1. ECOSISTEMA DEL DERECHO PENAL DIGITAL: DE LOS DELITOS INFORMÁTICOS AL METAVERSO

Mª Isabel González Tapia

Universidad de Córdoba[1]

I. LA WEB ESTÁTICA 1.0: LO DIGITAL COMO OBJETO Y COMO INSTRUMENTO DE COMISIÓN DELICTIVA: DELITOS INFORMÁTICOS

La era digital, en lo que se refiere a la red internet, se inició con la denominada web 1.0 o web estática, denominada así por tratarse de un entorno en el que la comunicación emisor y usuario era unidireccional. Es decir, en la internet originaria las páginas web eran estáticas, pues se alojaba solo la información que el emisor quería difundir y el usuario únicamente podía acceder, imprimir y almacenarla en su ordenador o en los dispositivos de almacenaje. Las empresas, las instituciones o algunos otros canales ofrecían su información y los usuarios eran meros consumidores pasivos de un contenido ya dado, que no podían modificar ni interactuar con él.

En este contexto surgen nuevos objetos merecedores y necesitados de la tutela penal relacionados con lo digital. Así, más allá de la protección convencional de los elementos puramente patrimoniales, por ejemplo, los ordenadores y sus componentes (*hardware*) o dispositivos externos de almacenamiento o de la propia instalación informática, surgieron nuevas realidades que debían ser también protegidas. Así, por ejemplo, surgió la necesidad de proteger la propia información digital: los datos, los documentos electrónicos, frente a su borrado, manipulación o el mero acceso o conexión indebidas; los programas de ordenador o software o los sistemas informáticos (hardware + software + datos + usuarios).

De otra parte, con lo digital surgieron también nuevas formas de ataque a bienes jurídicos y objetos materiales ya existentes, pero que se veían ahora puestos en peligro o lesionados de forma muy diferente, mediante estas nuevas tecnologías informáticas de la información y de la comunicación (TICs). Por ejemplo, podía cometerse una estafa a través de una transferencia informática no consentida, mediando únicamente "moneda escritural" y dispositivos electrónicos; o fabricarse

---

[1] Mª Isabel González Tapia es profesora Titular de Derecho Penal en la Facultad de Derecho y CCEE. Universidad de Córdoba (España). Contacto: fd1gotam@uco.es.

programas de ordenador cuya finalidad específica fuera la de facilitar ese tipo de fraudes o dañar los dispositivos; o se podía vulnerar la privacidad del usuario a través del acceso no autorizado a sistemas, registros o dispositivos electrónicos[2].

Lo digital, por tanto, surgió como nuevo objeto de tutela y como nuevo medio o instrumento para la comisión de delitos, dando lugar a una categoría delictiva específica de comportamientos delictivos que se denominaron los delitos informáticos. Progresivamente, se incorporaron al Código Penal las infracciones cometidas contra los sistemas informáticos, como los daños en datos, programas y documentos electrónicos o el sabotaje de redes…; y las infracciones cometidas *mediante* tales sistemas informáticos, como el acceso ilícito a un servidor, registro o fichero; la interceptación de comunicaciones, el acceso a fotografías o correos electrónicos privados alojados en dispositivos informáticos, vulnerando las claves de acceso...[3] .

En esta etapa germinal de la Web 1.0, la adaptación del Derecho Penal podría decirse que fue bastante simple y limitada, a pesar de exigir la actualización de algunas figuras delictivas preexistentes y la tipificación de algunas otras nuevas. Sin embargo, con independencia del medio informático, los nuevos riesgos definidos, el contexto, la superestructura y la dimensión de la criminalidad informática se mantenían en gran medida dentro los cauces tradicionales. Los delincuentes y las victimas seguían siendo personas físicas o jurídicas, la extensión era todavía esencialmente estatal y los delitos conservaban su expresión en el mundo físico, con daños reales a bienes y servicios.

II. LA WEB SOCIAL 2.0: EL CONTEXTO DIGITAL COMO ESPACIO DELICTIVO: LOS CIBERDELITOS

Podemos considerar que la Web 2.0 representa un escenario en el que, más allá del acceso a los contenidos difundidos de la internet inicial, se evoluciona hacia una plataforma dinámica, colaborativa y social. En esta etapa la palabra clave es la interacción, entendida como la comunicación recíproca entre usuarios y emisores. Así, los emisores pueden ser tanto difusores de contenido como proveedores de servicios para los consumidores (consumidor electrónico), por ejemplo, para compras *online* o para poder realizar una transferencia en la banca electrónica. Además, los usuarios, además de receptores pasivos, pueden ser también creadores de contenido e interactuar masivamente con otros usuarios. A este entorno pertenecen los *blogs* (sitios web personales), las *wikis* (páginas web colaborativas), los servicios de alojamiento de archivos en la *nube*, así como una amplia gama de servicios en línea, pagos en línea y comercio electrónico. En este momento irrumpen también las redes sociales en su formato original, dentro de las que el usuario

---

[2] Cfr. Miró Llinares, F., *El cibercrimen: fenomenología y criminología de la delincuencia en el ciberespacio*, Marcial Pons, Madrid, 2012, por todos.

[3] Cfr., ampliamente, el trabajo seminal de González Rus Rus, J.J., "Precisiones conceptuales y político-criminales sobre la intervención penal en internet", en AAVV *Cuadernos penales José María Lidón. Nº 4: Delito e informática: algunos aspectos*, Publicaciones de la Universidad de Deusto, Bilbao, 2007, pp. 13-40.

posee un perfil propio o *identidad digital*, con la que comparte información e interactúa masivamente con otras personas.

Con esta transición a la Web 2.0, el Derecho penal comenzó a reconocer que lo digital no era únicamente un medio técnico, sino todo un entorno propio y diverso, de donde surgían[4] riesgos inexplorados para la comisión delictiva, derivados de la interacción masiva y de la identidad digital; nuevas formas de comportamiento lesivo y una dimensión específica de la criminalidad, masiva y sin fronteras, fundamentalmente en los delitos de expresión y contra el patrimonio. Así pues, la expansión de la llamada Web 2.0 significó un cambio sustancial en la manera de concebir la criminalidad en el espacio digital y a una nueva categoría delictiva acumulativa de la anterior: los ciberdelitos, definidos por el Convenio de Budapest (2001)[5].

La noción de ciberdelito es más amplia que la de delito informático. Mientras que el delito informático es una categoría general, que se conecta con la idea de lo digital como "cosa" y como medio comisivo; el concepto de ciberdelito es más amplio, puesto que, además de lo puramente informático, abarca y *se enfoca ya en el entorno*, global e interconectado, en el que se desarrolla la nueva criminalidad[6]. Así, el concepto de ciberdelito, como el término utilizado, pone el foco en la red, en el ciberespacio y en la interacción masiva *online*, esto es, en un contexto global donde las personas socializan, realizan actividades económicas y desarrollan también una parte esencial de su vida, con una conectividad permanente. Al propio tiempo, la noción de ciberdelito se conecta también con la idea de *ciberseguridad*, de una radical importancia ante el uso masivo de la red por parte de los ciudadanos. Y ello, porque la evolución de la red y los avances en las tecnologías digitales tiene como contrapartida la aparición, a su vez, de herramientas informáticas muy sofisticadas y potentes, que amenazan la ciberseguridad, poniendo en riesgo la *integridad* de la estructura de internet, la *confidencialidad* de los datos y la *disponibilidad* de los servicios electrónicos, la *propiedad* intelectual…

Como ciberdelitos característicos de esta fase podrían mencionar algunos ejemplos de diversas categorías. Así, entre los delitos patrimoniales, especialmente *piratería* de los derechos de autor, el *fraude* electrónico en sus múltiples variantes, con técnicas como el *phishing* o suplantación de páginas o correos, para obtener datos personales de los usuarios, como el DNI, números de tarjetas o contraseñas bancarias…; estafas en el comercio electrónico, como las ventas falsas de productos inexistentes o la manipulación de los medios de pago usados en las plataformas p.e. *Paypal* o similares… También ataques *ransonware* de sistemas

---

[4] Considerando que se trata de un entorno especialmente propicio, Agustina, J.R., «Cibercriminalidad y perspectiva victimológica: un enfoque general explicativo de la cibervictimización», *Cuadernos de Política Criminal*, 2014, 114(III), pp. 143-178.

[5] Council of Europe, *Budapest Convention on Cibercrime. European Treaty Series No. 185*, 2001, fecha de consulta en https://rm.coe.int/1680081561

[6] Romeo Casabona, C.M., "De los delitos informáticos al cibercrimen. Una aproximación conceptual y político-criminal", en AAVV-Romeo Casabona, C.M. (Coord.), *El cibercrimen: nuevos retos jurídico-penales, nuevas respuestas político-criminales*, Granada, Comares, 2006, pp. 1-42, p. 9.

informáticos, mediante la introducción clandestina (generalmente enlaces infectados) de *software* malicioso (*malware*) en el sistema del usuario, que bloquea el acceso a los datos o al propio sistema hasta que el usuario efectúa un pago o rescate; o delitos de usurpación de estado civil, mediante *suplantación la identidad digital*, especialmente en la banca en línea, tras la obtención ilícita de las credenciales de acceso, para apropiarse indebidamente del dinero del usuario o hacer cualquier tipo de negocio o transacción en su nombre... Así mismo, también como ejemplos esenciales de ciberdelitos, podríamos mencionar algunos se dirigen a atentar contra la intimidad o privacidad del usuario, impulsados por la capacidad de difusión masiva de las redes sociales (*viralización*) y por la ingente cantidad de datos disponibles, que, o bien han sido obtenidos a través de medios ilícitos: *phishing*, *malware* instalado en los equipos, interceptación de comunicaciones (ataque *MitM)*, redireccionamiento (*pharming*), filtrado de contraseñas…; o bien los propios usuarios han vertido voluntariamente en la red (*huella digital*). Por otra parte, en un lugar también destacado, los delitos contra la libertad e indemnidad sexual de los menores, colectivo especialmente vulnerable y ampliamente victimizado en el entorno digital, que favorece, entre otras, el *anonimato* del victimario. Así, entre los nuevos comportamientos que han emergido destacan el *grooming*, que consiste en contactar, acosar o manipular a un menor con una finalidad sexual, a través de internet; o la *sextorsión*, que consiste en conseguir el acceso sexual con un menor u otros beneficios, mediante chantaje, basado en la amenaza de difundir imágenes íntimas obtenidas por diferentes medios, generalmente a través del propio *grooming*. Ambos fenómenos suelen desarrollarse sobre todo en redes sociales, en plataformas de mensajería o en entornos de juego en línea, donde el delincuente se hace pasar muchas veces por menor para generar confianza y obtener material íntimo o concertar encuentros presenciales. A estas formas principales de ciber-criminalidad podrían añadirse otras como la difusión no consentida de imágenes sexuales, la producción y distribución de material pornográfico con menores o la creación de pornografía infantil simulada, mediante técnicas digitales de manipulación de imagen.

Finalmente, en esta larga lista, ni siquiera exhaustiva, merecen mención especial los denominados delitos de expresión, categoría compleja que englobaría infracciones como las injurias y calumnias, los delitos de apología, enaltecimiento o escarnio propios del denominado "*discurso del odio*" y, en general, cualquier modalidad delictiva cuya conducta típica pueda realizarse a través del lenguaje, oral, escrito o visual. Ejemplos paradigmáticos de ello serían las campañas coordinadas de acoso, difamación o cancelación pública, a través de redes sociales (*ciberacoso*), que se ven favorecidas por el anonimato y la sensación de impunidad; los mensajes intimidatorios dirigidos contra personas por motivos raciales, sexuales o ideológicos, o la difusión de contenido falso o contenido privado para para dañar la reputación ajena, con un potencial de expansión y daño mucho mayor que en los medios tradicionales, dada también la extraordinaria dificultad de erradicar los contenidos lesivos de la red o de ejercer en la práctica el "*derecho al olvido*". En esta misma línea, también deben considerarse los fenómenos de radicalización y

apología del terrorismo en línea, que utilizan la red como medio de captación, adoctrinamiento y difusión de mensajes violentos…[7].

Desde la perspectiva de la adaptación del Derecho Penal, a diferencia de lo señalado con respecto a la criminalidad informática, el ajuste a este nuevo contexto y la cobertura de sus riesgos asociados no está siendo fácil ni, mucho menos, simple. Podría considerarse, ciertamente, que el Derecho Penal y, sobre todo, la dogmática penal, estarían aún en proceso de definir y desarrollar las adaptaciones dogmáticas necesarias, por ejemplo, redimensionado bienes jurídicos tradicionales como la intimidad con la "intimidad en red"; el estado civil, con la "identidad digital"; la integridad moral, con el ciberacoso; o la indemnidad sexual de los menores en los entornos virtuales; los daños a la salud mental derivados del uso abusivo de juegos online, de las redes sociales o de la exposición a contenidos lesivos...

Desde la perspectiva de la persecución efectiva de estos delitos y de la competencia jurisdiccional, la interacción de los usuarios en un "espacio sin fronteras" como es internet, puso también sobre la mesa un serio problema por la intrínseca extraterritorialidad de las infracciones *online*.

De una perspectiva dogmática, entre otras dificultades, en este nuevo entorno surgió ya otra dinámica trascendental para el Derecho Penal: la interacción masiva y en red, de múltiples usuarios permanentemente interconectados, que permite no solo la creación conjunta de contenidos, sino también un uso intensivo de las plataformas sociales. Con ello, como se ha anticipado, el usuario pasó de ser mero receptor pasivo de contenidos ajenos, a ser *agente del entorno digital*. Además, pasó de ser un usuario más o menos individualizado a ser algo así, si se me permite la expresión, como *usuario-red* de una *sociedad-red*[8], y formar parte de un "ente" complejo de usuarios interconectados actuando al unísono, pero, a la vez, desconectados entre sí desde el punto de vista subjetivo. A la evidente problemática de coautoría accesoria por este tipo de acciones colectivas, debe sumarse la específica problemática de la participación en el hecho a través de la mera amplificación del daño que ha hecho un tercero, que generalmente se realiza sin connivencia, lo que da lugar a una *responsabilidad distribuida* que necesariamente debe articularse dogmáticamente[9]. Además, otro aspecto central de esta clase de infracciones es que la gravedad del resultado se incrementa precisamente por la acumulación y difusión masiva de tales conductas, por la exposición pública sin parangón y cuasi perpetua que sufre el usuario y por la partici-

---

[7] Puede consultarse, entre otros, Fernández Teruelo, J.G., *Cibercrimen. Los delitos cometidos a través de internet – estafas, distribución de pornografía infantil, atentados contra la propiedad intelectual, daños informáticos, delitos contra la intimidad y otros delitos en la Red-, Constitutio Criminalis Carolina, Oviedo, 2007;* o AAVV-Romeo Casabona, C.M. (Coord.), *El cibercrimen: nuevos retos jurídico-penales, nuevas respuestas político-criminales*, Granada, Comares, 2006.

[8] Vid. la obra seminal de Castells, M., *La era de la información: economía, sociedad y cultura. Volumen I: La sociedad red.* Madrid: Alianza Editorial, 2006, *passim.*

[9] Sobre este tipo de problemas, pero enfocado desde la perspectiva del concepto de bien jurídico, Kuhlen, L., "Bienes jurídicos y nuevos tipos de delito", en AAVV-Robles Planas, R. (edición española), *Límites al derecho penal: Principios operativos en la fundamentación del castigo*, Atelier, Barcelona, 2013, pp. 225-235.

pación masiva de usuarios que amplifican el daño con "*likes*", comentarios , citas y "*retuits*"; es decir, por el concepto de *viralidad*, Este aspecto de la dimensión del daño a la víctima es clave en el el ciberacoso o el *ciberbullying*, dada la *persistencia del daño* por la dificultad de erradicar los contenidos lesivos en las redes y en los dispositivos privados y por la *difuminación de la frontera entre lo virtual y la vida cotidiana de la víctima.*

A todo lo anterior, debe añadirse la ya conocida dificultad para regular y perseguir penalmente los delitos de expresión, respecto de los que debe guardarse un delicado equilibrio con las libertades fundamentales, en particular con la libertad de expresión, por cuanto en muchas ocasiones la línea divisoria entre el discurso protegido y el discurso de odio sancionable no es clara. Pero a la vez, desde una perspectiva opuesta, las redes sociales han supuesto el entorno perfecto para la proliferación de difamaciones o cancelaciones públicas, incluso de carácter masivo, que no sólo lesionan directamente el honor o la autoestima de una persona, sino que, de forma indirecta, pueden también restringir su libertad de expresión, al generar un "*efecto silenciador*" [10].

Como puede apreciarse de este breve repaso, todo este escenario ha generado, particularmente con respecto a delitos contra bienes jurídicos individuales, retos dogmáticos y político-criminales, que todavía hoy no están del todo resueltos y que afectan tanto a la imputación objetiva, como a la autoría y a la culpabilidad, dado que la acción delictiva se diluye en entornos distribuidos territorialmente y muchas veces anónimos. Este *impás* de la dogmática ha hecho preguntarse a la doctrina más autorizada, si la no deberíamos plantear su tratamiento jurídico como un Derecho Penal de *tercera velocidad*[11].

III. LA TRANSICIÓN HACIA LA WEB 3.0: LA WEB SEMÁNTICA Y LA IRRUPCIÓN DE LA INTELIGENCIA ARTIFICIAL PERSUASIVA

III.1. La Web semántica: hacia una red de datos global interoperable

Con la evolución de la Web 2.0 hacia la Web 3.0 se incorporan en la evolución del contexto digital dos nuevos factores clave para comprender la criminalidad digital actual y la que pudiera estar por venir. A los elementos ya conocidos de acceso a internet, de lo informático y de lo interactivo, se añade ahora un salto cualitativo en la Inteligencia Artificial disponible (en adelante IA) y el desarrollo de la denominada Web Semántica. A ello debe añadirse el impulso

---

[10] Sobre estos problemas, el trabajo clave y de avance de Agustina, J.R., "Nuevos retos dogmáticos ante la cibercriminalidad", en *Estudios Penales y Criminológicos*, Vol. 41 (2021), Artículos doctrinales, Páginas 705-777, en https://revistas.usc.gal/index.php/epc/article/view/7433 (fecha de consulta 10-10-2025), particularmente, pp. 746 y ss, particularmente sobre la problemática de la imputación objetiva y de la responsabilidad distribuida, pp. 757 y ss.

[11] Agustina, J.R., "Nuevos retos dogmáticos ante la cibercriminalidad", cit., p. 714. Y más ampliamente sobre este argumento en pp. 715 a 722.

derivado de los avances de la revolución neurotecnológica, que se abordarán más adelante.

Así, con la Web 2.0, la expansión de *redes sociales* como Facebook, de *plataformas de vídeo* como Youtube y de *motores de búsqueda* como Google… generaron un volumen masivo de datos de interacción de los usuarios, que permitió el desarrollo de la IA en el formato *machine learning,* en particular los modelos supervisados, que aprendían a partir de grandes cantidades de datos etiquetados por personas. A partir de ahí, y apoyada en los avances neurocientíficos y en una estructura basada en redes neuronales artificiales, la IA evoluciona hacia el denominado *deep learning*, que se asienta sobre tres pilares esenciales[12]. El primero es una capacidad computacional extraordinaria, que permite procesar y analizar cantidades ingentes de datos (*big data*), extrayendo correlaciones invisibles para la percepción humana. El segundo, una capacidad predictiva también sin precedentes, derivada de la detección de patrones múltiples y de la posibilidad de formular predicciones, que, a su vez, se retroalimentan en ciclos de aprendizaje continuo (sin supervisión). El tercero, una capacidad operativa masiva y, al mismo tiempo, personalizada, que posibilita tanto la acción a gran escala como la adaptación, por ejemplo, en la publicidad personalizada en tiempo real en función de las preferencias, hábitos de consumo o estados emocionales del usuario.

Pero para que pudiera producirse este salto cualitativo de la inteligencia artificial, la Web Semántica introdujo otro cambio esencial, al incorporar una arquitectura tecnológica dirigida a estructurar los datos y la información digital (*segmentación* y *codificación semántica*) de manera que pudieran ser inteligibles para las máquinas. De esta forma, se posibilitó también que los sistemas de IA evolucionaran hacia modelos más contextuales y refinados, con algoritmos capaces ya de realizar inferencias contextualizadas, relacionando la información dispersa en la red (interoperable) y generando recomendaciones más personalizadas y precisas[13]. La aspiración de esta etapa fue, por tanto, que el entorno web fuera algo así como una base de datos global, donde el contenido procedente de múltiples fuentes fuera interoperable. De esta forma, las máquinas por sí solas podrían ya cruzar, relacionar e interpretar los datos de forma automáti-

---

[12] A los efectos de lo que sigue, nos remitiremos a la definición de IA proveniente del Reglamento de la UE , que define al sistema IA como "un sistema basado en una máquina que está diseñado para funcionar con distintos niveles de autonomía y que puede 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". Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y se modifican los Reglamentos (CE) n.º 300/2008, (UE) n.º 167/2013, (UE) n.º 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144, así como las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial).Publicado en DOUE L, núm. 168, de 12 de julio de 2024, pp. 1–144 y en el BOE n.º 177, de 23 de julio de 2024, pp. 8 564–8 707.

[13] Berners-Lee, Tim, James Hendler y Ora Lassila. «The Semantic Web: A New Form of Web Content That Is Meaningful to Computers Will Unleash a Revolution of New Possibilities». *Scientific American* 284, n.º 5 (2001): 34–43. https://doi.org/10.1038/scientificamerican0501-34.

ca, para dar respuestas coherentes, contextualizadas y cada vez más personalizadas para el usuario.

Este avance tecnológico ha permitido el despliegue de aplicaciones para el usuario que están transformando nuestra experiencia digital y, por tanto, también, nuestra vida cotidiana. Así, por ejemplo, destacan los *algoritmos de recomendación*, que seleccionan y priorizan contenidos en función de los intereses, preferencias, hábitos de navegación y de consumo del usuario, por ejemplo de *Netflix, Spotify* o *Amazon*; o los *asistentes* de voz, que se integran en nuestras tareas domésticas o en nuestras búsquedas, como *Siri o Alexa*; o los *sistemas de publicidad personalizada*, que ajustaban los mensajes comerciales según los perfiles conductuales y emocionales detectados en el usuario[14]…Como puede fácilmente intuirse, todos estos avances técnicos que hacen de puente hacia la Web 3.0, dan un paso más en la integración entre humanos y sistemas digitales, pues a partir de este momento la interacción en la red ya no es sólo humana ni puramente social; sino también híbrida, automatizada e inteligente.

A su vez, sobre ello volveremos a continuación, a partir de esta etapa, sientan las bases para la aparición de nuevos riesgos sobre la privacidad, manipulación y control del ciudadano, asociados al uso masivo de datos, algoritmos de recomendación y sistemas de IA capaces de modelar la conducta humana. Y es que en este contexto emerge también la IA persuasiva, que desplaza el eje de la IA de lo técnico hacia lo manipulativo. Se trata de sistemas IA diseñados para captar la atención de usuario y para influir en sus decisiones, mediante estrategias basadas en la observación constante de su *huella digital* o *datos de interacción*: *cliks*, *likes*, tiempo de uso, productos adquiridos, páginas web visitadas, consultas realizadas, geolocalización... Además, desde el punto de vista económico, todo este entramado viene a consolidar un nuevo modelo basado en la *hipervigilancia* del ciudadano y en la explotación sistemática de los datos personales de los usuarios (*monetización*) y en su control manipulativo. Este modelo fue acuñado y caracterizado magistralmente por Shoshana Zuboff, en 2019, denominándolo Capitalismo de la Vigilancia[15]. En tal modelo, el dato es el valor de referencia y la información personal diseminada en la red, una materia prima muy valiosa, para predecir y modificar y dirigir su conducta, con fines de control social o comerciales. A partir de ahí, como correlato lógico inevitable, la captación de la atención será una estrategia vital para las plataformas, luchando por mantener al usuario conectado y generando información. Como veremos a continuación, ello ha dado lugar a nuevos riesgos sociales y jurídicos, que se proyectan especialmente sobre la autonomía individual, la privacidad y la salud mental de menores y adolescentes.

---

[14] Cfr. Helberger, Natali; Karppinen, Kari; D'Acunto, Luigi. «Exposure diversity as a design principle for recommender systems». Information, Communication & Society, vol. 21, n.º 2, 2018, pp. 191-207. DOI: 10.1080/1369118X.2016.1271900. Milano, Silvia; Taddeo, Mariarosaria; Floridi, Luciano. «Recommender Systems and Their Ethical Challenges». AI & Society, vol. 35, 2020, pp. 957-967. DOI: 10.1007/s00146-020-00950-y.

[15] Zuboff, S., *La era del capitalismo de la vigilancia: La lucha por un futuro humano frente a las nuevas fronteras del poder*, . Barcelona, Paidós, 2020, para la edición española.

III.2. La IA persuasiva en el contexto de la "Economía de la Atención": la manipulación algorítmica

La IA persuasiva es una forma de inteligencia artificial que está diseñada para influir en las personas, interviniendo de forma dirigida en las decisiones, las emociones y/o en el comportamiento de los usuarios[16].

Tuvo su origen en la Web 2.0, cuando el desarrollo de las redes sociales puso al descubierto una ingente cantidad de datos y de informaciones de los ciudadanos que la IA podía procesar. Su despegue se produjo, sin embargo, a partir del momento en que la web semántica permitió hacer inteligibles y procesables todos esos datos, puesto que ya se los podía tratar y estructurar para ofrecer respuestas más precisas, contextualizadas y con un mayor grado de personalización para el usuario. A partir de ahí, se perfeccionaron los sistemas de recomendación, haciendo que el sistema no sólo pudiera recomendar un determinado producto, por ejemplo, sino adaptar dicha sugerencia al momento y a la forma concreta que resulta más propicia para el usuario, tras el análisis de los patrones conductuales y situacionales de éste, que el sistema es ya capaz de identificar. En este punto, en definitiva, la IA persuasiva evoluciona desde un modelo basado en correlaciones simples hacia otro modelo más profundo, sustentado en inferencias semánticas, con respecto al usuario concreto.

En el momento actual, la IA persuasiva es un factor clave para entender la naturaleza de una nueva clase de riesgos que deben preocupar al Derecho Penal. Estos riesgos serían los derivados de la exposición a contenidos digitales lesivos, los relacionados con la fragmentación y degradación de la atención humana, los inherentes a las adicciones digitales y, como eje articulador de todos ellos, los conectados con la manipulación algorítmica del ser humano. Y es que todos estos riesgos son la consecuencia natural del concepto, los fines y las estrategias centrales de la denominada *economía de la atención*[17], modelo que, como ya se ha anticipado, se inserta dentro en la estrategia propia del *Capitalismo de la Vigilancia* y que se centra en la captación de la atención del usuario, que se convierte en el valor más preciado por el que compiten deliberadamente las plataformas. Como es lógico, dentro de este modelo económico, la IA persuasiva juega un doble papel esencial.

Así, de una parte, es el medio o estrategia fundamental *para captar la atención* del usuario y mantenerlo conectado el mayor tiempo posible, por ejemplo, con diseños adictivos que explotan las vulnerabilidades cognitivas, emocionales y del

---

[16] Pionero en esta idea fue Fogg, J.G., *Persuasive technology: using computers to change what we think and do*, Morgan Kaufmann Publishers, Amsterdam Boston, 2003. También, proponiendo un marco básico para su diseño: Oinas-Kukkonen, H.; Harjumaa, M., «Persuasive Systems Design: Key Issues, Process Model, and System Features», *Communications of the Association for Information Systems*, vol. 24, 2009, en https://aisel.aisnet.org/cais/vol24/iss1/28.

[17] Desarrollado por Davenport, T.H.; Beck, J.C., *The attention economy: understanding the new currency of business*, Harvard Business School Press, Boston, Mass, 2001. Davenport, T.H.; Beck, J.C., *The attention economy: understanding the new currency of business*, Boston : Harvard Business School Press, 2001, en http://archive.org/details/attentioneconomy00thom. También en Davenport, T.H.; Beck, J.C., *The attention economy: understanding the new currency of business*, Boston : Harvard Business School Press, 2001, en http://archive.org/details/attentioneconomy00thom.

desarrollo del ser humano… Desde la Web 2.0, las redes sociales y los grandes servicios en línea han ido desarrollando esta lógica mediante un diseño algorítmico adictivo y técnicas de gamificación, basado en *notificaciones constantes*, *contenidos efímeros*, el botón "*like*", el "*scroll infinito*", los *sistemas de recompensa intermitente* y la *explotación de otras vulnerabilidades* particularmente de niños y adolescentes. El objetivo de estos diseños algorítmicas es siempre el mismo: capturar y retener al usuario el mayor tiempo posible, utilizando *segmentación emocional, microtargeting y personalización*, de modo que cada persona recibe narrativas a su medida, que, con el tiempo, acaba asumiendo como su propia realidad. Este fenómeno da lugar a las conocidas "*burbujas de información*", que refuerzan preferencias previas y/o inducen procesos de polarización a conveniencia de la plataforma o de sus clientes[18].

El segundo papel que juega la IA persuasiva dentro de este modelo económico es el de mecanismo de introducción del mensaje manipulador. Una vez abierta la puerta con la captación de la atención de usuario, se insertan mensajes personalizados que se introducen estratégicamente de forma sutil, casi invisible, en la experiencia digital. Es decir, la persuasión que se pretende no se trata de lograr mostrando anuncios que el usuario perciba claramente como una "cuña" publicitaria, sino que opera de forma más sutil, modelando progresivamente el contexto informativo del usuario, predisponerlo internamente, para finalmente orientarlo hacia determinadas decisiones. Por ejemplo, en YouTube, el sistema no se conforma con entregar el vídeo solicitado, sino que despliega una *cadena de recomendaciones calculadas* para inducir un consumo prolongado y, en ciertos casos, para dirigir la atención hacia contenidos específicos comerciales o ideológicos[19].

Como puede apreciarse, la IA persuasiva es un sistema de influencia programada, cuya manipulación es, podría decirse así, *de naturaleza estadística-cognitiva*, más que emocional. Su estrategia se basa en la captación de la atención, en la configuración paulatina y soterrada de las preferencias del usuario para, de esta forma, orientar el proceso de toma de decisión. Su entorno natural, además, serían las redes sociales y otros servicios digitales masivos, puesto que los algoritmos empleados en estas plataformas se nutren de la hipervigilancia sin precedentes de una multitud de usuarios. A partir de este conocimiento exhaustivo, los sistemas pueden generar respuestas altamente personalizadas, que incrementan la eficacia de los mensajes comerciales, ideológicos o políticos. Como dijera James Williams, las redes sociales funcionan máquinas globales de persuasión, a escala industrial (*industrial persuasion*)[20].

---

[18] Pariser, Eli. El filtro burbuja: Cómo la web decide lo que leemos y lo que pensamos. Madrid: Taurus, 2017, *passim*. Zuiderveen Borgesius, Frederik J. et al. «Should We Worry About Filter Bubbles?». Internet Policy Review, vol. 5, n.º 1, 2016. DOI: 10.14763/2016.1.401, entre otros.

[19] Vid, en una obra clave de Lanier, Jaron. Diez razones para borrar tus redes sociales de inmediato. Madrid: Taurus, 2018. También el documento fundamental para la comprensión de lo que sucede en las redes sociales: Orlowski, Jeff (dir.). The Social Dilemma [documentary film]. Exposure Labs / Netflix, 2020.

[20] Williams, J., *Stand out of our Light: Freedom and Resistance in the Attention Economy*, Cambridge University Press, Cambridge, 2018, en https://www.cambridge.org/core/books/stand-out-of-our-li-

Teniendo en cuenta el concepto, las finalidades y las estrategias operativas de la IA persuasiva, y las plataformas que la utilizan, podría extraerse sin dificultad el conjunto de riesgos que se plantean y que, a mi modo de ver, requieren un análisis político-criminal; es decir, una reflexión en orden a decidir si el Derecho Penal debe o no intervenir para su control y en qué términos, concretando necesidad de tutela de los bienes jurídicos subyacentes.

Como riesgo general de la IA, sea de la clase que sea, debe mencionarse, obviamente, el problema de privacidad del ser humano y de sus datos e informaciones sensibles. Los sistemas de IA persuasiva se alimentan de una recogida masiva y continua de datos personales y de comportamiento del usuario, (datos de interacción) tanto explícitos: lo que el usuario publica, busca o consume; como implícitos: tiempos de lectura y uso, pausas, movimientos que realiza, lenguaje y expresiones utilizadas, geolocalización…; y tanto directos, como indirectos o extraídos de terceros. Estos datos son procesados mediante algoritmos capaces de inferir características personales, preferencias, emociones o incluso rasgos de personalidad… Sin duda, no es preciso ahondar en el significado que ello tiene en hacer del usuario un ciudadano cada vez más transparente, con un derecho a la privacidad usurpado por el mercado[21]. Más aún, como veremos, inmediatamente a continuación, con la IA empática, la realidad virtual y las neurotecnologías propias de la Web 3.0, el desafío de la intimidad es mucho mayor, puesto afecta también a la privacidad mental, la última frontera de la intimidad del ser humano.

Aparte de ello, en el contexto específico de la IA persuasiva, deben señalarse dos fuentes de riesgo específicas: los derivados de la economía de la atención y los derivados de la finalidad misma de la IA persuasiva.

En este sentido y, como se ha anticipado, los riesgos derivados de un modelo que busca captar la atención del usuario y mantenerlo "enganchado" todo el tiempo que sea posible, explotando las vulnerabilidades del ser humano, tiene como primer riesgo el derivado de la exposición abusiva a contenidos que pueden ser lesivos, por ejemplo, porque favorezcan conductas suicidas, autolesivas, de adoctrinamiento terrorista o de trastornos de la alimentación Así se reconoce, por ejemplo, en los art. 143 bis o 156ter o 361bis, con respecto a contenidos relacionados con el suicidio, las autolesiones o los trastornos alimenticios. Así mismo, el diseño adictivo de las redes y plataformas puede producir daños a la salud mental, especialmente entre menores y adolescentes. En este sentido, se han reportado daños como la ansiedad, el estrés, dismorfia, pérdida de concentración, patrones de uso compulsivo y, lo que también es un daño social muy relevante, daños estructurales en niños y jóvenes derivados de fragmentación de la atención[22].

---

ght/3F8D7BA2C0FE3A7126A4D9B73A89415D., como obra paradigmática, particularmente Capítulo 1.

[21] Cfr., Zuboff, S., *El Capitalismo de la Vigilancia,* cit. p. 21.

[22] Más ampliamente, González Tapia, M.I, "Persuasión algorítmica y adicción tecnológica: las nuevas tabaqueras". *Revista de Victimología / Journal of Victimology*, n.º 19, 2025, pp. 43-76. DOI: 10.12827/RVJV.19.02

Además, derivado del propio su propio concepto, la IA persuasiva es un mecanismo idóneo y eficaz de influencia. Por ello, su riesgo específico es: la manipulación de los ciudadanos, cuya importancia la reconoce el propio art. 5.1. del RIA, al albergar entre las conductas directamente prohibidas, de una forma ciertamente restrictiva, conductas manipuladoras cometidas mediante sistemas IA[23]. A diferencia de la simple influencia de base informativa, la manipulación persuasiva implica ya una intervención programada, oculta, de base estadística (pero también personalizada o segmentada), que incide en los procesos cognitivos del individuo, mediante la presentación selectiva y personalizada de estímulos, en una burbuja informativa, que orienta sus decisiones sin su plena conciencia ni consentimiento.

IV. LA WEB 3.0 INMERSIVA O EXPERIENCIAL: LA CRIMINALIDAD "NEURODIGITAL"

Se suele considerar que la Web Semántica ha sido un punto de inflexión esencial para la transición de la web 2.0 a la Web 3.0, permitiendo dar un paso crucial hacia una web todavía más inteligente. El valor ya no está solo en el contenido de los datos, sino en sus conexiones y en su significado contextual, así como en la experiencia digital personalizada que se presenta al usuario. Frente a las etapas anteriores, caracterizadas por la interacción social y el intercambio de información, esta nueva fase introduce una dimensión experiencial e inmersiva que transforma la relación entre el usuario y la tecnología[24].

La IA, la Realidad Extendida (en adelante, XR, por sus siglas en inglés) y las neurotecnologías comienzan a converger, creando entornos donde la mente, la emoción y el comportamiento del individuo se integran progresivamente en el ecosistema digital. En este contexto surge el entorno y el germen de lo que, en mi opinión, podría denominarse *criminalidad neurodigital*, que podría ser entendida como el conjunto de conductas ilícitas que utilizan o afectan directamente a los procesos mentales, emocionales o neuronales del ser humano mediante sistemas inteligentes o interfaces tecnológicas. Esta nueva categoría, en cualquier, no vendría sustituir a la criminalidad digital tradicional, sino que simplemente viene a acumularse a la anterior, incorporando una dimensión cognitiva y afectiva: la integridad psíquica, la libertad de pensamiento, la autonomía cognitiva del individuo y hasta su propia identidad.

Cuando se habla de Web 3.0, no obstante, conviene subrayar que no se trata todavía de una realidad plenamente implantada, sino de un modelo en desarrollo, que se está fundamentando, en primer lugar, sobre un *deep learning* todavía

---

[23] Sobre este tema, un reciente trabajo en González Tapia, M. I. «Tutela penal frente a la manipulación algorítmica: análisis político-criminal al artículo 5.1 del Reglamento de Inteligencia Artificial», en R. Bartoli & I. Benítez Ortúzar (Dirs.), *Derecho Penal y Garantías Constitucionales*, Madrid: Dykinson, 2025, pp. 355-377.

[24] Arif Hassan, M.; Jamshidi, M. B.; Manh, B. D.; Chu, N. H.; Nguyen, C.-H.; Nguyen, Q. H.; et al. «Enabling Technologies for Web 3.0: A Comprehensive Survey». *arXiv* preprint, 2023. DOI: 10.48550/arXiv.2401.10901.

más avanzado y en el surgimiento de modelos generativos de IA, como Chatgpt, que están suponiendo una verdadera disrupción, puesto que ya son capaces de producir lenguaje, imágenes, música o entornos virtuales, completos y extremadamente realistas y de "creación propia". Además, en este modelo de red se integran otros sistemas IA, que captan y analizan los datos biométricos, conductuales y emocionales de los usuarios, lo que ha abierto paso, hablaremos a continuación de ello, a la denominada IA afectiva y a la IA empática[25]. Paralelamente, la Web 3.0 también se define por la integración de la la denominada XR, que englobaría la Realidad Aumentada (en adelante, AR), la Realidad Mixta (en adelante, MR) y la Realidad Virtual (en adelante, VR). Así mismo, en la configuración de dicho entorno se están desarrollando también otras herramientas tecnológicas, basadas en el *blockchain* y en la descentralización, que están abriendo las puertas a una nueva economía digital, basada en activos digitales (*tokens*), en una *identidad digital* única (interoperable), en organismos descentralizados (DAOs - Decentralized Autonomous Organization) y *contratos inteligentes*… y, en conexión con ello, también se están desarrollando *mundos virtuales* hacia, quien sabe, la fundación de un verdadero metaverso[26].

Como puede fácilmente intuirse a partir de esta breve descripción de la noción de la Web 3.0, los riesgos que se proyectan sobre los derechos de los ciudadanos y sobre el propio entorno institucional vigente son múltiples y evidentes. Sin ánimo exhaustivo, podría mencionarse una erosión todavía más profunda de la privacidad, puesto que la capacidad de los sistemas para perfilar conductas, emociones y datos biométricos multiplica los riesgos tradicionales; la intensificación de la manipulación emocional y cognitiva del ciudadano, que trataremos a continuación; la creación de identidades digitales falsas o alteradas, como formas avanzadas de usurpación de identidad; la difusión de información falsa o ilícita y los fenómenos de desinformación, ya que la Web 3.0 facilita la creación de entornos informativos personalizados, donde los algoritmos pueden reforzar creencias preexistentes y difundir *fake news* o *deepfakes*; nuevas formas de agresión o daño digital dirigidas contra los avatares o los activos virtuales de los usuarios; discriminación algorítmica, derivada de decisiones automatizadas en sectores como el bancario, el laboral o el acceso a servicios; y fraudes patrimoniales de nueva generación, basados en técnicas como la suplantación de voz (*deep voice*)…

En el presente análisis, sin embargo, nos centraremos exclusivamente en los riesgos derivados de la IA afectiva y empática, así como de la RV, por constituir ejemplos paradigmáticos de las nuevas formas de vulnerabilidad y manipulación

---

[25] Vid. la referencia originaria en Picard, Rosalind W. Affective Computing. Cambridge (Massachusetts): The MIT Press, 1997. DOI:10.7551/mitpress/1140.001.0001. ISBN electrónico: 978-0262281584. Para entender la transición de la IA afectiva a la empática, también Cowie, Roddy, Ellen Douglas-Cowie, Nicola Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz y John G. Taylor. "Emotion Recognition in Human-Computer Interaction." IEEE Signal Processing Magazine 18, n.º 1 (2001): 32–80. https://doi.org/10.1109/79.911197.

[26] Huynh-The, Thien, Thippa Reddy Gadekallu, Weizheng Wang, Gokul Yenduri, Pasika Ranaweera, Quoc-Viet Pham, Daniel Benevides da Costa y Madhusanka Liyanage. "Blockchain for the Metaverse: A Review." Future Generation Computer Systems 143 (2023): 401–419. https://doi.org/10.1016/j.future.2023.02.008.

que emergen en esta fase avanzada de la evolución digital, en contraste con las características propias de la etapa anterior.

IV.1. IA afectiva e IA empática: de la persuasión algorítmico-cognitiva a la emocional

La IA afectiva (*affective computing*) es un tipo de sistema inteligente diseñado para mantener una interacción emocional coherente, contextualizada y personalizada con el usuario. Se trata de un campo de estudio muy relevante, cuyo fin es permitir que los sistemas inteligentes reconozcan, sientan, infieran e interpreten las emociones humanas[27]. Para lograrlo, este tipo de sistemas no se limitan únicamente a detectar emociones en el usuario (*affect recognition*); sino que también están diseñados para generar respuestas que imitan tal naturaleza, simulando las emociones humanas (*affective responses*). Además, sus respuestas son adaptables y dinámicas, según requiera el contexto y el *feedback* que proporcione el usuario en tiempo real[28].

A partir de esta mínima conceptualización, podría identificarse una clara distinción entre la IA persuasiva y la IA afectiva. Así, a diferencia de la IA afectiva, la IA persuasiva no requiere necesariamente detectar emociones. Su núcleo funcional reside en influir sobre la atención, las preferencias y las decisiones del usuario, a partir de datos de interacción, como clics, historial de navegación, tiempos de visualización o patrones de consumo digital. Este tipo de sistemas, predominantes desde la Web 2.0, actúan sólo en el plano cognitivo, orientando lo que el usuario piensa, elige o prefiere. La IA afectiva, sin embargo, amplía ese campo de acción sumando y centrándose en la segunda dimensión de la racionalidad humana, que es, además, clave en la toma de decisiones. La IA afectiva, en definitiva, pone el foco en lo emocional, tratando de influir en el usuario, sobre todo, a través de la incidencia sobre lo que siente, teme o desea; en la experiencia emocional en el entorno digital. Esta estrategia es, a mi juicio, un salto cualitativo en cuanto al grado de intrusión en la privacidad del usuario y en cuanto a la capacidad de influencia que el sistema IA puede desarrollar sobre el ser humano. Por esta razón, la IA afectiva vendría a suponer un manifiesto incremento del riesgo de manipulación del ser humano, con una estrategia persuasiva mucho más profunda, más personalizada y, por tanto, más efectiva.

Además, debe precisarse que la IA afectiva sería una categoría general y la IA empática una forma más avanzada de ella y deben delimitarse. Así, la IA afectiva

---

[27] R. Picard la define como un tipo de computación que se relaciona, proviene o influye deliberadamente en la emoción u otros fenómenos afectivos del usuario, y en la que se los términos emoción y afecto se utilizan de forma indistinta. Así, Picard, *Affective Computing*, cit. p. 3.

También, como referencia clave, cfr. Calvo, Rafael A., Sidney D'Mello, Jonathan Gratch y Arvid Kappas. The Oxford Handbook of Affective Computing. Nueva York: Oxford University Press, 2015. https://doi.org/10.1093/oxfordhb/9780199942237.001.0001.

[28] Ma, Haiwei, y Svetlana Yarosh. "A Review of Affective Computing Research Based on Function-Component-Representation Framework." IEEE Transactions on Affective Computing 14, n.º 2 (1 de abril-junio de 2023): 1655-1674. https://doi.org/10.1109/TAFFC.2021.3104512.

en sentido general es más limitada, puesto que se orienta exclusivamente al reconocimiento emocional, es decir, a identificar indicadores de alegría, tristeza, enfado o miedo mediante el análisis de datos biométricos, faciales, vocales o fisiológicos. A partir de ahí, estaría diseñada para responder de forma coherente y adaptativa, emulando una interacción emocional básica. Sin embargo, la IA empática va mucho más allá, puesto que no sólo reconoce las emociones, sino que, además, además, responde coherentemente al contexto emocional del usuario y *simulando comprensión y conexión afectiva*, simulando empatía, a través de la modulación de la voz, la forma y el fondo del mensaje. Su finalidad, por diseño, es establecer una "relación" con el usuario: generar en el usuario la sensación de ser comprendido, de estar acompañado y de tener intimidad y confianza, creándose una experiencia de interacción íntima y personalizada. Por ejemplo, un asistente emocional *(IA companions)* como Replika está diseñado, precisamente, para dar apoyo emocional o sanitario, incidiendo directamente en la esfera afectiva del individuo; en lo más íntimo[29].

IV.2. Riesgos derivados de la IA afectiva y de la IA empática: la manipulación emocional

La expansión de la IA afectiva y empática introduce una nueva categoría de riesgos, derivados de la explotación emocional del ser humano; un fenómeno que, si se me permite, entiendo más allá de la idea de manipulación, con ser ésta radicalmente importante. El núcleo del problema radica en el significado y en las consecuencias derivadas de explotar la vulnerabilidad emocional del ser humano, su necesidad de socializar con otras personales, a través de herramientas diseñadas para modular sus afectos, estados de ánimo y sus decisiones, sin su plena conciencia ni su consentimiento informado[30].

Sin ánimo de ser exhaustiva, dado que el espacio no lo permite, un primer gran riesgo, que aparece evidente, es que este tipo de sistemas suponen un incremento evidente en erosión de la privacidad del ciudadano, afectándose su privacidad mental. Si la Web 2.0 implicó ya una exposición masiva de datos personales, la IA afectiva multiplica ese riesgo al registrar y procesar señales biométricas y emocionales, como la voz, el rostro, la respiración, el ritmo cardíaco, microexpresiones, forma de caminar o teclear… que permiten inferir sentimientos, estados psicológicos o incluso pensamientos incipientes, por ejemplo, autolíticos[31]. De este modo, la vida emocional del individuo se convierte en objeto de análisis y

---

[29] Sharkey, Amanda, y Noel Sharkey. "Children, the Elderly, and Interactive Robots." IEEE Robotics & Automation Magazine 18, n.º 1 (marzo 2011): 32–38. https://doi.org/10.1109/MRA.2010.940151.

[30] En este sentido, McStay, Andrew. "Emulated Empathy: Can Risks Be Countered by a Soft-Law Standard?" IEEE Transactions on Technology and Society 6, n.º 3 (septiembre 2025): 250-256. https://doi.org/10.1109/TTS.2025.3573875.

[31] Dewitte, Pierre. "Better Alone Than in Bad Company: Addressing the Risks of Companion Chatbots Through Data Protection by Design." Computer Law & Security Review 54 (septiembre 2024), Art. nº 106019. Disponible en línea: https://www.sciencedirect.com/science/article/abs/pii/S0267364924000852.

explotación económica, dando lugar a lo que algunos autores han denominado *neurocapitalismo*[32].

Un segundo riesgo sería el de la manipulación, que ahora es todavía más intensa, por dirigirse tanto lo cognitivo como a lo racional. Mientras la IA persuasiva tradicional influía en la atención y en las preferencias, la IA empática es capaz de alterar activamente el estado emocional del usuario para orientar su conducta, lo que hace que la manipulación sea mucho más eficaz. Además, el diseño de la IA empática genera también el peligro de generar dependencia emocional en el usuario y vínculos de confianza artificiales, distorsionados y socialmente disfuncionales. Por ejemplo, un usuario que recurre incesantemente a su *chatbot* de compañía, para obtener consejo y orientación de cualquier aspecto relevante de su vida, puede acabar desarrollando dependencia emocional o estableciendo una relación que, para el humano, es de amistad o amorosa, lo que puede reducir o desplazar otras relaciones reales.

IV.3. Convergencia disruptiva de la Inteligencia Artificial y de la Realidad Extendida: hacia una criminalidad inmersiva y sensorial

El concepto de realidad extendida[33] es un término paraguas utilizado para englobar todas aquellas formas de interacción digital que expanden o modifican la percepción del usuario de la realidad física y que divergen entre ellas por su grado de inmersión y de integración entre lo físico y lo digital. Así, la XR incluye tres grandes modalidades. En primer lugar, la realidad aumentada (AR), que consiste en la mera superposición de elementos digitales sobre la percepción del mundo físico, mediante teléfonos móviles, gafas inteligentes o visores, pero sin que el usuario esté aislado de su entorno, por ejemplo, el juego PokémonGo. En segundo, como situación intermedia, estaría la realidad mixta (MR), en la que los objetos virtuales no solo se superponen, sino que interactúan con el entorno físico y con el usuario de manera coherente, por ejemplo, Microsort Hololens. La tercera situación, con el mayor grado de inmersión en un entorno totalmente simulado sería VR, que puede entenderse como una interfaz avanzada de interacción humano-computadora, que permite al usuario sumergirse en experiencias digitales tridimensionales mediante dispositivos tecnológicos. Entre ellos destacan las pantallas y lentes para visión estereoscópica, auriculares con sonido envolvente, sensores de movimiento en cabeza y cuerpo, controladores de mano y dispositivos hápticos (guantes, chalecos o trajes) que transmiten señales vibrotáctiles. El objetivo es generar la ilusión de presencia, esto es, la vivencia subjetiva de

---

[32] Pykett, John. "Neurocapitalism and the New Neuros: Using Neuroeconomics." Journal of Economic Geography 13, n.º 5 (2013): 845–863. https://doi.org/10.1093/jeg/lbt026. Griziotti, Giorgio. Neurocapitalism: Technological Mediation and Vanishing Lines. Londres: Minor Compositions, 2019. Disponible en acceso abierto: https://www.minorcompositions.info/wp-content/uploads/2018/09/neurocapitalism-web.pdf., entre otros.

[33] Mann, Steve, Havens, John C., Iorio, Jay, Yu, Yuan, and Furness, Tom. (2018). All reality: Values, taxonomy, and continuum, for virtual, augmented, eXtended/MiXed (X), mediated (X, Y), and multimediated reality/intelligence. In Presented at the AWE 2018.

estar dentro del entorno e interactuar en primera persona mediante un avatar[34]. En la actualidad, los dispositivos más extendidos son los cascos (*head-mounted displays* -HMDs) y auriculares, que constituyen el estándar de referencia en materia de inmersión.

En un epígrafe posterior haremos referencia también a la RV, con ocasión de la clarificación del concepto de y de neurointervención. Podría aquí anticiparse que, en sentido estricto, la RV no sería una tecnología, pero que sí podría ser considerada como tal, en sentido funcional. La VR recoge datos biométricos, emocionales y conductuales que permiten inferir estados mentales y, lo más relevante, influir en ellos. Puede generar emociones intensas, activar la memoria espacial y producir cambios en la conducta del usuario[35]. Por esta razón, se utiliza en intervenciones clínicas y terapéuticas: por ejemplo, para reducir la ansiedad y el dolor…

La VR inmersiva es, además, inseparable de la IA[36]. Los entornos digitales solo alcanzan una experiencia plenamente inmersiva mediante la integración de dispositivos no invasivos (como HMDs) con sistemas de IA capaces de procesar voz, lenguaje natural, patrones de visión y movimiento, así como expresiones emocionales. Gracias a esta combinación, el entorno virtual puede reaccionar de manera dinámica, ofreciendo una experiencia altamente verosímil[37].

De entre los múltiples sistemas IA que son necesarios, en todos los niveles, para hacer posible la RV inmersiva, me gustaría destacar la íntima relación que tiene con la IA persuasiva o, si se prefiere, destacar que la RV es un medio especialmente idóneo para ejercer persuasión a través de la IA. La VR es un entorno altamente persuasivo por su capacidad de generar presencia y emoción en el usuario[38] y la IA permite lograrlo, lo que la convierte en una forma avanzada de IA persuasiva.

Además, también me gustaría destacar que dicha persuasión se ejerce fundamentalmente a través de la evocación de emociones en el usuario. Por esta razón, en mi opinión, la RV podría llegar entenderse también como un medio, un entorno propicio o, incluso, una especie de modalidad de IA afectiva[39]. La VR puede funcionar como modalidad de IA afectiva, en la medida en que los sis-

---

[34] Cfr. Bailenson, Jeremy. (2018). Experience on demand: What virtual reality is, how it works, and what it can do. 1st Ed. New York: W. W. Norton & Company, Inc.

[35] Riva, G.; Wiederhold, B. K.; y Mantovani, F. «Neuroscience of Virtual Reality: From Virtual Exposure to Embodied Medicine». Cyberpsychology, Behavior & Social Networking, vol. 22, n.º 1, 2019, pp. 82-96. DOI: 10.1089/cyber.2017.29099.gri.

[36] Huynh-The, Thien, Pham, Quoc-Viet, Pham, Xuan-Qui, Nguyen, Thanh Thi, Zhu, Han, and Kim, Dong-Seong. (2022). Artificial intelligence for the metaverse: a survey, February. Available at: http://*arxiv*.org/abs/2202.10336.

[37] Slater, M., Gonzalez-Liencres, C., Haggard, P., Vinkers, C., Gregory-Clarke, R., Jelley, S., et al. (2020). The ethics of realism in virtual and augmented reality. Front. Virtual Real. 1:1. doi: 10.3389/frvir.2020.00001

[38] Slater, M., and Sanchez-Vives, M. V. (2016). Enhancing our lives with immersive virtual reality. *Front. Robot. AI* 3:74. doi: 10.3389/frobt.2016.00074

[39] Cfr. Latoschik, Marc Erich, and Carolin Wienrich. "Coherent and Plausible Artificial Intelligence in Virtual Reality." *Frontiers in Virtual Reality* 2 (2021): 679943. https://doi.org/10.3389/frvir.2021.679943.

temas que la componen pueden detectar y responder a emociones del usuario (por voz, gestos, ritmo cardíaco o expresiones). Esta combinación hace que la VR sea, efectivamente, medio tecnológico que amplifica las capacidades persuasivas y emocionales de la IA y, por tanto, un entorno de persuasión emocional personalizada[40].

Como se ha dicho, la IA afectiva o empática marca un punto de inflexión: sus sistemas no solo responden a las emociones humanas, sino que las anticipan, modulan e incluso las inducen, configurando experiencias personalizadas que buscan maximizar la conexión emocional con el usuario. Por ello, su potencial persuasivo se vuelve mucho más profundo, como el desarrollado en redes sociales y, sobre todo, con modelos generativos como ChatGPT o Character.AI.

Ahora bien, esta *IA afectiva o empática textual*, cuando se combina con entornos inmersivos, es un modelo todavía más avanzado y profundo, puesto que se nutre del acceso a datos biométricos, emocionales, conductuales y fisiológicos, que variarán según el grado de inmersión. En estos supuestos, podríamos hablar de una incipiente *IA afectiva o empática inmersiva, de naturaleza sensorial, emocional y fisiológica*. Cuando la IA persuasiva se integra con la XR, el tipo de interacción que se establece con el usuario consiste en sumergir sensorial y emocionalmente al usuario en escenarios digitales personalizados, donde los estímulos visuales, sonoros y hápticos se combinan para generar experiencias de persuasión intensas y multisensoriales. El entorno deja de ser un medio de comunicación para convertirse en una realidad experiencial, en un ecosistema inmersivo híbrido que puede ser creado completamente o que puede ampliar la precepción de la realidad física cotidiana. En dichos espacios virtuales el usuario se ve involucrado sensorial y emocionalmente, resultando un impacto directo sobre su atención, sobre la memoria y sobre sus respuestas afectivas. Por ello, se puede decir que la RV inmersiva es una neurotecnología en sentido funcional, no invasiva y transformativa, cuya estrategia central es la inducción de emociones[41].

Considerado todo lo anterior, la misma tecnología que permite innovaciones muy positivas en educación, salud o entretenimiento, genera también riesgos para los derechos fundamentales muy importantes; riesgos que pueden exacerbarse en un futuro próximo, cuando se generalice su uso o se pongan productos terapéuticos a disposición de los consumidores. Uno de los riesgos más preocupantes es el de manipulación, cognitiva y emocional y el de la persuasión subliminal, es decir, aquella que opera, directamente, puenteando el control consciente del usuario. Además de las posibilidades de engaño que permite, la RV por su propio funcionamiento, es una tecnología especialmente idónea para el denominado *nudging*, esto es, para influenciar la toma de decisiones mediante la

---

[40] En este sentido, el texto de referencia en cuanto a los riesgos de la RV, Madary, M., y Metzinger, T. K. «Real Virtuality: A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology». *Frontiers in Robotics and* AI, vol. 3, 2016, art. 3, pp. 1-23. DOI: 10.3389/frobt.2016.00003.

[41] Cfr. más ampliamente González-Tapia, M.I., "Virtual Emotions and Criminal Law", *Frontiers in Psychology 14* (2023). https://doi.org/10.3389/fpsyg.2023.1260425.

modulación de la atención, la emoción y el contexto perceptivo del individuo[42]. Esta capacidad para moldear el comportamiento, sin recurrir a la coacción ni a la oferta directa, representa un claro riesgo para los consumidores y usuarios[43] . De otra parte, otro riesgo específico de la RV inmersiva es la cibervictimización en el ámbito de las agresiones sexuales inmersivas, que son relativamente frecuentes en la red, especialmente en mundos virtuales con interacción social libre o en modos de juego multijugador. Cuando las víctimas describen sus experiencias, expresan la idea de que dichos ataques se sienten como si fueran reales. Sin embargo, la dificultad jurídica radica, no obstante, en concretar si el "contacto virtual" puede o no ser considerado equivalente a un contacto físico, lo que depende realmente del grado de inmersión de la experiencia. Y, por supuesto, otro ámbito especialmente sensible en la RV es el de la privacidad, debido a la intensa explotación de datos biométricos, emocionales y conductuales que hace posible esta tecnología. La XR, en todas sus modalidades, requiere registrar en tiempo real una gran cantidad de información íntima del usuario para construir la experiencia inmersiva y adaptar los estímulos del entorno virtual. Según el diseño, el grado de inmersión y los dispositivos que el usuario porte, esta tecnología puede captar y procesar expresiones faciales, tono y ritmo de voz, movimientos oculares, dilatación pupilar, frecuencia cardíaca, patrones de respiración, postura corporal y microgestos, entre otros indicadores fisiológicos. A partir de estos datos, la IA puede inferir estados emocionales, niveles de atención, motivaciones e incluso intenciones futuras, lo que convierte a la XR en un espacio de observación constante del cuerpo y la mente del usuario y en fuente de una información muy precisa y valiosa sobre cómo es y qué le gusta al usuario. Con razón se ha dicho que la RV es algo así coo el "Santo Grial" de la publicidad[44]. Por último, dada su gran utilidad para el aprendizaje y el entrenamiento, se ha reportado la posibilidad de su uso con fines terroristas o del crimen organizado, para entrenar tácticas violentas o ciberataques simulados[45]…

IV.4. Convergencia disruptiva del binomio IA y neurotecnologías: los neurodelitos

La revolución de la IA se entrelaza hoy con la paralela revolución de la neurobiología y de sus tecnologías aplicadas, dando lugar a una convergencia sin pre-

---

[42] Steinert, S., and Friedrich, O. (2020). Wired emotions: ethical issues of affective brain–computer interfaces. *Sci. Eng. Ethics* 26, 351–367. doi: 10.1007/s11948-019-00087-2

[43] Cfr. Slater, M., y Sánchez-Vives, M. V. «Enhancing Our Lives with Immersive Virtual Reality». *Frontiers in Robotics and AI*, vol. 3, 2016, pp. 1-47. DOI: 10.3389/frobt.2016.00074. También, Riva, G.; Wiederhold, B. K.; Mantovani, F. «Zero Sales Resistance: The Dark Side of Big Data and Artificial Intelligence». *Cyberpsychology, Behavior & Social Networking*, 2022.

[44] Heller, B., and Bar-Zeev, A. (2021). The problems with immersive advertising: in AR/VR, nobody knows you are an Ad., *J. Online Trust Safety 1*. doi: 10.54501/jots.v1i1.21

[45] Cfr. EUROPOL. *Policing in the Metaverse: What Law Enforcement Needs to Know*. An Observatory Report from the Europol Innovation Lab. Luxemburgo: Publications Office of the European Union, 2022. DOI: 10.2813/81062.

cedentes[46]. Dentro de la evolución del contexto digital, la irrupción de las neurotecnologías debe situarse en la web 3.0, donde la interacción del ser humano con lo digital no se limita ya a las pantallas, sino que empieza a integrarse también con tecnologías inmersivas, con implantes y con neurointerfaces[47]. En este contexto puede situarse la emergente categoría de los *neurodelitos*[48], destinada a englobar aquellas conductas delictivas que tienen por objeto, medio o resultado la interferencia ilícita en los procesos neuronales humanos o en los sistemas técnicos que los registran, estimulan o modulan.

Las neurotecnologías podrían definirse como el conjunto de dispositivos, técnicas y sistemas diseñados para, directamente, registrar, monitorizar, interpretar y/o modificar la actividad del sistema nervioso y, más en concreto, la actividad cerebral. En términos generales, suelen distinguirse dos grandes tipos. Por un lado, las neurotecnologías invasivas, que requieren la implantación de electrodos o interfaces dentro de cerebro y que permiten una conexión directa con las redes neuronales, como ocurre en la estimulación cerebral profunda (DBS), utilizada para el tratamiento del Parkinson o en el implante coclear, para restaurar la función auditiva. Por otro lado, las neurotecnologías no invasivas son aquellas que operan desde el exterior del cuerpo sin intervención quirúrgica, mediante técnicas como la electroencefalografía (EEG), la estimulación magnética transcraneal (TMS) o las interfaces cerebro-computadora (BCI) basadas en realidad virtual... [49].

El salto cualitativo se produce, de una parte, por la irrupción de tecnologías que, directa o indirectamente, son capaces de registrar, procesar e interpretar la

---

[46] Yuste, R., Goering, S., Arcas, B. A. y otros. "Four ethical priorities for neurotechnologies and AI". *Neuron*, vol. 94, n.º 4, 2017, pp. 732–737. DOI: 10.1016/j.neuron.2017.05.019. Savage, N. "How AI and neuroscience drive each other forwards". Nature, vol. 571, n.º 7766, 2019, pp. S15–S17. DOI: 10.1038/d41586-019-02212-4.

[47] Cfr. Council of Europe. *The Privacy and Data Protection Implication of the Use of Neurotechnology and Neural Data from the Perspective of Convention 108*, por Eduardo Bertoni y Marcello Ienca. Estrasburgo: Consejo de Europa, 2024. Disponible en: https://globalfreedomofexpression.columbia.edu/wp-content/uploads/2024/08/Expert-Report-Neuroscience.docx.pdf.

[48] Utilizando ya este término, Demetrio Crespo, E., García de la Torre García, F. "Neurotechnologies, Human Rights and Criminal Law", en AAVV-López Silva (Ed.), *Contextualizing Neuroprotection: Latin American Perpectives on the impact of Neurotechnological Development in live and society.* Cham (Suiza), Springer Nature, 2025, (Colección Ethics of Science and Technology Assessment, vol. 54). DOI: 10.1007/978-3-031-96055-0.

[49] El concepto de neurointervención fue definido por Vincent, Nadelhoffer y McCay como toda técnica capaz de cambiar la mente de un individuo (y con ello el comportamiento), a través de la modificación de su cerebro (2020, pp. 6-7). Se trata de un término amplio, que abarca tanto los procedimientos médicos diseñados para alterar estados mentales patológicos, como aquellas prácticas tecnológicas o sociales que inciden de manera indirecta en el procesamiento neuronal y cognitivo del sujeto. Se trata, pues, de un concepto funcional y finalista, pues lo relevante no es el medio sino el resultado obtenido: la modificación de la actividad cerebral. En este sentido, una neurointervención puede ser tanto una intervención médica directa, como la administración de un fármaco o la estimulación eléctrica de áreas cerebrales; como una influencia más indirecta, como ocurre con la publicidad subliminal o ciertas técnicas de manipulación algorítmica de emociones. Cfr. Vincent, Nicole A., Thomas Nadelhoffer y Allan McCay (eds.). *Neurointerventions and the Law: Regulating Human Mental Capacity.* Oxford: Oxford University Press, 2020; y Bublitz, Jan-Christoph. "My Mind Is Mine!? Cognitive Liberty as a Legal Concept.", en *Cognitive Enhancement*, editado por Elisabeth Hildt y Andreas G. Franke, págs. 233-264. Springer, 2013.

actividad neuronal del usuario, convirtiendo también al cerebro en una nueva fuente de datos con que alimentar el ecosistema digital. De otra, por tratarse de tecnologías que, además, son capaces de generar respuestas susceptibles de estimular dicha actividad.

Tecnologías como la estimulación craneal profunda pueden activar directamente redes cerebrales, generando estímulos artificiales, que el cerebro no distinguiría claramente de un estímulo natural[50]. Pero incluso sin una intervención invasiva, en experiencias de RV, por ejemplo[51], el cerebro puede reaccionar de forma similar a como lo haría con un estímulo real. Del mismo modo, la exposición repetida a experiencias violentas o estresantes en entornos virtuales puede dejar huellas que condicionan la respuesta futura del sujeto en situaciones reales, especialmente en sujetos vulnerables y en exposiciones intensas y prolongadas. Ello conduce a concluir que, a partir del estímulo artificial que introduce la neurotecnología, es posible crear asociaciones mentales dirigidas en el usuario, idóneas para inducir recuerdos o sentimientos de apego u odio hacia personas, situaciones o símbolos.

En definitiva, las neurotecnologías poseen la capacidad de modificar de forma directa, y sobre todo eficaz, nuestra actividad cerebral: lo que pensamos, lo que nos atrae o lo que nos genera miedo, las emociones que experimentamos, los rasgos de nuestra personalidad e, incluso, nuestro comportamiento.

Como fácilmente puede concluirse, tanto las técnicas invasivas como las no invasivas constituyen un factor ambiental de enorme potencial transformador del cerebro, capaces por ello de alterar los fundamentos mismos de la identidad y la libertad del usuario. Incluso modalidades menos invasivas, como la RV, comparten este poder de incidir en los procesos cognitivos y emocionales del individuo.

En este contexto, el problema de la privacidad se mantiene y se radicaliza con el desarrollo de las neurotecnologías, debido al riesgo de que los datos neurales y mentales (*neurodatos*) de los ciudadanos puedan ser monetizados y explotados económicamente; mucho más, ante la perspectiva de que las neurotecnologías salgan del campo estrictamente médico y terapéutico y sean puestas a disposición de los consumidores. En este contexto, el ciudadano puede convertirse en *transparente*, el cerebro dejar de ser la *última frontera* de la privacidad del ser humano y la esencia misma de la persona, convertirse en una fuente de información susceptible de captación, análisis y comercialización al mejor postor[52].

---

[50] Sapolsky, Robert M. *Behave: The Biology of Humans at Our Best and Worst.* Nueva York: Penguin Press, 2017, pp. 50 y ss.

[51] Slater, Mel, y María V. Sánchez-Vives. "Enhancing Our Lives with Immersive Virtual Reality." *Nature Reviews Neuroscience* 17, n.º 10 (2016): 605–615. https://doi.org/10.1038/nrn.2016.140.

[52] Kellmeyer, Philipp. "Big Brain Data: On the Responsible Use of Brain Data from Clinical and Consumer-Directed Neurotechnological Devices." *Neuroethics* 14 (2021): 83–98. DOI: 10.1007/s12152-018-9371-x. Gordon, Emily C.; Rao, Rajesh P. N.; y Yuste, Rafael. "Ethical Considerations for the Use of Brain–Computer Interfaces." *Nature Reviews Bioengineering 1*, n.º 3 (2024): 201–213. DOI: 10.1038/s44222-023-00034-6.

Proponiendo el *habeas cogitationem* como nuevo remedio jurídico con el que proteger la libertad de pensamiento frente a las neurotecnologías, Muñoz, José M., y Juan Ángel Marinaro. "You Shall

De otra parte, en este contexto, la frontera entre el cerebro y el entorno tecnológico comienza a diluirse en formas de interacción mente-máquina que están transformando la experiencia digital en un proceso compartido, como sucede, por ejemplo, en la RV inmersiva; y que implican un proceso incipiente de hibridación del ser humano. Esta nueva interacción permite imaginar la previsible aparición de formas inéditas de criminalidad, como, por ejemplo, el *neurohacking*, esto es, el acceso no autorizado y/o manipulación de implantes neuronales y de dispositivos de estimulación cerebral[53]. Tal conducta podría causar daños físicos en la estructura o funcionalidad del cerebro del usuario o bien en el propio dispositivo. Pero, además, podría permitir modular estados mentales, invocar emociones o directamente inducir comportamientos de cualquier naturaleza, sin un consentimiento informado por parte del usuario.

En definitiva, la confluencia entre IA y neurotecnologías representa uno de los fenómenos más disruptivos de la nueva era digital, por cuanto plantea el desafío de proteger la mente humana y de preservar su identidad. Nunca antes había sido tan necesario, porque nunca antes habían existido medios capaces de captar, descodificar y manipular la mente humana; ni medios para hibridar al ser humano con implantes que modifican las capacidades humanas o restauran funcionalidades dañadas. Como fácilmente puede concluirse, la integración progresiva de la IA y las neurotecnologías en el ecosistema de la Web 3.0 abre un escenario de riesgos inéditos para los derechos fundamentales, en particular para la libertad de pensamiento, la integridad mental y la autonomía personal, pero también riesgos sociales como los derivados de un acceso no equitativo a las tecnologías.

V. DISCUSIÓN. Y… ¿HACIA LA DISTOPÍA DE LA WEB 4.0? Y EL METAVERSO?

Este trabajo ha consistido en una descripción analítica sobre la evolución del entorno digital y su impacto en el Derecho Penal contemporáneo, desde los primeros ciberdelitos hasta los nuevos riesgos derivados de la IA, la RV y las neurotecnologías. El hilo conductor ha sido la transformación progresiva del espacio digital en un entorno de interacción emocional y cognitiva entre el ser humano y los sistemas inteligentes, donde la tecnología deja de ser un mero instrumento para convertirse progresivamente en un "agente", con capacidad de influencia en los ciudadanos y que pudiera estar adquiriendo progresivamente una incipiente "individualidad".

En primer lugar, se ha expuesto la evolución de la web desde su fase inicial hasta la actual Web 3.0, que está asentándose, destacando cómo cada etapa ha modificado la relación entre usuario y tecnología: comunicación unidireccional, interacción participativa, inmersión e interacción emocional. Este recorrido permite comprender que el centro del riesgo relevante en materia penal se ha desplazado

---

Have the Thought: Habeas Cogitationem as a New Legal Remedy to Enforce Freedom of Thinking and Neurorights." *Neuroethics* 17, n.º 1 (2024): 18. https://doi.org/10.1007/s12152-024-09551-8.

[53] Ienca, Marcello & Haselager, Pim (2016). Hacking the brain: brain–computer interfacing technology and the ethics of neurosecurity. *Ethics and Information Technology* 18 (2):117-129.

(o se acumulado, realmente) de la protección del patrimonio y la información personal, a la protección de los datos neurales, la libertad de pensamiento, la libre autodeterminación y la propia naturaleza o identidad humana.

Ha sido objeto de atención especial, la evolución paralela de la IA, desde una IA operativa a una IA persuasiva, como sistema diseñado para influir en el comportamiento humano, a través del tratamiento masivo de datos conductuales y emocionales. En este punto, en el trabajo se ha tratado de mostrar y deslindar distintas clases de IA persuasiva, en atención a diseño, al tipo de datos de los que se sirve y a la estrategia utilizada para realizar la persuasión, influencia o manipulación pretendida. Así, se identifica una evolución hacia la IA afectiva o empática, capaz no solo de reconocer emociones, sino también de anticiparlas o inducirlas y cuya estrategia de influencia se basa esencialmente en la vertiente emocional de la racionalidad humana. Esta progresión amplía exponencialmente el potencial de manipulación, especialmente en contextos de interacción continua como las redes sociales o los modelos generativos conversacionales.

El tercer bloque se ha centrado en la XR y, en particular, en la RV inmersiva que constituye una forma avanzada de entorno digital persuasivo, que también está en desarrollo y en proceso de implantación social. Se ha analizado como la VR, mediada inherentemente por IA, puede entenderse también como una modalidad funcional de neurotecnología no invasiva, cuya estrategia central es la inducción de emociones, y que ello la convierte en un instrumento idóneo para la persuasión emocional personalizada; en una especie también, si se me permite la expresión, de "IA afectiva".

En los distintos ámbitos, estadios de evolución de lo digital y tecnologías referidas, se ha ido describiendo el tipo de riesgos penales específicos que van emergiendo en cada uno de ellos; particularmente los que, de forma más evidente, los distinguen de la fase tecnológica anterior. Se ha destacado la erosión progresiva de la intimidad, desde lo personal o biométrico hasta la afectación última de la privacidad mental; pero se ha insistido especialmente en la evolución en las formas y en la intensidad de la manipulación del ciudadano que hace posible la IA afectiva, particularmente la empática; las neurotecnologías y también la RV inversiva. Además, se ha llamado la atención sobre otros como el *nudging* subliminal, los atentados contra la libertad sexual en los entornos inmersivos, las nuevas modalidades de delitos patrimoniales, las *ultrafalsificaciones*…que plantean claras divergencias con las modalidades 2D o las realizadas sin la intervención clave de la IA.

No se ha podido hacer un análisis exhaustivo del impacto dogmático que esta revolución de la IA, de las neurotecnologías y de la robótica; tampoco era el objetivo perseguido. No obstante, sí se ha tratado de proporcionar una base clarificadora para poder abordar dicho problema. A lo largo de estas páginas, en cualquier caso, creo que se ha puesto de manifiesto que el nuevo entorno digital y su evolución previsible hace necesario plantear la función que el Derecho Penal, en los términos que lo conocemos, debe cumplir. También creo que aparece evidente la necesidad de incorporar nuevos bienes jurídicos, como el referido a la libertad cognitiva, y, sobre todo, de actualizar nuevas formas de ataque a bienes

jurídicos preexistentes (como la privacidad) que no han necesitado hasta ahora o no han sido susceptibles hasta ahora de la tutela penal. Probablemente, de la naturaleza de los riesgos que se han ido mostrando, también parece surgir la necesidad de repensar los planteamientos y las categorías penales clásicas como la responsabilidad de las personas jurídicas, la autoría o la imputación objetiva, el dolo/imprudencia…

El reto para el Derecho Penal, sin embargo, no será solo tipificar nuevas conductas, sino adaptarse a un entorno que dista mucho de ser para el que fue diseñado y salvaguardar no sólo los bienes jurídicos esenciales para la convivencia, sino la propia naturaleza del ser humano, con su atributo esencial que es la libertad de autodeterminación.

Mirando hacia el futuro, la evolución hacia la Web 4.0 o Web Simbiótica, se abre un horizonte todavía más complejo, desafiante y tal vez distópico. Este nuevo ecosistema digital que ya se apunta como posible, se dirigiría hacia una integración casi orgánica entre humanos y sistemas inteligentes; es decir, hacia la hibridación del ser humano, donde la frontera entre el ser humano y la máquina se diluiría. A diferencia de las etapas anteriores, la Web 4.0 se caracterizaría por la conexión directa entre el cerebro y las tecnologías digitales mediante neurointerfaces, implantes corticales o *neurochips*, capaces de registrar y modificar la actividad neuronal, de modularla, permitiendo mejorar al ser humano o, incluso, darle capacidades metahumanas…En este contexto, la convergencia entre IA y neurotecnología no solo interpretaría los estados mentales, sino que podría llegar a escribir en ellos, transformando la experiencia humana en un proceso compartido con la máquina. Se trataría, en definitiva, de un salto cualitativo y estructural sin precedentes. Tal vez, además, en este nuevo entorno también exista un Metaverso, en el sentido teorizado por Matthew Ball[54], que supondría realmente un cambio de paradigma para el Derecho Penal, en tanto que un nuevo contrato social. Quién sabe si para entonces, alguno de los riesgos y de las necesidades apuntadas en este trabajo… seguirán siendo tales.

VI. REFERENCIAS

ROMEO CASABONA, C. M. (Coord.). *El cibercrimen: nuevos retos jurídico-penales, nuevas respuestas político-criminales*. Granada: Comares, 2006.

AGUSTINA, José R. Cibercriminalidad y perspectiva victimológica: un enfoque general explicativo de la cibervictimización, *Cuadernos de Política Criminal* 114 (III), 2014, pp. 143–178.

- Nuevos retos dogmáticos ante la cibercriminalidad. *Estudios Penales y Criminológicos* 41, 2021, pp. 705–777. Disponible en: https://revistas.usc.gal/index.php/epc/article/view/7433 (consulta: 10-10-2025).

ARIF HASSAN, M., JAMSHIDI, M. B., MANH, B. D., CHU, N. H., NGUYEN, C.-H., NGUYEN, Q. H., *et al.* Enabling Technologies for Web 3.0: A Comprehensive Survey. *arXiv preprint*, 2023. https://doi.org/10.48550/arXiv.2401.10901.

[54] Ball, Matthew. The Metaverse: And How It Will Revolutionize Everything. Liveright Publishing Corporation, 2022.

BAILENSON, J. *Experience on Demand: What Virtual Reality Is, How It Works, and What It Can Do.* Nueva York: W. W. Norton & Company, 2018.

BERNERS-LEE, T., HENDLER, J. y LASSILA, O. The Semantic Web: A New Form of Web Content That Is Meaningful to Computers Will Unleash a Revolution of New Possibilities. *Scientific American* 284, n. º 5, 2001, pp. 34–43. https://doi.org/10.1038/scientificamerican0501-34.

BUBLITZ, J.C. My Mind Is Mine!? Cognitive Liberty as a Legal Concept. En *Cognitive Enhancement*, editado por Elisabeth Hildt y Andreas G. Franke, Springer, 2013, pp. 233–264.

CALVO, R., D'MELLO, S., GRATCH, J. y KAPPAS, A. *The Oxford Handbook of Affective Computing.* Nueva York: Oxford University Press, 2015. https://doi.org/10.1093/oxfordhb/9780199942237.001.0001.

CASTELLS, M. *La era de la información: economía, sociedad y cultura. Volumen I: La sociedad red.* Madrid: Alianza Editorial, 2006.

COUNCIL OF EUROPE. *Budapest Convention on Cybercrime.* European Treaty Series n.º 185, 2001. Disponible en: https://rm.coe.int/1680081561.

-*The Privacy and Data Protection Implication of the Use of Neurotechnology and Neural Data from the Perspective of Convention 108*, por Eduardo Bertoni y Marcello Ienca. Estrasburgo: Consejo de Europa, 2024. Disponible en: https://globalfreedomofexpression.columbia.edu/wp-content/uploads/2024/08/Expert-Report-Neuroscience.docx.pdf.

COWIE, R., DOUGLAS-COWIE, E., TSAPATSOULIS, N., VOTSIS, G., KOLLIAS, S., FELLENZ, W. y TAYLOR J.G. Emotion Recognition in Human-Computer Interaction. *IEEE Signal Processing Magazine* 18, nº 1, 2001, pp. 32–80. https://doi.org/10.1109/79.911197.

DAVENPORT, T.H., y BECK, J.C. *The Attention Economy: Understanding the New Currency of Business.* Boston (Massachusetts): Harvard Business School Press, 2001.

DEMETRIO CRESPO, E. y GARCÍA DE LA TORRE GARCÍA, F., *Neurotechnologies, Human Rights and Criminal Law, en Contextualizing Neuroprotection: Latin American Perspectives on the Impact of Neurotechnological Development in Life and Society*, AA.VV.–López-Silva (Ed.). Cham (Suiza): Springer Nature, 2025. (Colección *Ethics of Science and Technology Assessment*, vol. 54). https://doi.org/10.1007/978-3-031-96055-0.

DEWITTE, P. Better Alone Than in Bad Company: Addressing the Risks of Companion Chatbots Through Data Protection by Design. *Computer Law & Security Review,* 54 (septiembre 2024), Art. n.º 106019. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0267364924000852.

EUROPOL. *Policing in the Metaverse: What Law Enforcement Needs to Know. An Observatory Report from the Europol Innovation Lab.* Luxemburgo: Publications Office of the European Union, 2022. https://doi.org/10.2813/81062.

FERNÁNDEZ TERUELO, J.G. *Cibercrimen. Los delitos cometidos a través de internet – estafas, distribución de pornografía infantil, atentados contra la propiedad intelectual, daños informáticos, delitos contra la intimidad y otros delitos en la Red.* Oviedo: Constitutio Criminalis Carolina, 2007.

FOGG, B. J. *Persuasive Technology: Using Computers to Change What We Think and Do.* Ámsterdam–Boston: Morgan Kaufmann, 2003.

GONZÁLEZ RUS, J.J., Precisiones conceptuales y político-criminales sobre la intervención penal en internet. En AAVV Cuadernos penales José María Lidón. Nº 4: Delito e informática: algunos aspectos, Publicaciones de la Universidad de Deusto, Bilbao, 2007.

GONZÁLEZ TAPIA, M.I. Persuasión algorítmica y adicción tecnológica: las nuevas tabaquera. *Revista de Victimología / Journal of Victimology* 19, 2025, pp. 43–76. https://doi.org/10.12827/RVJV.19.02.

-Tutela penal frente a la manipulación algorítmica: análisis político-criminal al artículo 5.1 del Reglamento de Inteligencia Artificial. En R. Bartoli e I. Benítez Ortúzar (Dirs.), *Derecho Penal y Garantías Constitucionales*, Madrid: Dykinson, 2025, pp. 355–377.

- Virtual Emotions and Criminal Law. *Frontiers in Psychology* 14, 2023. https://doi.org/10.3389/fpsyg.2023.1260425.

GORDON, E., RAO, C., RAJESH, P.N. y YUSTE, R. Ethical Considerations for the Use of Brain–Computer Interfaces. *Nature Reviews Bioengineering* 1, nº 3, 2024, pp. 201–213. https://doi.org/10.1038/s44222-023-00034-6.

GRIZIOTTI, G. *Neurocapitalism: Technological Mediation and Vanishing Lines*. Londres: Minor Compositions, 2019. Disponible en acceso abierto: https://www.minorcompositions.info/wp-content/uploads/2018/09/neurocapitalism-web.pdf.

HELBERGER, N., KARPPINEN, K. y D'ACUNTO, L. Exposure Diversity as a Design Principle for Recommender Systems. *Information, Communication & Society* 21, nº 2, 2018, pp. 191–207. https://doi.org/10.1080/1369118X.2016.1271900.

HELLER, B., y BAR-ZEEV, A. The Problems with Immersive Advertising: In AR/VR, Nobody Knows You Are an Ad. *Journal of Online Trust and Safety* 1, 2021. https://doi.org/10.54501/jots.v1i1.21.

HUYNH-THE, T., REDDY GADEKALLU, T., WANG, W., YENDURI, G., RANAWEERA, P., PHAM, Q.V., BENEVIDES DA COSTA, D. y LIYANAGE, M. Blockchain for the Metaverse: A Review. *Future Generation Computer Systems* 143, 2023, pp. 401–419. https://doi.org/10.1016/j.future.2023.02.008.

HUYNH-THE, T., PHAM, Q.V., PHAM, X.Q, NGUYEN, T.T., ZHU, H. y KIM, D.S. Artificial Intelligence for the Metaverse: A Survey. *arXiv preprint* (febrero 2022). Disponible en: http://arxiv.org/abs/2202.10336.

IENCA, M., y HASELAGER, P. Hacking the Brain: Brain–Computer Interfacing Technology and the Ethics of Neurosecurity. *Ethics and Information Technology* 18, nº 2, 2016, pp. 117–129.

KELLMEYER, P. Big Brain Data: On the Responsible Use of Brain Data from Clinical and Consumer-Directed Neurotechnological Devices. *Neuroethics* 14, 2021, pp. 83–98. https://doi.org/10.1007/s12152-018-9371-x.

KUHLEN, L. Bienes jurídicos y nuevos tipos de delito. En AA.VV.–Robles Planas, Ricardo (ed. esp.), *Límites al Derecho Penal: Principios operativos en la fundamentación del castigo*. Barcelona: Atelier, 2013, pp. 225–235.

LANIER, J. *Diez razones para borrar tus redes sociales de inmediato*. Madrid: Taurus, 2018.

LATOSCHIK, M.E. y WIENRICH, C. Coherent and Plausible Artificial Intelligence in Virtual Reality. *Frontiers in Virtual Reality* 2, 2021, 679943. https://doi.org/10.3389/frvir.2021.679943.

HAIWEI, M. y YAROSH, S. A Review of Affective Computing Research Based on Function-Component-Representation Framework. *IEEE Transactions on Affective Computing* 14, nº 2 (abril–junio 2023), pp. 1655–1674. https://doi.org/10.1109/TAFFC.2021.3104512.

MADARY, M. y METZINGE, T.K. Real Virtuality: A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology. *Frontiers in Robotics and AI* 3, Art. 3, 2017, pp. 1–23. https://doi.org/10.3389/frobt.2016.00003.

MANN, S., HAVENS, J.C., IORIO, Y., JAY, Y. y FURNESS, T. All Reality: Values, Taxonomy, and Continuum, for Virtual, Augmented, eXtended/MiXed (X), Mediated (X, Y), and Multimediated Reality/Intelligence. Ponencia presentada en AWE 2018.

MCSTAY, A. Emulated Empathy: Can Risks Be Countered by a Soft-Law Standard? *IEEE Transactions on Technology and Society* 6, n.º 3 (septiembre 2025), pp. 250–256. https://doi.org/10.1109/TTS.2025.3573875.

MILANO, S., MARIAROSARIA, T. y FLORIDI, L. Recommender Systems and Their Ethical Challenges, *AI & Society* 35, 2020, pp. 957–967. https://doi.org/10.1007/s00146-020-00950-y.

MIRÓ LLINARES, F. *El cibercrimen: fenomenología y criminología de la delincuencia en el ciberespacio.* Madrid: Marcial Pons, 2012.

MUÑOZ, J.M. y MARINARO, J. A. You Shall Have the Thought: *Habeas Cogitationem* as a New Legal Remedy to Enforce Freedom of Thinking and Neurorights, *Neuroethics* 17, nº 1, 2024, pp. 18. https://doi.org/10.1007/s12152-024-09551-8.

OINAS-KUKKONEN, H. y HARJUMAA, M. Persuasive Systems Design: Key Issues, Process Model, and System Features. *Communications of the Association for Information Systems* 24, 2009. Disponible en: https://aisel.aisnet.org/cais/vol24/iss1/28.

ORLOWSKI, J. (dir.). *The Social Dilemma* [documental]. Exposure Labs / Netflix, 2020.

PARISER, E. *El filtro burbuja: Cómo la web decide lo que leemos y lo que pensamos.* Madrid: Taurus, 2017.

PICARD, R.W. *Affective Computing.* Cambridge (Massachusetts): The MIT Press, 1997. https://doi.org/10.7551/mitpress/1140.001.0001. ISBN electrónico: 978-0262281584.

PYKETT, J. Neurocapitalism and the New Neuros: Using Neuroeconomics, *Journal of Economic Geography* 13, nº 5, 2013, pp. 845–863. https://doi.org/10.1093/jeg/lbt026.

RIVA, G., WIEDERHOLD, B.K. y MANTOVANI, F. Neuroscience of Virtual Reality: From Virtual Exposure to Embodied Medicine. *Cyberpsychology, Behavior & Social Networking* 22, nº 1, 2019, pp. 82–96. https://doi.org/10.1089/cyber.2017.29099.gri.

- Sales Resistance: The Dark Side of Big Data and Artificial Intelligence. *Cyberpsychology, Behavior & Social Networking*, 2022.

ROMEO CASABONA, C.M. De los delitos informáticos al cibercrimen. Una aproximación conceptual y político-criminal. En AA.VV.–Romeo Casabona, C. M. (Coord.), *El cibercrimen: nuevos retos jurídico-penales, nuevas respuestas político-criminales*, Granada: Comares, 2006, pp. 1-42.

SAVAGE, N. How AI and Neuroscience Drive Each Other Forwards. *Nature* 571, nº 7766, 2019, S15–S17. https://doi.org/10.1038/d41586-019-02212-4.

SHARKEY, A. y SHARKEY, N. Children, the Elderly, and Interactive Robots. *IEEE Robotics & Automation Magazine* 18, nº 1 (marzo 2011), pp. 32–38. https://doi.org/10.1109/MRA.2010.940151.

SLATER, M., GONZALEZ-LIENCRES, C., HAGGARd, P., VINKERS, C., GREGORY-CLARKE, R., JELLEY S., et al. The Ethics of Realism in Virtual and Augmented Reality. *Frontiers in Virtual Reality* 1, 2020, p. 1. https://doi.org/10.3389/frvir.2020.00001.

SLATER, M., y SÁNCHEZ-VIVES, M.V. Enhancing Our Lives with Immersive Virtual Reality. *Frontiers in Robotics and AI* 3, 2016, p. 74. https://doi.org/10.3389/frobt.2016.00074.

- Enhancing Our Lives with Immersive Virtual Reality" *Nature Reviews Neuroscience* 17, nº 10, 2016, pp. 605–615. https://doi.org/10.1038/nrn.2016.140.

SAPOLSKY, R.M. *Behave: The Biology of Humans at Our Best and Worst.* Nueva York: Penguin Press, 2017, pp. 50 y ss.

STEINERT, S. y OLIVER, F. Wired Emotions: Ethical Issues of Affective Brain–Computer Interfaces. *Science and Engineering Ethics* 26, 2020, pp. 351–367. https://doi.org/10.1007/s11948-019-00087-2.

VINCENT, N.A., NADELHOFFER, T. y MCCAY, A. (eds.). *Neurointerventions and the Law: Regulating Human Mental Capacity.* Oxford: Oxford University Press, 2020.

WILLIAMS, J. *Stand Out of Our Light: Freedom and Resistance in the Attention Economy*. Cambridge: Cambridge University Press, 2018. Disponible en: https://www.cambridge.org/core/books/stand-out-of-our-light/3F8D7BA2C0FE3A7126A4D9B73A89415D.

YUSTE, R., GOERING, S., AGÜER, ARCAS, B. *et al.* Four Ethical Priorities for Neurotechnologies and AI. *Neuron* 94, nº 4, 2017, pp. 732–737. https://doi.org/10.1016/j.neuron.2017.05.019.

ZUIDERVEEN BORGESIUS, F.J., et al. Should We Worry About Filter Bubbles? *Internet Policy Review* 5, nº 1, 2016. https://doi.org/10.14763/2016.1.401.

ZUBOFf, S. *La era del capitalismo de la vigilancia: La lucha por un futuro humano frente a las nuevas fronteras del poder.* Barcelona: Paidós, 2020