementales y

[7] Los elementos más relevantes de los sistemas mencionados en esta sección (Compas, Amazon Mechanical Turk, Captcha, Ushahidi, Duolingo, AlphaZero, Watson [inteligencia artificial], Cyc, Open_Mind_Common_Sense, DBpedia, GMT-3 y Ciudad Inteligente) están bien descritos en los artículos correspondientes de la *Wikipedia*. La versión en inglés puede ser más completa técnicamente que la de castellano.

usan una IA muy básica para el logro de sus tareas, por ejemplo, evitar obstáculos y construir un mapa del espacio (Angle y Brooks, 1990).

Otros sistemas autónomos de este nivel pueden tener modelos de comportamiento más sofisticados, como los traductores de texto automáticos o los asistentes para el diagnóstico automático por imágenes, pero su pericia se limita a una tarea bien delimitada. En estos últimos ejemplos, el modelo de decisión del sistema puede ser muy complejo y frecuentemente ininteligible, pero como el alcance de su autonomía está confinado a una tarea concreta, su competencia debería ser validable objetivamente. Por ello, la autonomía de estos sistemas debe ser, en principio, gobernable como otros tipos de agencia no artificial.

Sin embargo, hay sistemas de IA que aun diseñándose para realizar una tarea bien definida y realizarla competentemente, pueden conllevar problemas de delegación que tienen causas bien identificables, pero no siempre atendidas correctamente. Esencialmente, la problemática radica en delegar en el sistema un rango de decisión que excede el ámbito que se instrumenta en el sistema. Una manifestación de esta problemática son los sistemas complejos de gestión cuyas decisiones automáticas responden a una interpretación incorrecta de los procesos que gestionan. Por ejemplo, el sistema de gestión de surtido de pedidos en los centros de reparto de Amazon que se adecúan correctamente a las características físicas y operativas de los robots que mueven la mercancía, pero no así a las necesidades fisiológicas de los trabajadores humanos que son también parte del proceso (Sainato, 2020).

Otro problema recurrente sucede cuando el usuario le atribuye al sistema una capacidad de decisión sobre situaciones no previstas o inadecuadamente consideradas en el universo de su competencia. Por ejemplo, ignorando los sesgos en la fuente de datos sobre el que se construye el algoritmo que determina la probabilidad de reincidencia delictiva, como en el caso del sistema Compas.[8]

Otro problema frecuente radica en un exceso de autonomía cuando el sistema autónomo se utiliza para tomar decisiones que rebasan el universo de situaciones que se instrumentan en el sistema. Esta excesiva delegación de autonomía se suele agravar porque el sistema no tiene la competencia suficiente para fundamentar las decisiones. El caso citado para determinar la probabilidad de reincidencia (Compas) no solo tiene el problema de sesgo mencionado, sino que hay jueces cuya decisión de conceder la libertad bajo palabra se determina por la incorrecta probabilidad de reincidencia, sin además tomar en cuenta las condiciones atenuantes que no están previstas en ese algoritmo (Washington,

[8] *Correctional Offender Management Profiling for Alternative Sanctions* (COMPAS) es un programa ampliamente utilizado por los tribunales en Estados Unidos (New York, Wisconsin, California o Florida, entre otras jurisdicciones) para calcular el grado de reincidencia criminal. Se calcula que desde 1998 ha realizado más de un millón de calificaciones. Ha sido criticado desde el principio por incorporar un sesgo racial a sus resultados (Dressel y Farid, 2018).

2018). Otro ejemplo es la utilización de sistemas de reconocimiento facial para determinar la identidad de las personas –que suelen proporcionar un número excesivo de falsos reconocimientos positivos– y utilizar esta inadecuada identificación automática para tomar acciones que perjudiquen a los sujetos identificados. Este es el caso de la detención de sospechosos o la calificación de la reputación automática de ciudadanos en el conocido «sistema de crédito social» en la República China (Botsman, 2017; Zuboff, 2019).

5.3.2. Nivel 2

Los sistemas autónomos tienen una autonomía teleológica –o de misión– en la que el sistema articula múltiples tareas o acciones instrumentales u operativas interconectadas conducentes a un objetivo predeterminado. Un buen ejemplo de la complejidad que puede darse en este nivel es el de los sistemas que optimizan el uso de aviones en las aerolíneas comerciales: incluyen la selección de rutas, el plan de mantenimiento, determinación de puertas de embarque, asignación de asientos e inclusive las estrategias de descuento para la venta óptima de pasajes.[9]

En este tipo de sistemas autónomos se dan dos problemas adicionales de agencia, que muchas veces se mezclan de manera inextricable. El primero es el problema de atribución de responsabilidades; el segundo la opacidad de las decisiones basadas en «cajas negras». En estos sistemas intervienen delegaciones de autonomía y de responsabilidades de distintos actores y en distintos subsistemas. Por ello es difícil calcular la contribución individual a un daño de cada uno de los subsistemas y más aun la culpabilidad asociada a cada uno de los subsistemas. A esta dificultad suele añadirse el hecho de que las decisiones que toman cada uno de los subsistemas –y más aun la agregación de estos– pueden en sí mismas ser difíciles de explicar o de justificar, ya sea porque utilizan algoritmos o técnicas de decisión que por sí mismas son opacas –como el aprendizaje profundo– o porque se basan en componentes ajenos que no se pueden inspeccionar, ya sea por razones tecnológicas o por reservas de secreto industrial.[10]

---

[9] Otro ejemplo que sirve además para ilustrar las consecuencias morales que conlleva la delegación de autonomía teleológica es el de las armas autónomas (que deciden sobre el plan y la ejecución misma de una misión, como bombardear un objetivo militar). En este ejemplo, las implicaciones morales y la asignación de responsabilidades son ciertamente problemáticos, inclusive en el caso «restringido» en el que la decisión terminal de «apretar el gatillo» se reserva a un operador humano. El Principio n.º 18 de la Conferencia de ASILOMAR (2017) dice expresamente que «debe evitarse la carrera de armamentos respecto a armas autónomas letales» (Future of Life Institute, s.a.). La necesidad de control humano está presente en los veintitrés principios.

[10] Por ello, no se pueden explicar adecuadamente las decisiones que toma el sistema y por lo tanto es complicado justificar o impugnar las decisiones (por ejemplo, los sistemas que determinan el precio de descuento que se le ofrece a un cliente). Para contender con este problema, la UE ha propuesto adoptar el principio de «explicabilidad» para las aplicaciones de IA. Los artículos 13-15 y 22 del nuevo Reglamento de Protección de Datos (*General Data Protection Regulation*, GDPR) exigen que los procesadores de datos personales provean «la razón que está detrás o los criterios adoptados para alcanzar la decisión» cuando una decisión automática ha tenido lugar. Bibal *et al*. (2020) han distinguido cuatro niveles de «explicabilidad» en la legislación europea.

Por otra parte, está el difícil problema de anticipar y gestionar las consecuencias no deseadas de la manera en que el sistema autónomo alcanza los objetivos para los que fue diseñado.[11]

5.3.3. Nivel 3

En este nivel de autonomía conviene incluir sistemas que desbordan los confines que definen la competencia y la responsabilidad de los niveles 1 y 2. Aunque la frontera entre este y los dos niveles anteriores es difusa, el que un sistema sea de nivel 3 propiamente puede deberse a tres motivos:

– El primero es por el uso de una inteligencia mixta (artificial y natural), que corresponde a diversas formas de *crowd-processing* (Poblet *et al*., 2012). Los sistemas inteligentes de este tipo están basados en una arquitectura relativamente estándar (como el Amazon Mechanical Turk) que permite la descomposición de un problema en subproblemas, la cual se complementa con la correspondiente composición automática de las soluciones parciales, las cuales, sin embargo, se obtienen mediante procesos que requieren la intervención de inteligencia humana (Con Ahn *et al*., 2003). Esta hibridación es relevante, primeramente, porque al incorporar agentes humanos en la solución de subproblemas, el sistema aprovecha la inteligencia humana general para algunas tareas o misiones y la puramente artificial para otras. Un ejemplo es el caso del sistema Captcha para discriminar robots de humanos (Amazon se refiere a esto como «inteligencia artificial artificial»); otro es Ushahidi, para los mapas de catástrofes humanitarias (Poblet *et al*., 2018). La hibridación es relevante en segundo lugar porque la utilización de un sistema de coordinación en línea permite incorporar un número inmenso de personas en una misma actividad colectiva. Lo importante es que las decisiones requieren una inteligencia general y que la coordinación artificialmente inteligente permite incorporar una cantidad de decisiones que sería inmanejable sin ella; como Duolingo, que suscribe literalmente a millones de humanos para la enseñanza de idiomas. En esta subclase del Nivel 3, los problemas de asignación de responsabilidad y adecuación del diseño que atañen a los niveles 1 y 2 subsisten, pero además se añaden consideraciones asociadas a la representatividad, la privacidad, la equidad y la seguridad de las personas que intervienen en su funcionamiento y utilización.

– El segundo motivo para distinguir este nivel ocurre cuando una arquitectura muy sofisticada de sistemas inteligentes autónomos se reutiliza en nuevos dominios sin que requiera adecuaciones sustanciales. Aquí no se incluirían las clásicas máquinas de inferencia o los

[11] Se han propuesto dos ejemplos ideales para ilustrar esta problemática. Uno es el sistema autónomo programado para optimizar la producción de clips, que termina destruyendo al planeta para tener suficiente materia prima (Bostrom, 2014).

sistemas basados en conocimiento que las utilizan como su soporte de decisión, cambiando las heurísticas o hechos específicos de un dominio. Tampoco se incluirían las múltiples aplicaciones «estándares» de las redes neuronales convolutivas (aplicadas a diagnóstico, reconocimiento de estilos, etc.) ni los sistemas de traducción de lenguaje natural basados en aprendizaje bayesiano; estos corresponderían al nivel 1 o 2, ya que el artefacto básico es un componente autocontenido y que es parametrizable, o simplemente se entrena nuevamente para que sea utilizable en otro dominio. Sin embargo, sí conviene reconocer un nivel mayor de autonomía en programas como AlphaZero y su sucesor, MuZero, que solo llegan a adquirir su destreza para realizar tareas específicas de gran complejidad gracias a la concurrencia de capacidad masiva de cómputo y una muy sofisticada combinación de técnicas diversas de IA como solución cooperativa de problemas, aprendizaje artificial, percepción de patrones y razonamiento automático. Lo importante es que esa misma arquitectura se puede reentrenar por sí misma para otras tareas igualmente complejas (Silver *et al*., 2018).[12]

– El tercer elemento diferencial de sistemas en este nivel es cuando el dominio de competencia del sistema es abierto, lo que conlleva la solución genérica de problemas y por tanto un ámbito de delegación difícil de acotar a priori. Es el caso de sistemas como Cyc, un sistema para fundamentar el razonamiento basado en el sentido común y con ello dotar de generalidad a los sistemas expertos convencionales (Lenat *et al*., 1986). Cyc contiene una ontología y una base de conocimientos de dimensiones colosales que pretenden describir «cómo funciona el mundo» y que son resultado de esfuerzos cooperativos y continuados de la comunidad de IA y de una muchedumbre de voluntarios. Otros ejemplos análogos son Open Mind Common Sense (basado en frases en lenguaje natural, a diferencia del conocimiento representado formalmente de Cyc) y DBpedia (que extrae información de las entradas de *Wikipedia*). El impacto más significativo de este tipo de sistemas inteligentes es cuando se incorporan en sistemas autónomos que funcionan sin un dominio de conocimientos acotado. El caso más emblemático de esta subclase es el programa Watson, que fue diseñado para «comprender» y resolver acertijos en lenguaje natural que abarcan temas de cultura general. El sistema incorpora ese tipo de sistemas y una gran cantidad de datos no estructurados, además de diversos sistemas inteligentes para tareas como la comprensión y generación de lenguaje natural, algoritmos para comportamiento estratégico, múltiples formas de aprendizaje automáti-

[12] Los artículos sobre Amazon Mechanical Turk, Captcha, Ushahidi, Duolingo y AlphaZero en la *Wikipedia* hacen una buena descripción de estos sistemas. La versión en inglés suele ser técnicamente más precisa que la española.

co y una compleja arquitectura de solución cooperativa de problemas (Ferrucci, 2012).[13]

5.3.4. Nivel 4

Incluye sistemas que llevan a cabo un repertorio de acciones socialmente complejas dentro del mundo real de manera autónoma y que, desde el punto de vista conceptual, puede aducirse que en un cierto nivel de abstracción exhiben agencia moral (Floridi, 2013). Aquí caerían sistemas autónomos cuyas primeras encarnaciones ya existen como prototipos suficientemente convincentes o inclusive en versiones primitivas que demuestran la factibilidad de su construcción futura. Es el caso de, por ejemplo, los vehículos autónomos, los robots para asistencia (pacientes, personas mayores, discapacitados) o los programas como GMT-3 (Brown *et al*., 2020) que, más allá de la capacidad de Watson para organizar conocimiento común, son capaces de integrar distintas formas de percepción y acción alrededor de un comportamiento cuya racionalidad incluye una capacidad de aprender y realizar competentemente tareas que no estaban previstas en el diseño original.[14]

5.3.5. Nivel 5

Finalmente, se ha postulado la posibilidad de desarrollar sistemas artificiales con «inteligencia general», es decir, capaces de desempeñar cualquier tarea con el nivel de profundidad y generalidad que un humano pudiera llegar a realizar (López de Mántaras y Meseguer, 2017, p. 148; Woolbridge, 2020). Se entiende que esta caracterización supone una capacidad de aprendizaje, descubrimiento, creatividad y un grado de conciencia análogo al humano, lo que incluiría introspección, voluntad y juicio moral; ello constituye una realización artificial muy semejante a la autonomía y a la responsabilidad moral de los seres humanos. Tiene sentido dudar si acaso esto sea factible y, más aun, si fuere deseable.

---

[13] Watson se programó originalmente para competir en el programa televisivo *Jeopardy*, en el que tres concursantes humanos deben proporcionar respuestas a unas claves que son leídas por el conductor del programa. Las respuestas se deben dar correctamente y tan rápidamente como sea posible: el primer concursante que oprime un pulsador una vez que la clave ha sido leída en su totalidad puede responder y si su respuesta es correcta gana puntos; si no, los pierde y el primero de los otros concursantes que oprima el botón tiene la opción de responder. Las claves se presentan de forma ambigua y la respuesta debe expresarse en forma de una pregunta que sería respondida inequívocamente por la clave. Por ejemplo, si el programa fuese en castellano, un ejemplo podría ser que el conductor televisivo dijese: «Categoría "alimentos". Por trescientos euros. La clave es: "Elemental, querido Watson", cuya respuesta correcta sería: "*¿Cuál es el queso favorito de Sherlock Holmes*?"». En junio de 2011 el programa Watson compitió en condiciones exactamente iguales a las de los concursos diarios contra los dos concursantes más exitosos en la historia del programa y ganó (Markoff, 2011).

[14] Para lograr ese nivel de competencia racional, los creadores de GMT-3 aducen que el sistema goza de cierto nivel de autoconciencia y razonamiento moral; inclusive que puede mentir (conscientemente) para lograr sus objetivos (Metz, 2020; puede verse también una «entrevista» al sistema en Elliott (2020).

5.3.6. Sistemas multiagente

Aunque originalmente la IA se ocupó de procesos inteligentes de los individuos, en los últimos treinta años también se ha ocupado de la interacción inteligente, es decir, de los procesos racionales sociocognitivos y de la coordinación social (Noriega *et al*., 2016). El paradigma de esa inteligencia artificial social es el de agente, como apuntábamos al inicio de esta sección, la cual toma una forma particularmente significativa cuando la actividad colectiva de diversos agentes se da dentro de lo que se denomina un sistema multiagente (Woolbridge, 2020). Para realizar esta actividad colectiva se añade la noción clave de coordinación social, y con ella cobra relevancia la necesidad de gobernanza de las interacciones en lo que respecta al contexto compartido por los agentes (Andrigetto *et al*., 2013; Aldewereld *et al*., 2016).

Una consecuencia trascendental del paradigma de agente y de los sistemas inteligentes autónomos es la creación de sistemas multiagente *híbridos,* que dan pie a espacios de interacción social en línea, en los que participan entidades autónomas naturales y artificiales dentro de una realidad aumentada digitalmente. El ideal de las «ciudades inteligentes» es un buen ejemplo de estos espacios híbridos de interacción (Kominos, 2009). La trascendencia de estos espacios híbridos es que, además de los niveles de autonomía que puedan tener los sistemas autónomos artificiales incluidos, el nivel de autonomía de los agentes humanos que participan en tales interacciones corresponde, evidentemente, al Nivel 5.

5.4. Cómo contender con la disruptividad de la IA: la ética y los derechos

Como sucede con otras tecnologías que resultan disruptivas, una buena parte de la preocupación con la IA radica en el factor humano. Es decir, la necesidad de anticipar y contender con las consecuencias indeseables de la falta de responsabilidad, torpeza o dolo de individuos, organizaciones y gobiernos que desarrollan, poseen y usan las tecnologías de IA.

Dado que ciertamente esta preocupación atañe también a otras tecnologías para las cuales este problema ya se ha abordado previamente, una manera de atender a la inquietud es adecuar provechosamente las salvaguardas que funcionan para tales tecnologías al caso de la IA. Sin embargo, la adecuación de esas salvaguardas genéricas no siempre basta. Desde el punto de vista de una deontología del diseño y programación de sistemas IA, es necesario adaptarlas y desarrollar nuevas salvaguardas para contender con riesgos que provienen del objetivo propio de la IA. Es decir, de la implementación de comportamiento racional en sistemas artificiales y, más específicamente, cuando esos sistemas artificialmente inteligentes gozan de una cierta autonomía.

Postulamos que para contender con este riesgo es necesario establecer la gobernanza de tales sistemas. Creemos que sería provechoso analizarla para cada uno de los niveles de autonomía, pero llevar a cabo este tratamiento más

detallado excede los límites del presente capítulo. Vamos a limitarnos, pues, a una presentación general.

Para contender con los aspectos problemáticos del desarrollo o diseño de sistemas IA –sin entrar aquí tampoco en la relevante cuestión de su uso concreto– es posible adoptar tres estrategias: i) acciones genéricas dirigidas a diseñadores, usuarios y público (por ejemplo manifiestos, organizaciones, priorización de proyectos, políticas públicas de I+D+i, o formación curricular de ingenieros); ii) guias de diseño normado por valores, para hacer que los valores humanos sean parte de las consideraciones de diseño; y iii) finalmente, la estrategia más radical propuesta, que consiste en usar a la propia inteligencia artificial para resolver el problema de la moralidad en los sistemas artificiales: concretamente, diseñar los sistemas autónomos para que tomen decisiones y se conduzcan, en conformidad con valores humanos, de forma que la autonomía de los sistemas IA esté alineada con valores humanos –idealmente, de manera demostrable.[15]

En lo que sigue vamos a asumir simplemente de forma más genérica que la gobernanza de la IA se centra en el interés por los problemas éticos que la regulación mediante algoritmos puede incrementar o crear.[16] Hay en estos momentos un amplio consenso acerca de la necesidad de aplicar principios éticos en el desarrollo de la Inteligencia Artificial. El año 2019 ha visto nacer una multitud de manifiestos y declaraciones al respecto, en la línea adoptada por los estándares de la Consejo de la Organización para la Cooperación y el Desarrollo Económico (OCDE)[17] y del Institute of Electrical and Electronics Engineers (IEEE).[18] Esto ha sucedido de forma simultánea en Europa[19], Estados Unidos y Australia[20]. Incluso la católica Academia Pontificia para la Vida, Microsoft, IBM, la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO) y el Gobierno italiano, han firmado el *Llamamiento para una Ética de la Inteligencia Artificial*, un documento creado para apoyar el enfoque ético de la Inteligencia Artificial, en el marco de la Asamblea Plenaria de dicha Academia, celebrada en el Vaticano del 26 al 28 de febrero de 2020.[21]

---

[15] Esta última estrategia fue formulada por S. Russell en la forma de un desafío (Russell 2017). La misma inquietud ha suscitado una cantidad considerable de actividad que queda bien ilustrada en la reciente reseña de Toljeimer *et al*. (2020).

[16] Véase un elenco representativo de los más de cincuenta manifiestos y códigos éticos existentes relativos a AI y ética en: https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/

[17] Véase: https://www.oecd.org/going-digital/ai/principles/

[18] Véase: https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

[19] Véase:https://publications.jrc.ec.europa.eu/repository/bitstream/JRC113826/ai-flagship-report-online.pdf

[20] Véase: https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles

[21] Véase: http://www.academyforlife.va/content/pav/en/events/intelligenza-artificiale.html

Los gigantes tecnológicos ya habían apostado por esta estrategia global hace un poco más de tiempo[22], bajo el influjo de las discusiones sobre las *General Data Protection Regulations* (GDPR) que empezaron en Europa en 2012 y generaron un gran debate en torno a las protecciones debidas (hasta su aprobación por el Parlamento Europeo el 4 de abril de 2016 y la definitiva entrada en vigor de este Reglamento el 24 de mayo de 2018). En julio de 2019, el *Ciclo de Sobreexpectativas para la Inteligencia Artificial* de Gartner situó a la ética digital en la cima de estas expectativas, mientras la gobernanza mediante AI se hallaba aun –como la Inteligencia Artificial general– al principio de la pista de lanzamiento. De la misma manera, las estimaciones de Gartner cifraban de cinco a diez años el tiempo de desarrollo para que esta tecnología ética y regulativa adquiriera la madurez suficiente para una aplicación industrial.[23] Esta calificación no solo se ha mantenido en los *Ciclos* siguientes (2020 y 2021), sino que la ética digital «mediante el diseño» (*by design*) se contempla como un elemento proactivo de creación de valor más allá del mero cumplimiento y se recomienda la creación de una sección de ética digital en todos los planes de proyectos y documentos estratégicos.[24]

El recurso a la ética es extremadamente importante, puesto que es la única dimensión regulativa por consenso a nivel global relativa a valores fundamentales. Sin embargo, puede ser utilizado también como expediente justamente para el incumplimiento de las regulaciones, como ha revelado el reciente escándalo del laboratorio ético del MIT.[25] Hay que contar con la tendencia de las corporaciones a evitar pérdidas económicas a cualquier coste y a presionar a gobiernos e instituciones de investigación para lograrlo. Como veremos más tarde se trata de una «presión en cascada», porque los propios Estados democráticos tienden a trasladar la presión a ciudadanos, expertos e investigadores. Pero la tendencia hacia una regulación ética imbuida en los sistemas responde a una preocupación genuina de las empresas, los gobiernos, los filósofos y los propios científicos computacionales.

Se han sucedido entre los investigadores los manifiestos por una IA responsable (*Responsible AI*), confiable (*Trustworthy AI*) y capaz de rendir cuentas a la ciudadanía (*Accountable AI*). El *Onlife Manifesto* (Floridi *et al*., 2015), el *Manifesto for a conscientiuos design* of *hybrid online social systems* (Noriega *et al*., 2016), y la reciente *A research agenda for hybrid intelligence* (Akata *et al*., 2020) constituyen buena prueba de ello.

---

[22] Véase, por ejemplo, *Everyday Ethics for Artificial Intelligence* (IBM, 2014) y los resultados de *FATE: Fairness, Accountability, Transparency, and Ethics in AI,* el grupo de trabajo creado por Microsoft para este mismo objectivo.

[23] Véase los informes de Gartner, Judah (2019a y 2018b), Judah y O'Kane (2019).

[24] Véase los informes de Gartner, de Hamer *et al.* (2020) y Buytendijk *et al.* (2020). Este último recomienda la adopción de una ética del cuidado basada en cuatro pilares: empatía, responsabilidad, competencia y confianza.

[25] Antiguos miembros del laboratorio han denunciado explícitamente que «el discurso de una "IA ética" estaba estratégicamente alineado con los esfuerzos de Silicon Valley por evitar la ejecución de sanciones jurídicas para limitar las tecnologías cuestionables» (Ochigame, 2019).

Cabe mencionar los intentos de Floridi y Cawls (2019) y Dignum (2019) de ofrecer una perspectiva sintética general de los principios éticos para AI.[26] Hay tres aspectos que nos gustaría destacar en esa integración:[27]

– En primer lugar, la importancia de la explicabilidad (Floridi *et al*., 2018; Floridi y Cawls, 2019). El uso de lenguajes artificiales puede resultar en decisiones opacas, difíciles de entender y contar. Esta es una característica del aprendizaje automático –especialmente en el uso de *deep learning*–. Las inferencias computacionales con datos masivos (*Big Data*) responden a una razón numérica distinta de las de las decisiones humanas. Por ejemplo, en analítica jurídica, los algoritmos de aprendizaje automático clasifican, buscan y cuentan según sus propios requisitos y especificaciones. El resultado constituye una nueva información diferente de la extraída de los casos originales que se usaron como entrada. Como observa Vanderstichele (2019, p. 47-48), el proceso de creación de algoritmos «utiliza una metodología que es al menos parcialmente diferente de la metodología utilizada en la adjudicación». No puede ser clasificado ni como precedente –con validez jurídica– ni como hecho. Y sin embargo sus resultados (*outcome of Legal Analytics*, ocLA) tienen valor normativo. Esto debe ser objeto de explicación detallada en cada caso.

– En segundo lugar, la aplicación de principios éticos no puede hacerse en el vacío. Siempre hay intereses en juego que afectan a la transparencia de las actuaciones públicas y a la vulnerabilidad de los propios investigadores.[28]

---

[26] Floridi y Cawls (2019) identifican un marco general que consta de cinco principios fundamentales: beneficencia, no maleficencia, autonomía, justicia y explicabilidad. La «explicabilidad» se entiende en el sentido de la epistemología de la inteligibilidad (como respuesta a la pregunta «¿cómo funciona?») y en el sentido ético de rendición de cuentas (como respuesta a la pregunta: «¿quién es responsable de la forma en que funciona?»). Dignum (2019, p. 3) ofrece una visión integrada: «La IA representa un esfuerzo coordinado para comprender la complejidad de la experiencia humana en términos de procesos de información. No se refiere solamente a cómo representar y usar lógicamente la información, sino también a las cuestiones de cómo se observa (visión), se mueve (robótica), se comunica (habla y lenguaje) y se aprende (memoria, razonamiento, clasificación)».

[27] Los investigadores han señalado los diversos aspectos de los algoritmos que pueden resultar problemáticos, como: i) prueba no concluyente; ii) prueba inescrutable (efectos de «caja negra»); iii) prueba mal dirigida; iv) resultados injustos; v) efectos transformadores; y vi) trazabilidad (Morley *et al*., 2019).

[28] El caso más reciente (mientras escribimos estas líneas) es el de Vanessa Teague. No será el último. Vanessa, junto con un equipo de la Universidad de Melbourne, solo tardó ocho horas en desanonimizar la base de datos de expedientes médicos supuestamente segura que el Departamento de Salud había puesto a disposición pública en Australia. Dos millones y medio de expedientes médicos quedaron al descubierto. Los investigadores informaron puntualmente de buena fe a este Departamento del Gobierno federal (Culnane *et al*., 2016; Culnane y Leins, 2020). Fue una actuación ética que se reveló arriesgada: después de las presiones recibidas por la Universidad por parte del Departamento de Salud, la investigadora acaba de renunciar a su posición como profesora (N/A, *The Guardian,* marzo 2019).

– En tercer lugar, la actuación en base a principios éticos tiene siempre una dimensión pública, *i.e.* política, y necesita un nivel intermedio que defina las condiciones pragmáticas –procedimentales y de contenido– de las actuaciones. Hay una distancia considerable entre los principios y su posible implementación en el diseño y la programación (Morley *et al*., 2019). Y esto afecta también al ejercicio de derechos y a la forma en que las distintas administraciones han incorporado rápidamente técnicas de IA para la gestión de conocimiento, como se mostrará a continuación.

5.5. Del derecho a la gobernanza jurídica digital: la emergencia de un nuevo espacio público

Hace tiempo que administraciones, gobiernos y Estados utilizan técnicas de IA para la gestión de políticas públicas, así como para la gestión de las bases de datos de las que disponen con información privada de los ciudadanos. Se trata de una información individualizada y granular, que empezó primero en los modelos de gobernanza corporativa (COBIT[29], por ejemplo) y que se extendió más tarde a la propia arquitectura empresarial de las corporaciones, en la actualidad completamente digitalizada (TOGAF[30], por ejemplo).

Hay que distinguir dos aspectos distintos en el proceso de extensión: i) los procedimientos automáticos de control y monitorización interna de la ejecución de procesos (*compliance*); y ii) la concepción misma de la arquitectura (exportable a los Estados y a la red de agencias públicas) de las corporaciones.

La historia, a grandes rasgos, es la siguiente. La automatización de la adhesión a la ley (*compliance*) ha ido en aumento en los últimos veinte años, debido en primer lugar a la actuación de grandes corporaciones como Enron (con Arthur Andersen) o WorldCome (Segal, 2019; Hayes, 2020) contra los intereses de los consumidores e inversores a principios del siglo xxi; y, en segundo lugar, a la crisis financiera de 2008 (O'Neil, 2016). Los legisladores en USA reaccionaron mediante el endurecimiento de las auditorías y el establecimiento de mayores penas contra el fraude –la promulgación de leyes como el *Sarbane-Oaxley Act* (conocida como *Corporate and Auditing Accountabi-*

[29] COBIT (Control Objectives for Information and Related Technologies) se refiere a la estructura y conjunto de instrumentos creado por ISACA (Information Systems Audit and Control Association) para la gestión de procesos de IT. COBIT identifica cinco procesos: «Evaluate, Direct and Monitor (EDM); Align, Plan and Organize (APO); Build, Acquire and Implement (BAI); Deliver, Service and Support (DSS); and Monitor, Evaluate and Assess (MEA)» (véase: https://www.isaca.org/resources/cobit).

[30] The Open Group Architecture Framework (TOGAF) es una arquitectura empresarial modular basada en la ingeniería de la gestión y del conocimiento. Ofrece un marco para el diseño, la planificación y la aplicación de instrumentos para la gobernanza de los sistemas de información (véase: https://www.opengroup.org/togaf).

lity, Responsibility, and Transparency Act, 2002) y, más tarde, la *Foreign Account Tax Compliance Act* (2010). El sector bancario ha adoptado (hasta el momento) medidas más suaves, como el *Basel III Accord of the Basel Committee on Banking Supervision*, de carácter voluntario y preventivo.[31]

Sea como fuere, esta situación ha incentivado la adopción de medidas de control internas y externas en las corporaciones y empresas.[32] La aplicación de estas medidas empezó a denominarse *Compliance by Design* (CbD), *by Detection* (CbDt) y *by Default* (CbDf), dependiendo del momento de aplicación de las medidas previstas.[33] En CbD, la verificación de conformidad es automática o semiautomática y se realiza por adelantado, con anterioridad y dentro de la etapa de ejecución misma para minimizar riesgos, ahorrar recursos y aumentar la eficacia y eficiencia del proceso de gestión. Esto ha fomentado la creación de múltiples formalismos y plataformas para modelar, operar y verificar procesos (desde lenguajes de representación como BPML o Akoma Ntoso; de modelado de dominios como OWL y ejecutables como RuleM*;* hasta lógicas diversas como variantes de las lógicas de *entrada salida* (I/O) o las lógicas derrotables (*defeasible*) (Casanovas *et al*., 2018).

Los lenguajes tienen un componente –o una referencia– semánticos, puesto que se trata de planificación y, por lo tanto, en las regulaciones hay que definir el marco de referencia, los agentes, los objetos y las acciones que son esperables, posibles, permitidas, facultativas o directamente prohibidas. En los próximos desarrollos de la Web de Datos (Web 4.0), los agentes autónomos de *software* podrán interaccionar entre sí y con los humanos, dando lugar a una arquitectura híbrida, donde la identidad y la identificación de los sujetos, objetos y acciones estarán ya especificadas como entidades singulares en la red (Casanovas *et al*., 2016; Francesconi, 2018; Akata *et al*., 2020). La dimensión semántica refiere y formaliza el contenido de las dimensiones social y jurídica. Estas dos dimensiones, hasta hace poco, se expresaban solamente en la lengua natural, no mediante código y lenguajes formales (Poblet *et al*., 2019).

---

[31] Véase el reciente marco integrado de estándares (Basel Framework, 2019) del BCBS en: https://www.bis.org/basel_framework/index.htm

[32] En términos generales, el cumplimiento puede entenderse como conformidad con los requerimientos normativos previamente definidos o conformidad con un conjunto de restricciones (regulatorias). *Regulatory Compliance* denota un conjunto de requerimientos previamente seleccionado. Este conjunto también puede definirse de distintas formas. Por ejemplo, los estándares de la Organización Internacional de Normalización (ISO/IEC) establecidos para sectores industriales y comerciales exigen la conformidad de los procedimientos y procesos comerciales con las leyes, regulaciones, estándares, prácticas recomendadas o requisitos similares. De acuerdo con la ISO / IEC 27002, «La organización debe identificar y documentar sus obligaciones a autoridades externas y a otros terceros en relación con la seguridad de la información, incluidos propiedad intelectual, registros [comerciales], privacidad, información relativa a la identificación personal y criptografía».

[33] El cumplimiento por detección (CbDt) implica una verificación de conformidad durante o después de la ejecución de las reglas. Por lo tanto, si se detecta una conducta no conforme, el proceso necesita ser rediseñado. Por el contrario, el cumplimiento mediante diseño (CbD) significa que el conjunto de reglas se tiene en cuenta en la etapa de diseño.

En este escenario, el derecho no queda inmune. Se produce una tensión entre la forma horizontal de entender la aplicación de las regulaciones y una forma vertical. En la primera, la estabilización de los sistemas emerge de las interacciones (transacciones, contratos, acuerdos, formas de *crowdsourcing*, etc.). En la segunda, las condiciones de la autoridad se reflejan en la arquitectura regulativa misma, imponiéndola verticalmente, por ejemplo «regimentando» los sistemas o creando *softward* de gestión administrativa de la interoperabilidad de los servicios y de la identidad digital que presuponen esquemas binarios de autoridad/obediencia o de obligaciones automáticamente ejecutables sobre los ciudadanos. Los instrumentos se dividen en obligatorios y no obligatorios, de forma discreta. [34]

Cabe notar que existen al menos dos presupuestos para esta arquitectura: i) requiere un sistema de acceso controlado mediante la identidad digital; y ii) su generalización puede implementarse mediante el uso de sistemas de *compliance* para reducir los costes y minimizar los riesgos.

Para describir, entender y finalmente regular el espacio público emergente, los instrumentos jurídicos tradicionales (aunque importantes) ya no son suficientes, puesto que los instrumentos digitales tienen su propia dinámica. Se trata de instrumentos de gobernanza jurídica, más que de aplicación del derecho tal y como los hemos conocido hasta ahora. Esto último tiene riesgos. En algunas formulaciones, la prohibición se iguala con la imposibilidad, *i.e.* como diría C. A. Petri, «las cosas que no nos gustan no deben ser prohibidas, deben ser imposibles [de realizar]» (citado en Lohmann, 2013).

Es importante también señalar la tensión entre el espacio público institucionalizado desde las distintas corporaciones y administraciones nacionales e internacionales y el espacio público emergente de las transacciones en el mercado y en la sociedad digital global. ¿Hay alguna manera de armonizar estas dos tendencias? ¿Cómo conjugar los dos ejes, horizontal y vertical, del derecho para regular los distintos niveles de autonomía artificial? ¿Qué papel tendrá la autoridad frente a la descentralización en la implantación de sistemas de IA?

No tenemos una respuesta clara, por ahora. La posibilidad de una recentralización de los Estados-nación y de la recorporativización de las empresas –*i.e.* la posibilidad de dictaduras digitales y de oligarquías dominantes– ha sido mencionada varias veces en la literatura reciente (Shadbolt y Hampson, 2018; Zuboff, 2019).

Creemos que esta tensión puede resolverse mediante la creación de instituciones, verticales, horizontales o mixtas, que ayuden a tender puentes

---

[34] En la arquitectura de EIRA© (European Interoperability Reference Architecture), por ejemplo, alineada con e inspirada en TOGAF, se contemplan cuatro niveles de interoperabilidad: jurídica, organizativa, semántica y técnica. El nivel jurídico se integra como una capa específica y modular que asegura la «legalidad» de las actuaciones de la administración relativas a los servicios públicos EIRA (2019, p. 40).

entre las plataformas de servicios y las iniciativas y derechos de ciudadanos, consumidores y usuarios. Creemos también, que la IA puede contribuir a esta creación.

5.6. IA para gobernar la IA

En ese sentido hay dos grandes líneas de actividad en IA que responden a ese desafío. Por una parte, está la creación de espacios «híbridos» en línea (dentro de los que interactúan agentes humanos y artificiales) en los cuales la gobernanza de las interacciones es el factor determinante de su diseño; y por otra, está la llamada a imbuir valores en los sistemas de IA.

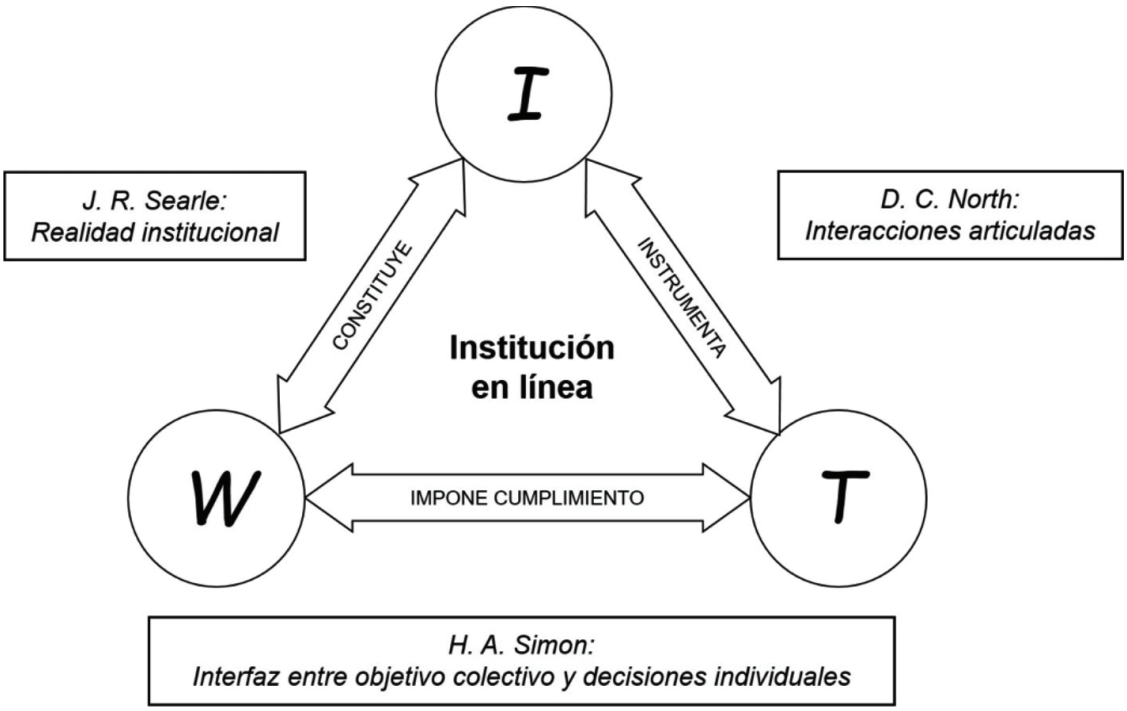
La primera línea responde a la cada vez más frecuente interacción social mediada por internet y en la que, a la vez que los humanos, participan sistemas artificiales inteligentes. Para este tipo de contexto se han desarrollado diversos marcos conceptuales y tecnológicos que abordan la necesidad de gobernar tales interacciones. El texto de Aldewereld *et al.* (2016), da una buena perspectiva de la variedad, grado de madurez y ámbitos de aplicación de estos esfuerzos. Desde un punto de vista abstracto, esos ejemplos pueden entenderse como «un *espacio social artificial* habitado por agentes autónomos, naturales o artificiales, cuyas interacciones suceden en línea, sujetas a las restricciones impuestas por el sistema».[35]

En el momento en que ese aspecto de gobernanza se vuelve explícito (hay unas reglas del juego claras que el propio sistema vigila que se cumplan), se les puede denominar «instituciones en línea» (*Online Institutions*) (Noriega *et al*., 2021a). Las instituciones en línea se denominan así porque recogen para ese entorno digital híbrido tres intuiciones básicas sobre lo que son las instituciones convencionales: 1) constituyen un *subcontexto W* del mundo social que se gobierna bajo sus propias convenciones, (Searle 1991); 2) Ese espacio está sujeto a un conjunto de *restriciones artifiiciales* (o intitucionales, *I*) que articula las interacciones de los agentes (North 1990); y 3) constituyen una *interfaz* entre los modelos de decisión de los individuos para la consecución de un objetivo colectivo (Simon 1996), que en el caso de las instituciones en línea es el substrato tecnológico (*T*) que permite la interacción en línea de los agentes. Cada una de esas tres intuiciones da lugar a un constructo funcional, formal, y computacional (*W,I,T*) que, para conformar una institución en línea, se combinan como se describe en el gráfico 5.1.

---

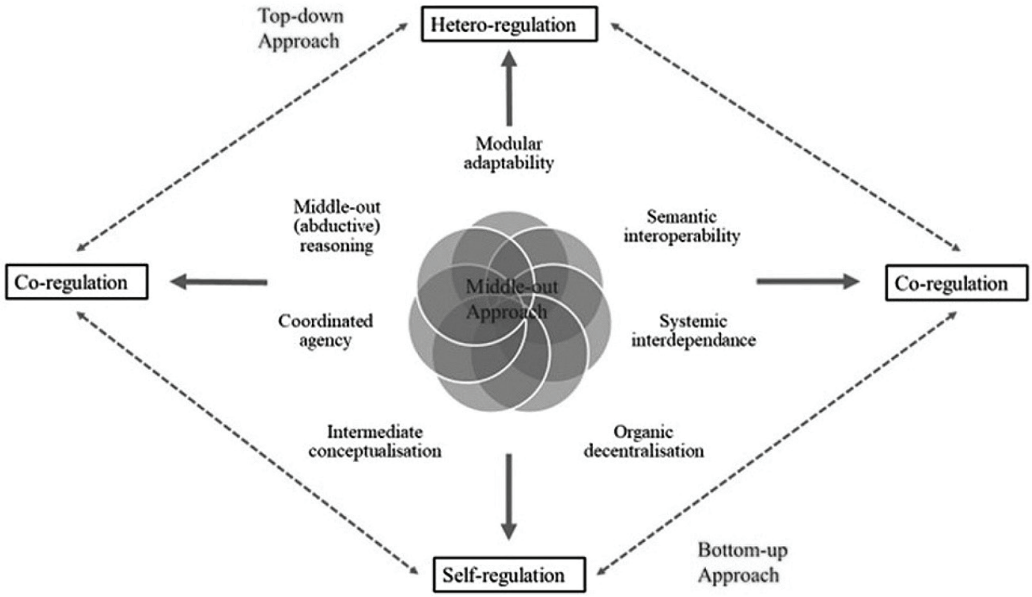
[35] Como se puede observar en los ejemplos de esos marcos que se describen en (Aldewereld *et al*., 2016), desde un punto de vista formal, algunos de esos marcos se pueden entender como sistemas multiagente normativos ya que la gobernanza se busca instrumentar básicamente a través de sistemas basados en normas (Adrighetto *et al*., 2013). Sin embargo, cabe destacar que otras propuestas derivan de la tradición más clásica de control –dentro de la perspectiva de los sistemas socio-técnicos (actividad humana apoyada en sistemas de cómputo) (Trist, 1985)– y su generalización a sistemas socio-cognitivos (agentes artificiales con algunas capacidades de razonamientos social).

Gráfico 5.1. *La perspectiva WIT de las instituciones en línea y sus interrelaciones.* (I: institucional, T: tecnológica, W: [*Working*] aplicaciones)



Fuente: Noriega, Padget y Verhagen (2021).

Gráfico 5.2. *La perspectiva de Middle-out approach para la gobernanza jurídica de la IA*



Fuente: Elaboración propia a partir de Pagallo, Casanovas y Madelin (2019), para encajar los mecanismos de auto, hetero y corregulación.

Para que esa institución en línea sea efectiva –para que las interacciones tengan efectos reales–, esta debe «anclarse» en un contexto socioeconómico específico para lo cual, la institución en línea debe ser compatible con las prácticas sociales, el entorno tecnológico y el contexto legal en el que se inserta. El anclaje tiene lugar a partir de las funciones pragmáticas del sistema al mismo tiempo que su gobernanza dentro de un escenario social determinado. Se trata, pues, de una aplicación «situada». Sucede en lo que se ha denominado «meso-nivel» (Poblet, Casanovas y Rodríguez-Doncel, 2019) o, desde el punto de vista metodológico, en el enfoque de dentro-afuera (*middle-out approach*) para la gobernanza jurídica de instrumentos de Inteligencia Artificial (Pagallo, Casanovas y Madelin, 2019) (véase gráfico 5.2).

En el nuevo estadio de la red, esta forma institucional y modular de concebir los instrumentos de regulación podría articular los comportamientos y las acciones. Se trata de la gestión de derechos en un marco que puede ser aplicado de forma flexible en múltiples situaciones y escenarios con una pluralidad de fuentes normativas, no solo estatales, sino también sociales (ética, protocolos, estándares, prácticas recomendadas y formas de *soft law*). Denominamos a este marco general Estado de derechos (en inglés, *metarule of law*), para diferenciarlo del Estado de derecho de los siglos XIX y XX.

La formulación de un marco regulativo semejante presenta problemas específicos que son distintos de los de la formulación política de las instituciones de gobierno.[36] Imbuir las protecciones de derechos humanos, fundamentales, o de protección de datos (GDPR), en los sistemas computacionales no puede hacerse de forma completa (Koops y Leenes, 2014). Requiere una formulación semántica y la reinterpretación constante de valores, normas y principios, en una pragmática dinámica y variable (Casanovas, Rodríguez Doncel y González Conejero, 2017). Si algo hemos aprendido en la gestión de derechos es que pueden coexistir modelos de gobernanza automática basados en esquemas computables de derechos con instituciones que operen las gestiones de validación, monitorización y auditoría mediante lenguaje natural y comportamiento humano.

Desde este punto de vista, el uso de diversas lógicas y otros formalismos es un avance importante para fundamentar estos sistemas, pero se requieren decisiones meta-lógicas sobre valores y principios para que puedan ser aplicados dentro de un conjunto de instrumentos prácticos. Esquemas clásicos, como el de Hohfeld, siguen aun dando juego en las formulaciones de sistemas normativos multiagente y en la creación de nuevos lenguajes formales para el

[36] Uno de los problemas irresueltos es la determinación de las condiciones para crear ecosistemas regulativos (y jurídicos) estables. Otro de los problemas es cómo alinear las tecnologías de *blockchain* con las de la Web semántica (Engish, Auer y Domingue, 2016), o cómo tratar el problema de la autoridad (*oráculos*) y la resolución de conflictos (Allen *et al*., 2020; Poblet *et al*., 2020).

derecho.[37] El *LegalRuleML*[38] jurídico también ha sido desarrollado desde el punto de vista semántico, al mismo tiempo que *LegalXML.*[39] Los lenguajes de derechos existen desde hace veinte años y han sido mejorados con el uso.[40] El denominado *Open Digital Rights Language*, ODRL) ha devenido un nuevo estándar del W3C en 2018.[41] Se trata de un lenguaje que permite al usuario gestionar su interacción –sus *transacciones*– con las plataformas de servicios. Puede escoger un producto de consumo (música, fotografía…) y regular sus términos de uso, especificando qué quiere/puede y no quiere/puede hacer. Este es el mundo de los datos vinculados (*linked data*), algunos en abierto, otros con restricciones de carácter privado.

El nivel de complejidad que puede alcanzar la vinculación ha dado origen a la investigación más reciente, tanto en el ámbito de la computación como en el de la ciencia política y el derecho. Los ingenieros del conocimiento trabajan en la configuración de grandes repositorios de datos (vocabularios) para ser reutilizados y en la construcción de las ontologías computacionales que pueden gestionarlos. Es un proceso iterativo de reingeniería, para crear patrones de diseño a partir de las ontologías existentes (ODP, *Ontology Design Patterns*), para mejorar sus especificaciones, para crear nuevas ontologías locales y para reestructurar y facilitar la creación de sentido en situaciones específicas.[42] No sabemos aun si los ODP serán utilizados de forma masiva, puesto que la explotación de los datos puede hacerse mediante técnicas de IA que no requieren una guía semántica. Representan de momento una pequeña parte de la investigación en la web de datos y AI y deben articularse con ella para ofrecer resultados que puedan escalar a grandes bases de datos de forma automática, dotándolas de sentido.

Esto nos lleva a la segunda línea de trabajo, en la que se busca usar la propia IA para controlar la autonomía artificial; es decir, imbuir valores en los sistemas autónomos artificialmente inteligentes. Este objetivo puede aproximarse, *de facto*, mediante estándares, lineamientos y prácticas que encarnan los valores deseados, como se propone en el programa de *diseño alineado éticamente* (EAD) del *Institute of Electrical and Electronics Engineers* (IEEE 2019). Otra estrategia es programar los valores dentro de los propios sistemas autónomos artificiales. Para este caso conviene distinguir dos enfoques complementarios: 1) construir modelos de decisión para los sistemas autónomos

---

[37] Hay un largo etcétera de artículos dedicados a la reformulación del esquema de derechos de Hohfeld, desde el denominado A-Hohfeld language (Allen y Saxon, 1995) hasta el empleo de redes de Petri para su reconstrucción (Sileno, 2016).

[38] Véase: https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=legalruleml

[39] LegalXML ha sido aceptado como standard por OASIS (véase: http://www.legalxml.org/).

[40] Véase una descripción y una corta historia de su evolución en Casanovas (2015).

[41] Véase el gráfico del modelo de ORDL en Iannella y Villata (2018).

[42] Véase los ODP existentes que han sido sometidos a certificación en: http://ontologydesignpatterns.org/wiki/Submissions:ContentOPs. Ya existen algunos patrones de ontologías (ODP) –en la misma dirección apuntada por los ODRL. Por ejemplo, para la representación de facturas (Gangemi y Gómez), para licencias (Rodríguez Doncel *et al*.), para datos personales (Pandt), para reclamaciones de consumo (Santos, Rodríguez Doncel *et al*., 2019).

que incluyan consideraciones éticas (de forma que las decisiones que tomen los sistemas autónomos sean demostrablemente consistentes con unos valores dados); o (2) construir un entorno de interacción constreñido por una regulación que conduce a que esos valores se satisfagan. En ambos casos hay que enfrentar el arduo problema de hacer operacional la noción de valor, de manera que pueda ser implementado en un sistema y que pueda determinarse objetivamente si ese valor está siendo satisfecho y en qué grado. Este ensayo no da cabida a una discusión detallada de estos enfoques pero los recientes trabajos de Yazdanpanah *et al*. (2021) y de Ramchun *et al*. (2021), ilustran cabalmente la complejidad de elucidar los problemas de operacionalización y Noriega *et al*. (2021b) ilustra la manera de cómo imbuir los valores en una institución en línea.

5.7. A MODO DE CONCLUSIÓN

La IA es una tecnología disruptiva –irruptora, perturbadora, desestabilizadora– y nos plantea con crudeza la paradoja de Collingridge: no la comprendemos del todo, pero podemos anticipar algunos de sus efectos negativos y positivos. Y sabemos que ya están siendo extraordinariamente significativos.

La moraleja de la trágica historia escrita por Mary Shelley en 1818 es que el monstruo no es el producto de la electricidad y la cirugía plástica, sino de las confusas elecciones morales del Dr. Víctor Frankenstein y del contexto sociocultural en el que estas se materializaron. Por ello, conviene aclararnos sobre las preguntas morales que guíen la creación y los usos de la inteligencia artificial.

A riesgo de banalizar el desafío, lo inmediato e ineludible es un esfuerzo constante y colectivo para comprender: sensibilizarnos y educarnos. Para ello se da la necesidad de un «currículo de» a todos los niveles y en el que participen todas las especialidades. Ese currículum debería capacitarnos para asumir con responsabilidad las tareas de creación, explotación y uso de las contribuciones científicas y tecnológicas de la IA. A la par de este esfuerzo, es necesario adaptar las salvaguardas convencionales al caso de la IA y finalmente incorporar las consideraciones éticas como parte del diseño mismo de los sistemas inteligentes autónomos, de forma que el comportamiento de estos esté efectiva y demostrablemente alineado con los valores humanos.

Pero debemos ser conscientes de que el desafío tiene, por una parte, una dimensión temporal inusitada y, por otra parte, una dimensión ontológica capital. En otras innovaciones disruptivas, como la agricultura y la ganadería, la humanidad dispuso de decenas de miles de años para experimentar y adaptarse a la humanización de los recursos naturales y asimilar la inteligencia natural de los seres vivos. Se puede argumentar que el origen de la IA viene de lejos; que solo han transcurrido unos tres mil años desde la invención de la escritura y que solo pasaron unos cuatrocientos desde la introducción de la imprenta, hasta la fotografía y las telecomunicaciones. La inteligencia artificial existe solo desde hace unos sesenta años e Internet apenas llega a los treinta pero, sin

embargo, esa confluencia nos impone súbitamente un nuevo hábitat en el que ya comenzamos a vivir en una realidad aumentada.

Hemos afirmado en suma que, en el marco de la paradoja de Collingridge, para contender con los efectos negativos de la IA es necesario gobernar la autonomía de los sistemas artificialmente inteligentes. También es necesario reconocer que la vida en esa realidad aumentada conlleva un cambio fundamental en la forma de entender la identidad, la interacción social, la cultura, el territorio y el espacio. La gobernanza en ese hábitat, en el que además de seres vivos habitan sistemas artificialmente inteligentes, tendrá que ser necesariamente distinta a la que conocemos hoy. Ya tenemos muestras de cómo difiere de la forma tradicional de entender el derecho y las formas de regulación, pero también comenzamos a encontrar los nuevos enfoques y los recursos para instaurarla, en los que la propia inteligencia artificial deberá jugar un papel esencial.

Agradecimientos

Este ensayo se deriva de apoyos recibidos en diversos proyectos, entre ellos:

TIN2017-89758-R; (2018-2020) *CIMBVAL: Cómo imbuir valores en la coordinación de redes sociales híbridas*. Ministerio de Economía. Recercaixa 2017 (2018-2020) *AppPhil: Applied Philosophy for the Value-Based Design of Social Network Apps;* Fundación Cultural La Caixa. DER2016-78108-P *Meta-rule of Law* (2017-2019), Ministerio de Economía; LYNX. H2020 ID: 780602 (2018-2020); Australian Government, *Compliance by Design (CbD) and Compliance through Design (CtD)* (2018-2019), CRC D2D,Project DC160051; NGI: ONTOCHAIN, H2020, OntoROPA. ID:1481458; OPTIMAI H2020, ID: 958264 (2021-2023).

Bibliografía

Akata, Zeynep; Balliet, Dan; Maarten De Rijke, Maarten; Dignum, Frank; Dignum, Virginia; Eiben, Guszti; Fokkens, Atkens *et al*. (2020). «A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence». *Computer,* 53(8), pp. 18-28.

Aldewereld, Huib; Boissier, Olivier; Dignum, Virginia; Noriega, Pablo y Padget, Julián (eds.) (2016). *Social coordination frameworks for social technical systems*. Dordrecht: Springer International Publishing.

Allen, Darcy W. E. ; Lane, Aaron M. y Poblet, Marta (2020). «The governance of blockchain dispute resolution». *Harvard Negotiation Law Review*, 25(2), pp. 75-101. Disponible en: https://heinonline.org/HOL/LandingPage?handle=hein.journals/haneg25&div=5&id=&page=

Allen, Layman E. y Saxon, Charles S. (1995). Better language, better thought, better communication: the A-Hohfeld language for legal analysis. En: *Proceedings of the 5th international conference on Artificial intelligence and law* (pp. 219-228). New York: ACM.

Andrighetto, Giulia; Governatore, Guido; Noriega, Pablo y Van der Torre, Leon (eds.) (2013). *Normative Multi-Agent Systems*. Dagstuhl Follow-Ups, n.º 4.

Angle, Colin M. y Brooks, Rodney A. (1990). «Small Planetary Rovers». *IEEE International Workshop on Intelligent Robots and Systems* (pp. 383-388) , 27 de abril. Tsuchiura, Japan: July. Disponible en: http://people.csail.mit.edu/brooks/papers/small.pdf

Ashley, Kevin D. (2017). *Artificial Intelligence and Legal Analytics*. Cambridge: Cambridge University Press.

Bibal, Adrien; Lognoul, Michael; De Streel, Alexandre y Frénay, Benoît (2020). *Impact of legal requirements on explainability in machine learning*. arXiv preprint arXiv:2007.05479. Disponible en: https://arxiv.org/pdf/2007.05479.pdf

Bostrom, Nick (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Botsman, Rachel (2017). «Big data meets Big Brother as China moves to rate its citizens». *Wired UK*, 21 de octubre, pp. 1-11. Disponible en: https://www.wired.co.uk/article/chinese-government-social-credit-score-privacy-invasion

Brown, Tom B.; Mann, Benjamin; Ryder, Nick *et al*. (2020). *Language Models are Few-Shot Learners,* 22 de julio. Disponible en: arXiv:2005.14165

Buytendijk, Frank; Hare, Jim y Clougherty Jones, Lidia (2020). *Digital Ethics by Design: A Framework for Better Digital Business*, ID G00421878). Gartner Report, 31 de julio de 2019 [actualizado el 20 de octubre de 2020].

Casanovas, Pompeu (2015). «Conceptualisation of rights and meta-rule of law for the web of data». *Democracia Digital e Governo Eletrônico*, (12), pp. 18-41.

Casanovas, Pompeu; González-Conejero, Jorge y De Koker, Louis (2018). «Legal Compliance by Design (LCbD) and through Design (LCtD): Preliminary Survey». TERECOM-17@JURIX, *First Workshop on Regulatory Compliance*. Disponible en: http://ceur-ws.org/Vol-2049/05paper.pdf

Casanovas, Pompeu; Palmirani, Mónica; Peroni, Silvio; Van Engers, Tom y Vitali, Fabio (2016). «Semantic web for the legal domain: the next step». *Semantic Web*, 7(3), pp. 213-227.

Casanovas, Pompeu; Rodríguez-Doncel, Víctor y González-Conejero, Jorge (2017). The Role of Pragmatics in the Web of Data. En: F. Poggi y A. Ca-

pone (eds.). *Pragmatics and law: Practical and Theoretical Perspectives* (pp. 293-330). Cham: Springer.

Christensen, Clayton M. (1997). *The innovator's dilemma: when new technologies cause great firms to fail*. Boston, MA: Harvard Business School Press.

CNBC (2014). «Artificial intelligence could end mankind: Hawking». *CNBC*, 4 de mayo. Disponible en: https://www.cnbc.com/2014/05/04/artificial-intelligence-could-end-mankind-hawking.html

Collingridge, David (1980). *The Social Control of Technology*. New York: St. Martin's Pres.

Columbus, Louis (2018). «2018 Roundup of Internet of Things Forecasts and Market Estimates». *Forbes*, 13 de diciembre. Disponible en: https://www.forbes.com/sites/louiscolumbus/2018/12/13/2018-roundup-of-internet-of-things-forecasts-and-market-estimates/?sh=49e8e98a7d83

Culnane, Chris y Leins, Kobi (2020). «Misconceptions in Privacy Protection and Regulation». *Law in Context*, 36(2), pp. 1-12.

Culnane, Chris; Rubinstein, Benjamin y Teague, Vanesa (2016). «Understanding the maths is crucial for protecting privacy». *Pursuit, Engineering and Technology,* de septiembre. Disponible en: https://pursuit.unimelb.edu.au/articles/understanding-the-maths-is-crucial-for-protecting-privacy

Den Hammer, Pieter; Gove, Katie *et al*. (2020). *Digital Ethics: From Compliance Duty to Competitive Differentiator*. ID G00725235. Gartner Report, 17 de julio.

Dressel, Julia y Farid, Hany (2018): «The accuracy, fairness, and limits of predicting recidivism». *Science Advances*, 4(1): eaao5580. Disponible en: https://advances.sciencemag.org/content/advances/4/1/eaao5580.full.pdf

EIRA (2019). *An introduction to the European Interoperability Reference Architecture (EIRA©) v3.0.0*. Disponible en: https://joinup.ec.europa.eu/solution/eira/distribution/eira-v300-overview

Elliott, Eric (2020). *Wht It's Like To be a Computer: An Interview with GPT-3*. [Vídeo *online*] 18 de septiembre. Disponible en: https://youtu.be/PqbB07n_uQ4

English, Matthew D.; Auer, S. y Domingue, J. (2016). Block chain technologies & the semantic web: A framework for symbiotic development. En: J. Lehmann *et al*. (eds.). *Computer Science Conference for University of Bonn Students (*pp. 47-61).

EU High-Level Expert Group on Artificial Intelligence (2019a). *Ethics Guidelines for Trustworthy AI.* Publicaciones de la Unión Europea, 8 de abril. Disponible en: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

EU High-Level Expert Group on Artificial Intelligence (2019b). *Policy and Investment Recommendations For Trustworthy AI*. Publicaciones de la Unión Europea, 26 de junio. Disponible en: https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence

Ferrucci, David A. (2012). «Introduction to "This is Watson"». *IBM Journal of Research and Development*, 56(3-4), May-June, pp. 1-15. Doi: 10.1147/JRD.2012.2184356.

Floridi, Luciano (2013). *The ethics of information*. Oxford: Oxford University Press.

Floridi, Luciano *et al.* (2015). *The online manifesto: Being human in a hyperconnected era*. Cham: Springer.

Floridi, Luciano *et al.* (2018). «AI4People–An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations». *Minds and Machines*, 28(4), pp. 689-707.

Francesconi, Enrico (2018). «On the Future of Legal Publishing Services in the Semantic Web». *Future Internet*, 10(6), p. 48. Disponible en: https://www.mdpi.com/1999-5903/10/6/48

Future of Life Institute (2017). *ASILOMAR AI Principles.* Disponible en: https://futureoflife.org/ai-principles/

Hashmi, Mustafa; Casanovas, Pompeu y De Koker, Louis (2019). «Legal Compliance Through Design: Preliminary Results». TERECOM 2018@JURIX, *Second Workshop on Regulatory Compliance* (pp. 59-72), Groningen: CEUR-WS. Disponible en: http://ceur-ws.org/Vol-2309/06.pdf

Hayes, Adam (2020). «The Rise and Fall of WorldCom» (actualizado el 11 de enero). *Investopedia.* Disponible en: https://www.investopedia.com/terms/w/worldcom.asp

Huhns, Michael N. y Singh, Munindar P. (2005). «Service-oriented computing: Key concepts and principles». *IEEE Internet computing*, 9(1), pp. 75-81.

Iannella, Renato y Villata, Serena (eds.) (2018). *ODRL Information Model 2.2.* W3C, 15 de febrero. Disponible en: https://www.w3.org/TR/2018/REC-odrl-model-20180215/

IEEE (2019). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. En: *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. Disponible en: https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html

Judah, Saul (2019a). *7 Must-Have Foundations for Modern Data and Analytics Governance.* ID: G00433904. Gartner Report, 14 de octubre.

Judah, Saul (2019b). *Hype Cycle for Data and Analytics Governance and Master Data Management*. ID G00369901. Gartner Report, 10 de julio.

Judah, Saul y O'Kane, Bill (2019). *Data and Analytics Strategies Primer for 2019*. ID G00710094. Gartner Report, 5 de febrero. Disponible en: https://www.gartner.com/en/documents/3899971/data-and-analytics-strategies-primer-for-2019

Komninos, Nicos (2009). «Intelligent cities: towards interactive and global innovation environments». *International Journal of Innovation and Regional Development,* 1(4), pp. 337-355.

Koops, Bert-Jaap y Leenes, Ronald (2014). «Privacy regulation cannot be hardcoded. A critical comment on the "privacy by design" provision in data-protection law». *International Review of Law, Computers & Technology*, 28(2), pp. 159-171.

Kurzweil, Ray (2005). *The Singularity is Near*. New York: Viking Books.

Lenat, Doug; Prakash, Mayank y Shepherd, Mary (1986). «CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquistion [sic] Bottlenecks». *AI Magazine*, 6(4), pp. 65-85. Doi: 10.5555/13432.13435

Lohmann, Niels (2013). «Compliance by Design for Artifact-Centric Business Processes». *Information Systems, Special section on BPM 2011 conference*, 38(4), pp. 606-18.

López de Mántaras, Ramón y Meseguer González, Pedro (2017). *Inteligencia artificial, ¿Qué sabemos de?* Madrid: Los libros de la Catarata.

Luck, Michael; McBurney, P.; Shehory, Onn y Willmott, Steven N. (2005). *Agent technology: computing as interaction (a roadmap for agent based computing)*. Southampton: University of Southampton.

Markoff, John (2011). «Computer Wins on "Jeopardy!" Trivial, It's Not». *The New York Times*, 16 de febrero. Disponible en: https://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html

Metz, Cade (2020). «Meet GPT-3. It Has Learned to Code (and Blog and Argue)». *The New York Times*, 24 de noviembre. Disponible en: /www.nytimes.com/2020/11/24/science/artificial-intelligence-ai-gpt3.html?smid=em-share

Morley, Jessica; Floridi, Luciano; Kinsey, Libby y Elhalal, Anat (2019). *From what to how. An overview of AI ethics tools, methods and research to translate principles into practices. arXiv preprint arXiv:1905.06876.*

Noriega, Pablo; Padget, Julian y Verhagen, Harko (2021). Anchoring Online Institutions. En: P. Casanovas y J. J. Moreso (eds.). *Anchoring Institutions in a Semi-Automated World*. Cham: Springer. [En prensa].

Noriega, Pablo; Verhagen, Harko; D'Inverno, Mark y Padget, Julian (2016). A manifesto for conscientious design of hybrid online social systems. En: *Coordination, Organizations, Institutions, and Norms in Agent Systems* (pp. 60-78). LNAI 10315. Cham: Springer.

Noriega, Pablo; Verhagen, Harko; Padget, Julian y D'Inverno, Mark (2021b). «Ethical Online AI Systems through Conscientious Design». *IEEE Internet Computing*. [En prensa].

North, Douglass C. (1990). *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press.

Ochigame, Rodrigo (2019). «Ethical AI. How Big Tech Manipulates Academia to Avoid Regulation». *The Intercept*, 20 de diciembre. Disponible en: https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/

Okamoto, Karl S. (2009). *Legal Profession as Subject: A Bibliography*. Disponible en: https://www.semanticscholar.org/paper/Legal-Profession-as-Subject%3A-A-Bibliography-Okamoto/f207da8e0e-4ff219c57fa2cd9d826549c573d4ce

O'Neil, C. (2016). *Weapons of Math Destruction*. New York: Crown Books.

Pagallo, Ugo. (2013). *The laws of robots: crimes, contracts, and torts*. LGT, 10. Dordrecht: Springer.

Pagallo, Ugo; Casanovas, Pompeu y Madelin, Robert (2019). «The middle-out approach: assessing models of legal governance in data protection, artificial intelligence, and the Web of Data». *The Theory and Practice of Legislation*, 7(1), pp. 1-25.

Poblet, Marta; Allen, Darcy W.; Konashevych, Oleksii; Lane, Aaron M. y Díaz Valdivia, Carlos A. (2020). «From Athens to the Blockchain: Oracles for Digital Democracy». *Frontiers in Blockchain*, 3, p. 41. Doi: 10.3389/fbloc.2020.575662

Poblet, Marta; Casanovas, Pompeu y Rodríguez Doncel, Víctor (2019). *Linked Democracy. Foundations, Tools, and Applications.* Springer Briefs 750. Disponible en: https://www.springer.com/gp/book/9783030133627

Poblet, Marta; García Cuesta, Esteban y Casanovas, Pompeu (2018). «Crowdsourcing roles, methods and tools for data-intensive disaster management». *Information Systems Frontiers*, 20(6), pp. 1363-1379.

Poblet, Marta; Noriega, Pablo y Enric Plaza, E. (2014). «Crowd Intelligence: Foundations, Methods and Practices». *Proceedings of the Sintelnet WG5 Workshop on Crowd Intelligence: Foundations, Methods and Practices*. CEUR Workshop Proceedings 1148.

Ramchurn, Sarvapali D.; Sebastian, Stein y Jennings, Nicholas R. (2021). «Trustworthy human-AI partnerships». *iScience,* 24(8): 102891.

Rodríguez-Doncel, Víctor (2019). *DC25008: Compliance by Design (CbD) and Compliance through Design (CtD).* Disponible en: https://zenodo.org/record/3271506#.Xm79J3L1Y2w

Russell, Stuart (2017). «Provably beneficial artificial intelligence».*The Next Step: Exponential Life, BBVA-Open Mind 9*. Disponible en: https://www.bbvaopenmind.com/wp-content/uploads/2017/03/BBVA-OpenMind-book-The-Next-Step-Exponential-Life-1-1.pdf#p173

Russell, Stuart y Norvig, Peter (2002). *Artificial Intelligence. A modern approach*. New Jersey: Prentice Hall.

Sainato, Michael (2020): «"I'm not a robot": Amazon workers condemn unsafe, grueling conditions at warehouse». *The Guardian*, 5 de febrero. Disponible en: https://www.theguardian.com/technology/2020/feb/05/amazon-workers-protest-unsafe-grueling-conditions-warehouse

Searle, J. R. (2005). «What is an institution?». *Journal of Institutional Economics,* 1(1), pp. 1-22.

Segal, Troy (2019). «Enron Scandal: The Fall of a Wall Street Darling». *Investopedia* (actualizado el 29 de mayo). Disponible en: https://www.investopedia.com/updates/enron-scandal-summary/

Shadbolt, Nigel y Hampson, Robert (2018). *The digital ape: how to live (in peace) with smart machines*. London: Scribe Publications.

Sicular, Svetlana; Hare, Jim y Brant, Kenneth (2019). *Hype Cycle for Artificial Intelligence, 2019*. ID: G00369840. Gartner Report, 25 de julio.

Sileno, Giovanni (2016). *Aligning Law and Action*. Amsterdam: University of Amsterdam. [Tesis doctoral].

Silver, David; Hubert, Thomas; Schrittwieser, Julian; Antonoglou, Ioannis; Lai, Matthew; Guez, Arthur; Lanctot, Marc; Sifre, Laurent; Kumaran, Dharshan; Graepel, Thore; Lillicrap, Timothy; Simonyan, Karen y Hassabis, Demis (2018). «A general reinforcement learning algorithm that masters chess, shogi, and go through self-play». *Science,* 362(6419), 7 de septiembre, pp. 1140-1144.

Simon, Herbert A. (1996). *The sciences of the artificial*. 3.ª ed. MIT Press. [Traducción castellana: *Las ciencias de lo artificial*. Granada: Comares, 2007].

Steels, Luc y Brooks, Rodney (eds.) (2018). *The artificial life route to artificial intelligence: Building embodied, situated agents*. London: Routledge.

Taylor, Josh (2020). «Melbourne professor quits after health department pressures her over data breach». *The Guardian*, 7 de marzo. Disponible en: https://www.theguardian.com/australia-news/2020/mar/08/melbourne-professor-quits-after-health-department-pressures-her-over-data-breach

Tolmeijer, Suzanne; Kneer, Markus; Sarasua, Cristina; Christen, Markus y Bernstein, Abraham (2020). «Implementations in machine ethics: a survey». *ACM Computing Surveys (CSUR)* 53(6), pp. 1-38.

Trist, Eric (1981). «The evolution of socio-technical systems. a conceptual framework and an action research program». Ontario: Ministry of Labour. [Publicación ocasional].

Vanderstichele, Geneviève (2019). *The normative value of Legal Analytics. Is there a case for statistical precedent?* Oxford: Oxford University Press. Disponible en: https://ssrn.com/abstract=3474878 [Tesis doctoral].

Von Ahn, Luis; Blum, Manuel; Hopper, Nicholas J. y Langford, John (2003). CAPTCHA: Using Hard AI Problems for Security. En: E. Biham (ed.).*Advances in Cryptology – EUROCRYPT 2003*. (Lecture Notes in Computer Science, vol 2656). Berlin: Springer, Heidelberg. Disponible en: https://doi.org/10.1007/3-540-39200-9_18

Washington, Anne L. (2018). «How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica Debate». *Colorado Technology Law Journal*, 17(1), pp. 131-160.

WEF. World Economic Forum; Global Agenda Council on Employment (2014). *Matching Skills and Labour Market Needs: Building Social Partnerships for Better Skills and Better Jobs.* Disponible en: http://www3.weforum.org/docs/GAC/2014/WEF_GAC_Employment_MatchingSkills-LabourMarket_Report_2014.pdf

Zuboff, Shoshana (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Profile Books.

Wooldridge, Michael (2009). *An introduction to multiagent systems*. 2.ª ed. New York: John Wiley & Sons.

Wooldridge, Michael (2020). *The Road to Conscious Machines: The Story of AI*. London, UK: Pelican Books.

Yazdanpanah, Vahid; Gerding, Enrico H; Stein, Sebastian; Cirstea, Corina; Schraefel, M.C.; Norman, Timothy J. y Jennings , Nicholas R. (2021). «Different Forms of Responsibility in Multiagent Systems: Sociotechnical Characteristics and Requirements». *IEEE Internet Computing*. [En prensa].