

APROXIMACIÓN A LOS PATRONES OSCUROS DIGITALES (Y A LA RESPONSABILIDAD PENAL CORPORATIVA POR LOS MISMOS)

Fernando Miró Llinares

ÍNDICE

1. APROXIMACIÓN A LOS PATRONES OSCUROS DIGITALES	2
2. MÁS ALLÁ DE LA PERSUASIÓN: RASGOS ESTRUCTURALES, DAÑOS Y CONTEXTOS EMPRESARIALES DE LOS POD	7
<i>2.1. De la arquitectura de la decisión al patrón oscuro digital: tipologías y rasgos estructurales</i>	<i>7</i>
<i>2.2. Plataformas, entornos y servicios con POD.....</i>	<i>14</i>
<i>2.3. El daño de los POD: más allá del perjuicio económico, la erosión de la autonomía individual</i>	<i>17</i>
3. RESPONSABILIDAD PENAL EMPRESARIAL Y POD	21
<i>3.1. Preámbulo: cultura organizativa empresarial y economía del patrón oscuro digital</i>	<i>21</i>
<i>3.2. Marco regulatorio europeo administrativo y patrones oscuros: hacia un estándar de diligencia corporativa ex ante.....</i>	<i>23</i>
<i>3.3. Posibles delitos cometidos en el seno de la empresa y umbral penal de la manipulación</i>	<i>27</i>
a) La estafa y el engaño por diseño	28
b) Publicidad engañosa y engaño estructural a consumidores	29
c) Manipulación y abuso de mercado y gamificación del riesgo financiero	32
d) La responsabilidad por estos y otros delitos de la persona jurídica.....	34
<i>3.4. Programas de cumplimiento y prevención de los POD lesivos</i>	<i>35</i>
BIBLIOGRAFÍA	38

APROXIMACIÓN A LOS PATRONES OSCUROS DIGITALES (Y A LA RESPONSABILIDAD PENAL CORPORATIVA POR LOS MISMOS)

*Fernando Miró Llinares **

1. APROXIMACIÓN A LOS PATRONES OSCUROS DIGITALES

En los últimos años, una parte creciente de nuestra vida cotidiana se desarrolla en ciberlugares¹ en los que, de forma casi inadvertida, nuestras decisiones son orientadas, moduladas o incluso inducidas por el diseño de interfaces digitales. No se trata de un fenómeno marginal ni accidental, sino de una transformación profunda que afecta a individuos y a mercados: la conversión del diseño de las interacciones digitales en una poderosísima tecnología de poder blando que combina psicología, datos y algoritmos para dirigir la conducta de las personas con efectos individuales y masivos. Muchas de estas configuraciones están basadas en técnicas precedentes a la digitalización que tienen que ver con el denominado diseño persuasivo, técnicas concebidas para cambiar actitudes o comportamientos que hunden sus raíces en la psicología de la persuasión (Cialdini, 1984, 2007) y en la arquitectura de elección popularizada recientemente por Thaler y Sunstein (2009). Este diseño persuasivo opera regulando de manera sistemática el contexto en el que las personas toman decisiones, actuando menos sobre los contenidos y más sobre la forma en que se presentan las opciones: el orden en que aparecen, la facilidad o dificultad de acceso a cada una de ellas, los elementos visuales que atraen la atención o las señales sociales que inducen conformidad, etc. Desde hace décadas, el marketing y la publicidad han recurrido

* Catedrático de Derecho penal y Criminología de la Universidad Miguel Hernández. Director del Centro CRÍMINA para el estudio y prevención de la delincuencia.

⁰ El presente trabajo se ha realizado en el marco del Proyecto GamerVictim (CIPROM/2022/33) financiado por el programa PROMETEO 2023 de la Consellería de Innovación, Universidades, Ciencia y Sociedad Digital de la Generalitat Valenciana. Quiero dar las gracias a Aiala Tejada García de Garayo y a Mario Santisteban Galarza por su ayuda, en particular por sus interesantes aportaciones y discusiones sobre estos temas.

¹ Sobre el concepto de ciberlugar y sus implicaciones criminológicas, muy relevantes en el tema aquí tratado, véase Miró Llinares y Johnson (2018).

a estas técnicas para incrementar la conversión, orientar el consumo o fidelizar al cliente, transformando la experiencia de decisión en un escenario cuidadosamente diseñado para favorecer determinados comportamientos. Y dentro de esas prácticas, de ese continuo de la persuasión existen algunas que llegan a manipular la arquitectura misma de la elección, cruzando una frontera material al afectar directamente a la autonomía y agencia individual: se trata de diseños que inducen decisiones contrarias a los intereses o a las preferencias informadas del usuario, explotando sus sesgos cognitivos (como la urgencia, la escasez, la prueba social o la aversión a la pérdida (Cialdini, 1984, 2007)) y creando asimetrías de información en las plataformas del entorno digital. Aunque muchos de estos mecanismos y prácticas eran visibles antes fuera de los contextos digitales, por ejemplo, y de forma muy evidente, en entornos de juegos de azar y apuestas, donde el propio diseño incrementa la frecuencia e impulsividad de la conducta, ha sido la digitalización la que ha amplificado exponencialmente estas posibilidades de manipulación. Lo ha hecho al proporcionar, por un lado, un conocimiento granular y en tiempo real de los patrones de conducta y, por otro, capacidad para personalizar la arquitectura y escalar la influencia mediante segmentación algorítmica. Este funcionamiento circular entre interacción, extracción de datos y personalización, descrito en la literatura bajo el concepto de *dataveillance*² es lo que ha dado lugar a la categoría contemporánea de “*dark patterns*”: configuraciones de interfaz digital “diseñadas para subvertir o distorsionar la autonomía del usuario” (Gray et al., 2018a).

El término “*dark patterns*” fue acuñado por Brignull en 2010 para describir las tácticas engañosas que algunas páginas utilizaban para forzar suscripciones no deseadas o dificultar la cancelación de servicios³. Lo que en un primer momento pudo parecer un fenómeno aislado pronto mostró ser algo mucho más extendido. La identificación de familias estables de patrones manipulativos en distintos entornos como videojuegos, redes sociales, plataformas de inversión, entre otros, puso de manifiesto que todas estas prácticas respondían a una misma lógica: emplear el diseño de la interfaz como

² Entendido como la vigilancia continua y automatizada de la conducta del usuario a través de cada gesto digital, y su reutilización para ajustar dinámicamente los estímulos que recibe. Véase Büchi et al. (2022); y, también, específicamente en el ámbito de la publicidad: Strycharz et al. (2025).

³ La definición inicial de Brignull (2010), en una web, es enormemente sencilla pero el término ha acabado teniendo mucho éxito. Concretamente su frase es “*Deceptive patterns (also known as “dark patterns”) are tricks used in websites and apps that make you do things that you didn’t mean to, like buying or signing up for something.*”. Brignull H (2010) Véase, también, a efectos de conceptualización de Kollmer & Eckhardt (2022).

herramienta de manipulación para maximizar la conversión y los beneficios, sacrificando la transparencia y, lo que a mi juicio es aún más grave, la autonomía del usuario (véase Gray et al., 2018). En los últimos años, además, los primeros estudios han mostrado no solo la expansión de estos mecanismos⁴, sino también sus variadas tipologías (véase las clasificaciones de Gray et al., 2018, 2024), algunos de los rasgos concretos que diferencian la manipulación de la mera persuasión⁵ y variadas evidencias de su capacidad manipulativa por medio del control sutil del contexto de decisión más que por medio de la mentira explícita⁶. Y estas evidencias tuvieron su traslado a la respuesta regulatoria a estas prácticas, que pasó de, si no ignorarlas sí tratarlas como simples desviaciones del consumo, a concebirlas como un riesgo estructural para la autonomía de las personas en el entorno digital primero con las Guidelines 03/2022 del European Data Protection Board (2022) que identificaron de manera explícita los *deceptive design patterns* en redes sociales y después con la Digital Services Act (DSA) (European Parliament and Council, 2022)⁷ y la Digital Markets Act (European Parliament and Council, 2022) y culminando con la aprobación de la AI Act (European Parliament and Council, 2024) que considera como riesgo inaceptable determinadas formas de manipulación digital⁸.

El presente trabajo afronta el estudio de los “*dark patterns*”, “*dark comercial patterns*” y “*deceptive design patterns*”⁹ bajo el término analítico

⁴ El estudio de Arvind Narayanan y su equipo en Princeton, *Dark Patterns at Scale*, rastreó decenas de miles de sitios web y demostró que miles de comercios electrónicos utilizaban estos mecanismos de manera rutinaria, muchos de ellos proporcionados incluso por empresas especializadas que los ofrecían como servicio (Mathur et al., 2019).

⁵ Por ejemplo, autoras como Karen Yeung (2017) han explicado, además, cómo la personalización basada en big data “reconfigura” la arquitectura de elección hacia hiperempujones (*hypernudges*) capaces de moldear conductas de modo opaco y dinámico.

⁶ Así, por ejemplo Luguri y Strahilevitz (2021) proporcionaron, mediante experimentos controlados, la primera evidencia empírica del poder persuasivo de los patrones oscuros: los participantes expuestos a versiones “agresivas” de estas interfaces tomaban decisiones claramente contrarias a su interés, y lo hacían sin advertirlo.

⁷ La Digital Markets Act (DMA) limitó a las grandes plataformas el uso cruzado de datos y otras prácticas que amplifican esas asimetrías de poder informacional.

⁸ Todo ello dentro de un marco internacional en el que la Federal Trade Commission (FTC) en su informe *Bringing Dark Patterns to Light* (2022) y la OCDE en *Dark Commercial Patterns* (2022) establecieron un lenguaje común sobre la manipulación comercial digital, unificando categorías y terminología.

⁹ Todas ellas denominaciones que son frecuentemente usadas en la literatura como sinónimos. Creo que la clarificación conceptual en este campo resulta especialmente necesario a la luz del panorama descrito por Chang et al. (2024), cuya revisión sistemática muestra que casi la mitad de la literatura empírica sobre patrones oscuros carece de un marco teórico explícito o utiliza la teoría de forma meramente nominal.

de “patrones oscuros digitales” (en adelante, POD) que englobaría todas aquellas arquitecturas de diseño para la interacción y la decisión mediadas por datos que, aprovechando sesgos cognitivos, asimetrías de información o barreras de elección, comprometen la autonomía del usuario induciendo a este a comportamientos contrarios a sus intereses o a su voluntad informada, causando daños individuales y/o colectivos¹⁰. La preferencia por el término “oscuro” frente a “persuasivo” quiere dejar fuera las arquitecturas persuasivos transparentes y trazables, que pueden resultar en parte manipulativas y perjudiciales, pero sobre las que es posible, al menos ex ante, que la autonomía del individuo se imponga, y situar el centro de atención en los diseños manipulativos que lo impiden o lo complican sobremanera.

¹⁰ Esta definición es similar a otras que están comenzando a ser usadas como la de Kollmer y Eckhardt (2023) que los definen como “elementos de diseño de la interfaz de usuario que comprometen la autonomía del usuario al impedir decisiones informadas y que pueden conducir a resultados adversos para el usuario, como la invasión de la privacidad, la pérdida financiera y la adicción a la tecnología.”. Desde la primera utilización del término “*dark patterns*” por Brignull en 2010 como “trucos utilizados en sitios web y aplicaciones que te hacen hacer cosas que no tenías intención de hacer, como comprar o suscribirte a algo” el estudio y la conceptualización de los POD ha cambiado mucho. Para una revisión de las principales definiciones y de las claves definitorias de este concepto véase el interesante trabajo de Mathur et al. (2019) en el que hallan hasta 19 definiciones sobre *dark patterns* y extraen cuatro dimensiones fundamentales que se encuentran en las definiciones de *dark pattern* y que también están en la que yo he aportado: el tipo de características de la interfaz, los mecanismos mediante los cuales esa interfaz interfiere en el comportamiento del usuario, o la intención del diseñador que configura estas prácticas y, finalmente, los beneficios que obtiene el servicio y/o los perjuicios que sufre el usuario como resultado. Una cuestión relevante es si el patrón oscuro viene definido o no por la explotación de sesgos cognitivos o por la intención manipulativa del diseñador. A mi parecer sí en el primer caso y no, aunque normalmente así será, en el segundo. En una línea distinta, Isola y Esposito (2025) sostienen, creo que con razón en esa parte, que aquello que hace “oscuro” a un patrón no es ni la explotación de sesgos cognitivos ni la intención manipulativa del diseñador, sino su capacidad objetiva para distorsionar el comportamiento del usuario a través del diseño del entorno digital de elección. Los autores subrayan que la clave está en que el patrón opere como arquitectura de elección capaz de alterar la decisión del usuario, generando un entorno que “contamina” el proceso decisorio, en términos muy próximos a las categorías tradicionales de la Directiva sobre Prácticas Comerciales Desleales. De hecho, lo interesante de su aportación es, precisamente, desplazar el foco desde la psicología del usuario o las intenciones del diseñador hacia un criterio estrictamente jurídico-funcional: un patrón es oscuro cuando configura la interfaz de modo que sea apta para engañar, omitir información relevante, hostigar o ejercer influencia indebida, con independencia de si se apoya o no en heurísticos, y sin que sea necesario demostrar propósito manipulativo. Desde esta óptica, lo “oscuro” reside en la capacidad estructural del diseño para producir una distorsión comercialmente relevante, no en las motivaciones subjetivas de quien lo implementa. Lo que a mi parecer es difícil sostener, y de hecho en eso no entran los autores, es que sea posible un diseño con capacidad estructural para distorsionar la voluntad y que ello no sea conocido e incluso querido por el diseñador. También creo que la explotación de sesgos cognitivos si debe ser un elemento definitorio del patrón oscuro digital en cuanto que es ello lo que genera una capacidad potencial no ya de engaño sino de negación arquitectónica de la autonomía individual.

Por otra parte, la adición del término “digitales” a la traducción más literal sirve para subrayar la dimensión estructural de los patrones oscuros y su absoluta dependencia del ecosistema de datos que los hace posibles. Al apoyarse estas técnicas en un ecosistema digital que permite el uso de grandes cantidades de datos y, especialmente gracias a la IA, la interacción automatizada con los usuarios mediante pruebas iterativas, el “error” o “descuido” previsible del usuario pasa a ser el modelo de negocio y la forma de monetización del producto digital, convirtiéndose en una de las últimas expresiones del modo en que el capitalismo contemporáneo transforma el conocimiento sobre la conducta humana en infraestructura económica. Lo que Zuboff (2015, 2019) denominó “capitalismo de vigilancia”¹¹ ha evolucionado hacia una fase en la que el objetivo ya no es sólo prever lo que haremos, sino producir las condiciones para que lo hagamos. Las plataformas han pasado de observar a configurar entornos de decisión diseñados para orientar la conducta, de modo que la predicción se transforma en una forma de determinación de la conducta. Del “vigilar para prever” se ha pasado al “diseñar para inducir”, en un modelo que convierte la vulnerabilidad humana en un recurso estable de rentabilidad (Couldry y Mejias, 2019)¹² y que obliga a todos los actores del modelo económico a jugar con sus mismas reglas (Zhang et al., 2025)¹³. Esto último, el hecho de que los POD no constituyan sólo un problema de diseño ético individual, sino una forma de racionalidad económica y empresarial, es lo que da sentido al propósito de este trabajo de analizar la cuestión

¹¹ Que vendría a ser un sistema basado en la captura de los rastros de la conducta humana para convertirlos en valor económico mediante su predicción y control (Zuboff, 2015, 2019).

¹² Mediante experimentos continuos, segmentación algorítmica y optimización iterativa, las plataformas diseñan la experiencia del usuario para maximizar la conversión o la permanencia (Couldry y Mejias, 2019). Como señala Julie Cohen (2012) esta transformación se apoya en una arquitectura informacional que define lo visible y lo elegible, y convierte la infraestructura tecnológica en un medio de gobierno. A su vez, los *hiper-empujones* descritos por Karen Yeung (2017) ilustran cómo los sistemas de influencia automatizada ajustan en tiempo real los estímulos que guían nuestras elecciones. En este marco, los POD son la manifestación práctica de esa mutación: el punto donde la observación se convierte en manipulación estructural.

¹³ Que el modelo económico digital de las plataformas se basa en la creación de entornos que favorecen la manipulación y que obliga a quienes quieran participar en él a reproducir esos mismos patrones, puede observarse claramente en una investigación empírica reciente sobre plataformas de videojuegos generados por usuarios. El estudio muestra que las *funciones y herramientas de diseño incorporadas por la propia plataforma* facilitan la integración sistemática de mecanismos monetarios, psicológicos y sociales en miles de juegos distintos (Zhang et al., 2025). La reiteración de estos patrones en obras creadas por desarrolladores independientes sería un indicio más de que el diseño manipulativo no responde a decisiones particulares de cada creador, sino a los incentivos estructurales que impone la plataforma.

de la responsabilidad penal corporativa por el uso de estas técnicas.

2. MÁS ALLÁ DE LA PERSUASIÓN: RASGOS ESTRUCTURALES, DAÑOS Y CONTEXTOS EMPRESARIALES DE LOS POD

2.1. De la arquitectura de la decisión al patrón oscuro digital: tipologías y rasgos estructurales

En el núcleo de lo que hoy denominamos *POD* late la arquitectura de la elección, a partir de la idea de que el modo en que se presentan las opciones determina, en buena medida, la decisión que se toma. La posibilidad de orientar la acción del usuario a través de la configuración del entorno en el que decide (la organización de las alternativas, los ritmos de la interacción, la visibilidad de la información o la fricción necesaria para avanzar o desistir) no es, como tal, engañar al usuario, sino persuadirlo. De hecho, la misma lógica late detrás de lo que Thaler y Sunstein (2009) denominaron los *nudges*, traducido al español como “acicates”, defendidos como herramientas útiles e incluso éticas que también pretenden influir en la conducta de las personas a partir del conocimiento de sus heurísticos, aunque, en ese caso, en beneficio del propio sujeto, ayudándole a superar sesgos cognitivos previsibles sin coartar su libertad de elección. Efectivamente tanto el *acicate* y el patrón oscuro digital comparten la raíz metodológica del aprovechamiento de los sesgos y heurísticos documentados por Kahneman y Tversky (1979), así como de otras vulnerabilidades del usuario¹⁴. Sin embargo, divergirían, en primer lugar, en su finalidad y en el grado transparencia y, después, en sus consecuencias¹⁵. Los acicates buscarían corregir sesgos en beneficio del propio sujeto; mientras que los patrones oscuros pretenden explotarlos sistemáticamente en beneficio del diseñador. En otras palabras, mientras que el *nudge* pretende ayudar a decidir mejor, el patrón oscuro busca impedir que se decida libremente por medio de técnicas que buscan crear costes de reversión o asimetrías de fricción para anular de facto la capacidad del usuario para deshacer su decisión¹⁶.

¹⁴ Al fin y al cabo, tienen razón Rossi et al. (2024) cuando señalan que hay vulnerabilidades más allá de los sesgos cognitivos y que centrar lo que puede ser explotado en el ámbito digital en esto constituye una visión algo simplista que no capta cómo condiciones socioeconómicas, contextuales o relacionales también pueden incrementar la susceptibilidad a estos diseños.

¹⁵ En este sentido podría referenciarse el análisis clásico de Gray et al. (2018), quienes sostienen que los patrones oscuros aparecen cuando el valor del usuario es sustituido por el valor del accionista y cuando el conocimiento del comportamiento humano se aplica para inducir al usuario a actuar no a favor sino en contra de su propio interés.

¹⁶ En este sentido, para Mathur et al. (2021) un patrón se vuelve “oscuro” cuando la interfaz deja de ayudar a decidir y pasa a empujar, engañar o presionar al usuario para que haga algo que normalmente no haría y que beneficia sobre todo a la empresa, explotando sesgos, asimetrías

Pero, más allá de la diferente intención entre uno y otro, el paso, o salto, entre el “*nudge*” y el “*dark pattern*” tiene que ver con el propio ecosistema en el que se desarrolla la nueva arquitectura de elección¹⁷. La traslación de las técnicas de la arquitectura de elección al ecosistema digital, en el que las posibilidades de diseño de la interacción son muchísimo mayores que en el espacio físico, habría alterado profundamente la naturaleza de los propios acicates. Así lo sostuvo Yeung quien en 2017 ya hablaba de *hypernudges* para referirse a un nuevo tipo de arquitectura de elección nutrida del uso intensivo de datos y el aprendizaje algorítmico automatizado que sería dinámico y permitiría ajustar en tiempo real las condiciones de elección en función del comportamiento previo del usuario (Yeung, 2017). Así, lo que antes podía ser una técnica puntual de influencia se convierte en un mecanismo continuo de optimización del comportamiento, alimentado por datos y validado por métricas. Esto, en manos de las grandes plataformas convierte lo que en su formulación inicial pretendía ser una herramienta paternalista y pro-social, en una tecnología de manipulación más que de persuasión donde la arquitectura de la decisión busca para inducir elecciones que benefician al proveedor en detrimento del usuario. Así, el rasgo verdaderamente distintivo de los POD frente a sus antecesores analógicos no reside tanto en sus técnicas, muchas de las cuales son herederas del marketing conductual clásico, como en la infraestructura de datos y aprendizaje automático que los sostiene. Esta base tecnológica, impulsada por la IA, y la consiguiente oscuridad, en el sentido de falta de transparencia, supone en términos de imposibilitar la comprensión del actuar propio y la consiguiente negación de la capacidad potencial de actuar autónomo del individuo. Es el entorno digital el que permite observar, registrar y cuantificar cada interacción del usuario, de modo que acaba siendo el propio usuario quien con su propio actuar va afinando la manipulación adecuada para él¹⁸.

de información o fricciones ocultas en el diseño. Es la combinación de intención estratégica, desalineación de intereses y explotación sistemática de la posición de debilidad del usuario lo que, según los autores, distingue un “*dark pattern*” de un simple diseño persuasivo o torpe. En el mismo sentido, Luguri y Strahilevitz (2021) a partir de su estudio experimental señalan que la diferencia entre el *nudge* y el *dark pattern* no es sólo ética, sino funcional.

¹⁷ Asiste razón a Chang et al. (2024) en que hay un uso excesivamente simplificado del *Nudge* como teoría de referencia en la investigación empírica sobre patrones oscuros, muchas veces sin explicitar sus límites ni sus diferencias funcionales respecto a los mecanismos manipulativos. Su revisión confirma que buena parte de la literatura confunde técnicas de persuasión benigna con mecanismos de distorsión de la autonomía, lo que subraya la necesidad de distinguir entre acicates y patrones oscuros como categorías normativamente no equivalentes.

¹⁸ De hecho, en uno de los estudios se observa que los mecanismos de captura atencional sólo se hacen visibles al usuario cuando se eliminan o se alteran deliberadamente: al desactivar

Las primeras tentativas descriptivas intentaron inventariar los distintos tipos de POD¹⁹. La literatura reciente se ha orientado hacia la construcción de una gramática común que permita comprender su lógica y clasificarlos de modo coherente. Entre los esfuerzos más influyentes destaca el de Colin Gray, Yubo Kou y su equipo (Gray et al., 2018a, 2024), quienes propusieron una ontología del patrón oscuro concebida a partir del modo en que estos mecanismos se integran en los ciclos organizativos de diseño y experimentación de producto. En su planteamiento, cada patrón se define por la relación entre una intención (la finalidad de orientar la conducta del usuario en una dirección determinada), un mecanismo cognitivo (el sesgo psicológico que se explota para lograrlo) y un contexto organizativo (las estructuras, rutinas y objetivos comerciales que hacen posible su implementación). Conforme a esto habría cinco estrategias principales, 1) el Nagging, o fastidio persistente (técnicas que interrumpen repetidamente al usuario para que realice una acción no deseada (por ejemplo, cuadros emergentes constantes o repetición insistente de opciones de suscripción); 2) la obstruction u obstrucción deliberada (hacer más difícil de lo necesario una acción que va en contra del interés de la empresa, como cancelar una suscripción o borrar una cuenta); 3) el Sneaking (Acciones ocultas o a escondidas): incorporar funcionalidades o consecuencias no visibles al usuario (como añadir servicios sin que el usuario sea consciente o modificar preferencias por defecto), 4) la Interface Interference (Interferencia en la interfaz): manipular el diseño visual para sesgar la atención del usuario, por ejemplo, resaltando opciones favorables a la empresa con colores llamativos y ocultando las alternativas.(5) Forced Action (Acción forzada): obligar al usuario a realizar un acto previo o adicional para acceder al servicio, como introducir datos personales innecesarios o vincular cuentas.”. Todas estas formas de manipulación, señalaban

dinámicas como el *scroll* infinito, las recomendaciones personalizadas o el *autoplay*, se reduce de forma significativa la distracción y aumenta la percepción de control (Lee et al., 2025). Es decir, que la arquitectura digital no solo observa y aprende del usuario, sino que estructura su experiencia de forma que dificulta la toma de conciencia sobre los mecanismos que moldean su propio comportamiento.

¹⁹ Harry Brignull (2010) estableció una tipología informal basada en ejemplos cotidianos de engaño en el diseño, desde los formularios que ocultan la opción de darse de baja (*roach motel*) hasta las suscripciones que se renuevan automáticamente sin advertencia (*forced continuity*), cuyo valor radicó más en su potencia heurística que en su sistematicidad. En 2019, el estudio de Mathur et al. ofreció la primera taxonomía a gran escala, basada en el análisis de más de once mil páginas web comerciales, y agrupó los patrones en cinco grandes familias: *obstruction*, *sneaking*, *interface interference*, *forced action* y *misdirection*. Aquella investigación demostró que los patrones oscuros no eran anomalías marginales, sino un componente estructural del comercio digital, presente en más del 11 % de los sitios examinados.

los autores, estaban institucionalizadas como prácticas productivas, por lo que constituían auténticos dispositivos empresariales que emergían de la interacción entre objetivos económicos, conocimiento conductual y experimentación algorítmica (Gray et al., 2018a, 2024).

Estas aportaciones fueron asimiladas por los principales organismos internacionales de supervisión como la Federal Trade Commission (FTC) de Estados Unidos, en su informe *Bringing Dark Patterns to Light* (2022) y la Organización para la Cooperación y el Desarrollo Económicos (OCDE), en su documento *Dark Commercial Patterns* (2022)²⁰. En el ámbito europeo, las Guidelines 03/2022 del European Data Protection Board (2022), dedicadas específicamente a los *deceptive design patterns* (patrones de diseño engañoso) en redes sociales, ofrecen una de las clasificaciones más depuradas desde el punto de vista técnico y jurídico. En ellas se identifican seis grandes modalidades de interferencia: *overloading* (sobrecarga de información o estímulos), *skipping* (inducción a saltarse pasos relevantes), *stirring* (activación emocional o manipulación afectiva), *hindering* (dificultad o impedimento artificial), *fickle* (inconsistencia en las opciones ofrecidas) y *left in the dark* (dejar al usuario en la oscuridad informativa). Desde la perspectiva de la protección de datos el elemento que interesaba era la alteración del contexto en que se prestaba el consentimiento y la obstrucción de los mecanismos que permitirían revocarlo. Pero tanto en esta taxonomía como en las anteriores se encontraba un núcleo común: el núcleo de la manipulación no radica en la estética o la técnica del diseño persuasivo, sino en su efecto acumulativo sobre la capacidad del consumidor para tomar decisiones libres e informadas, de modo que la arquitectura digital se convierte en un instrumento de gobierno de la voluntad.

Teniendo en cuenta esto, creo que más relevante que escoger una taxonomía actualizada de todos los posibles mecanismos de diseño que entrarían dentro de los POD, sería dotarnos de una tipología de los rasgos de “oscuridad” de la arquitectura digital que nos permita reconocer cuándo, por la presencia de los mismos, un diseño o interfaz deja de ser simplemente persuasivo y se transforma en un patrón

²⁰ Donde estableció cuatro categorías fundamentales: *interface interference* (interferencia de la interfaz), *choice pressure* (presión de elección), *misrepresentation* (falsedad o distorsión informativa) y *obstruction* (obstrucción de la decisión o de la salida) (Federal Trade Commission, 2022). Por su parte, la Organización para la Cooperación y el Desarrollo Económicos (OCDE), en su documento *Dark Commercial Patterns* (2022), propuso una clasificación más compleja que combina vectores funcionales y psicológicos, distinguiendo entre *misdirection* (inducción o desvío de la atención), *social proof and urgency cues* (prueba social y señales de urgencia), *nagging and obstruction* (insistencia y obstrucción), *confirmshaming* (vergüenza por no confirmar) y *sneaking* (ocultamiento o inclusión encubierta de opciones no deseadas) (OECD, 2022a).

oscuro digital propiamente dicho. La que propongo a continuación, basándome en la literatura anteriormente citada, incluye una serie de rasgos o dimensiones del diseño digital que, cuando operan, lo suelen hacer de forma interrelacionada, en cuanto que constituyen la gramática de la manipulación digital que, en cada contexto, aparecerá plasmada de un modo u otro, pero siempre con algunos de estos elementos. En otras palabras: ninguno de estos rasgos da lugar por sí mismo a un patrón oscuro ni tiene que estar presente para que el diseño pueda ser calificado como tal: será la delineación conjunta de algunos de ellos lo que genera ese diseño manipulativo oscuro que determina la conducta.

El primer rasgo de oscuridad es la sistematización trazable de la manipulación. Los patrones oscuros no aparecen por azar ni por torpeza en el diseño: son el resultado de un trabajo minucioso de prueba y ajuste. Cada versión de una interfaz se experimenta con distintos grupos de usuarios, mediante pruebas A/B u otros métodos similares, y se evalúa su impacto en métricas como la tasa de conversión, el tiempo de permanencia o la frecuencia de clics (Mathur et al., 2019). Este seguimiento continuo no solo permite identificar cómo se comportan los usuarios, sino también qué manipulaciones funcionan mejor para orientar sus decisiones. El diseño deja así de ser una actividad creativa basada en la intuición para convertirse en una auténtica ingeniería del comportamiento, un laboratorio en permanente funcionamiento en el que la interfaz se afina para obtener el máximo rendimiento de nuestros errores, hábitos y automatismos. Por supuesto, puede haber manipulación sin este nivel de sistematicidad; pero, en ese caso, no estaríamos ante un patrón oscuro en sentido estricto.

El segundo rasgo, fundamental para entender la transición del diseño persuasivo al patrón oscuro digital, es la personalización algorítmica de la influencia. Mientras que los diseños tradicionales ofrecían la misma interfaz para todos, los actuales aprenden del propio usuario y adaptan su comportamiento en tiempo real. Cada clic, cada pausa, cada gesto de navegación se convierte en un dato que alimenta modelos capaces de calibrar, con precisión casi experimental, qué combinación de estímulos provoca la reacción deseada. Esta capacidad de ajuste fino permite no solo adaptar el diseño a los hábitos generales de una población, sino también explorar y explotar vulnerabilidades individuales: impulsividad, preferencias, inseguridades, ritmos de uso o momentos de mayor cansancio o distracción. El resultado es un entorno que ya no persuade, sino que interviene directamente en la conducta, moldeando la experiencia de cada persona mediante una

arquitectura hecha a su medida²¹. Algunos usuarios verán un mensaje de urgencia (“solo quedan dos unidades”), otros recibirán un estímulo de prueba social (“otros usuarios están viendo este producto”), y otros serán persuadidos mediante recompensas o descuentos personalizados, dependiendo de lo que previamente se haya visto que funciona mejor con ellos (Luguri & Strahilevitz, 2021).

El tercer rasgo de oscuridad es, precisamente, la opacidad estructural, la falta de transparencia a la que lleva el diseño con POD. Este rasgo también deriva de la dimensión “data-driven” de los nuevos diseños, la personalización de los mismos y, por tanto, la absoluta asimetría cognitiva entre quien diseña la plataforma y quien interactúa con ella. Mientras que los algoritmos “aprenden” del propio usuario, al usuario desconoce tanto el proceso como los criterios que guían la presentación de opciones, creándose una opacidad estructural que impide comprender el modo en que la decisión ha sido inducida (Luguri & Strahilevitz, 2021). Esto se materializa de distintas maneras, aunque una de las más habituales es la ocultación o dispersión calculada de la información relevante. Un ejemplo claro son los precios que se desglosan en pasos sucesivos para impedir que el usuario tenga una visión completa del coste hasta el final del proceso; otro, las cláusulas esenciales diluidas en textos excesivamente largos o formulados de modo que desalientan su lectura; o, también, la ubicación de los botones de rechazo en lugares poco visibles, escondidos entre varios niveles de menús o presentados de forma menos destacada que las opciones favorables a la empresa. Al mismo tiempo, el diseño genera una ilusión de claridad y de control: se ofrece abundante información secundaria o irrelevante, pero se retiene (o se oculta bajo capas de interfaz) aquella que es decisiva para comprender las implicaciones reales de la elección. El resultado es un entorno en el que el usuario no dispone en el momento adecuado de los datos necesarios para evaluar las alternativas y, por tanto, se ve impedido para llevar a cabo una deliberación racional plenamente informada.

Este rasgo de oscuridad está estrechamente vinculado con el siguiente: la llamada *fricción asimétrica*. Se trata, en esencia, de alteraciones del equilibrio entre las distintas opciones de interacción para inclinar la balanza hacia una de ellas. La arquitectura digital permite diseñar desde el principio entornos en los que cada decisión

²¹ En este sentido, creo que es muy interesante el reciente trabajo de Rossi et al. (2024) que va mucho más allá de la visión de la vulnerabilidad como una característica fija de grupos previamente clasificados y la entiende como un fenómeno micro, meso y macro, dependiente de factores personales, contextuales y estructurales, que resulta de la interacción entre disposiciones individuales, condiciones sociales y la arquitectura digital que modela el comportamiento.

tiene un coste (de tiempo, esfuerzo o atención) distinto, y en los que ese coste puede distribuirse de forma estratégica. Así, la manipulación no actúa únicamente a través de lo que se muestra u oculta, sino también mediante la propia forma en que están configuradas las opciones. Algunas decisiones se presentan como caminos rectos y sin obstáculos, en los que un solo clic basta para aceptar, continuar o contratar, mientras que otras se convierten en auténticos laberintos: cadenas de menús, confirmaciones sucesivas, ventanas emergentes o incluso la exigencia de realizar una llamada telefónica. Se simplifica lo que conviene a la plataforma y se complica lo que beneficia al usuario, generando un terreno de juego inclinado que desincentiva la elección autónoma y empuja suavemente hacia el resultado que la empresa desea. Estos patrones de diseño, ya identificados empíricamente (Mathur et al., 2019) y ya recogidos en el ámbito regulatorio (European Data Protection Board, 2022), tienen, en particular, la capacidad de afectar a la autonomía al reducir, hasta el punto de impedir, la libertad de desistimiento sobre las acciones ya iniciadas.

Finalmente, y quizás como el rasgo más parecido a un “cajón de sastre de la oscuridad del patrón digital” tendríamos la fundamentación del diseño a partir de heurísticos psicológicos para orientar la conducta del usuario. Lo que antes eran mecanismos estudiados por la psicología conductual y cognitiva (o incluso simples reglas mentales que usamos para decidir “pensando rápido”) se convierte ahora en una arquitectura de persuasión automatizada incrustada en la interfaz. Heurísticos como el de escasez (“solo quedan dos unidades”), el de urgencia (“oferta disponible durante cinco minutos”), el de prueba social (“otras personas están comprando esto ahora”), el de autoridad (“selección recomendada”) o el de consistencia (“ya has empezado, continúa”) pasan a funcionar como disparadores programados que actúan directamente sobre la emoción y el impulso. Esta lógica se intensifica en entornos gamificados, donde la interfaz adopta dinámicas propias del juego: puntos, logros, rachas de actividad, estallidos de colores o pequeñas celebraciones visuales. En esos contextos, el diseño no solo orienta ni sugiere: condiciona. Lo que comienza como una acción voluntaria termina convirtiéndose en un gesto casi automático, una respuesta-reflejo cuidadosamente inducida por el sistema.

Como se ha dicho, los POD pueden concretarse de muy distintas formas, pero generalmente lo hará con la presencia de algunos de estos rasgos, de esta gramática oscura del diseño manipulativo, que hace que las interfaces dejen de ser el supuesto medio neutro de interacción social que se nos dijo y se conviertan en una tecnología de dirección de la conducta. Y, de nuevo, su proliferación no puede considerarse

una casualidad, sino la lógica consecuencia de la interacción entre digitalización, conocimiento del actuar humano y la lógica de la maximización de ganancias (ESMA, 2023; Mathur et al., 2021; Luguri & Strahilevitz, 2021; King et al., 2019; Zagal et al., 2013)²². Si son los equipos de diseño los que construyen la arquitectura de la plataforma pensando en maximizar el impulso humano de las formas descritas o de otras y de evitar la reflexión; si esto les viene previamente determinado por modelos de monetización en los que el usuario condicionado es el producto; y si esto acontece en el contexto de una economía capitalista de regulación fragmentada, el patrón oscuro no puede ser visto como un mero accidente ético sino como un recurso productivo y un daño, a la vez.

2.2. Plataformas, entornos y servicios con POD

Los POD no aparecen de manera homogénea en el ecosistema digital, sino que se concentran allí donde la interacción entre usuario y plataforma puede traducirse en datos y estos, a su vez, en beneficio económico. Si en el apartado anterior se mostraban sus mecanismos estructurales, aquí interesa observar cómo, en distintos entornos empresariales, esas lógicas adquieren concreción y se convierten en prácticas organizativas normalizadas.

Uno de los ámbitos donde se ha encontrado más evidencia del uso de patrones oscuros es en las propias plataformas de redes sociales, donde el modelo económico se basa en la captación y retención de la atención del usuario. En cierta medida puede afirmarse que el producto de las redes sociales es precisamente el usuario o más bien su tiempo y sus interacciones dentro del ciberlugar, en cuanto que las formas de monetización principales, como la publicidad segmentada o el comercio embebido, siguen dependiendo de que aquél pase el mayor tiempo posible en la red social. Así, cada gesto, pausa o desplazamiento de pantalla constituye un fragmento de información valioso, transformado en predicción conductual, y en valor monetizable. Este ecosistema que Davenport & Beck (2001) ya describieron como “economía de la atención” demanda una arquitectura de diseño que busca condicionar las acciones del usuario para retenerle y seguir monetizando su tiempo en la red. Además, la investigación reciente sobre diseño adictivo confirma este enfoque estructural: los mecanismos de persuasión algorítmica no responden a decisiones

²² De hecho los análisis empíricos más recientes confirman que estos rasgos rara vez operan de forma aislada: en juegos móviles *free-to-play*, por ejemplo, los patrones suelen agruparse en arquitecturas combinadas (*DP combos*) y reforzarse mediante elementos que amplifican su efecto (*DP enhancers*), lo que subraya el carácter sistémico y no accidental del diseño manipulativo (Zhang et al., 2024).

aisladas, sino a modelos de negocio basados en maximizar atención y tiempo de uso, especialmente entre menores²³. Las funciones de reproducción automática, el desplazamiento infinito de contenidos (*infinite scroll*), las notificaciones que interrumpen los ciclos de concentración o los mensajes que apelan al temor a quedar excluido (el ya conocido FOMO, *fear of missing out*) configuran un entorno donde el usuario apenas puede decidir cómo responde a una interfaz y a unos estímulos que no son accidentales ni improvisados sino que han sido previamente diseñados para optimizar métricas como la duración de la sesión, la tasa de interacción o la frecuencia de retorno. Aún con el riesgo que conlleva cualquier generalización, puede afirmarse que en las redes sociales el patrón oscuro digital no es una táctica puntual sino la propia atmósfera en la que tiene que respirar el usuario al entrar en la interfaz.

Pero hay otros ámbitos en los que el modelo de negocio de las empresas depende del diseño de una interfaz pensada para que el usuario se comporte de una forma determinada en ella. Uno de ellos es el de la industria del videojuego, donde la transición desde el modelo de compra tradicional hacia los sistemas *free-to-play* y de servicios continuos (*live service*) ha transformado el entretenimiento en un espacio de monetización del comportamiento²⁴. En muchos juegos on line el jugador deja de ser comprador de un producto para convertirse en parte de un experimento conductual permanente: las recompensas variables, las cajas de botín (*loot boxes*), los eventos de tiempo limitado o la escasez programada explotan los principios de refuerzo intermitente identificados en la psicología conductual clásica de Skinner (Ferster y Skinner, 1957; Skinner, 1953, 1938) y posteriores.

²³ Así, González Tapia (2025) señala que la lógica económica de estas plataformas reproduce dinámicas similares a las ‘nuevas tabaquerías’, cuyo beneficio depende de generar hábitos compulsivos.

²⁴ Esta evolución se ha documentado de manera sistemática en un muy interesante estudio sobre transiciones de modelos de negocio (Hadan et al., 2024), que muestra que el paso de modelos *buy-to-play* a *free-to-play* ha supuesto la introducción de nuevos mecanismos de monetización con características claramente manipulativas (como, y me someteré aquí voluntariamente a los anglicismos sin buscar traducción, el *grinding*, *infinite treadmill*, *anchoring*, *fake scarcity*, *pay-to-skip* o *endowed progress*) que transforman profundamente la experiencia del jugador y su autonomía. También en el ámbito de los denominados juegos generados por usuarios (UGGs) (principalmente Roblox, Minecraft, Rec Room y Fortnite Creative), un ecosistema particularmente vulnerable a prácticas de diseño engañosas y económicamente explotadoras. En un estudio reciente basado en entrevistas a jugadores de estas plataformas se ha identificado una proliferación de prácticas de monetización predatoria, describiéndose experiencias de manipulación percibida, presión grupal, y dinámicas de gasto inducido que revelan la profundidad con la que estos patrones se insertan en la arquitectura económica del videojuego (Zhang et al., 2025).

Son varios los estudios empíricos (King et al., 2019; Zendle y Cairns, 2018) que han mostrado cómo estos mecanismos no solo incrementan el gasto, sino que generan dinámicas de dependencia psicológica similares a los derivados del juego patológico, en las que la decisión económica es reemplazada por un impulso emocional inducido, con las potenciales consecuencias que ello conlleva de que los daños no sean solo patrimoniales sino psicológicos por la propia erosión del control de la propia conducta.

Y muchos de los mecanismos psicológicos y estructuras de recompensa desarrollados por la industria del videojuego han sido trasladados, casi sin fricción, al ámbito financiero. La gamificación, concebida originalmente como una estrategia de implicación lúdica, se convierte aquí en un instrumento de manipulación económica. Las plataformas on line de *trading* minorista replican el ecosistema del juego: transforman la inversión, una actividad tradicionalmente racional y analítica, en una experiencia emocional y acelerada, guiada por estímulos visuales y refuerzos inmediatos; donde el cierre de una operación se celebra con confeti digital, los ránquines y los logros fomentan la competencia constante, y donde los sistemas de notificaciones convierten cada transacción en una pequeña dosis de gratificación²⁵. Todo ello responde a un mismo principio de diseño: inducir al usuario a actuar con mayor frecuencia y menor deliberación, replicando los mecanismos de refuerzo variable ya conocidos en los videojuegos (King et al., 2019; Zendle y Cairns, 2018), erosionando la percepción del riesgo por parte del inversor y desplazando la decisión económica hacia el impulso emocional²⁶. En torno al trading financiero ha proliferado, además, un ecosistema paralelo de plataformas, formación y asesoramiento que, bajo la apariencia de educación sobre la inversión bursátil, reproduce idéntica estructura de diseño. Los cursos, simuladores y comunidades de *traders* prometen control y autonomía, pero en realidad refuerzan una ilusión de maestría que alimenta la sobreinversión y la exposición repetida al riesgo²⁷. En este contexto, los *POD* se manifiestan no sólo en la interfaz (en los botones, en los colores o en las trayectorias de navegación), sino en la estructura misma de la experiencia, concebida para sostener

²⁵ Esa configuración, en estudios recientes, aparece asociada a conductas excesivas y adictivas: uso problemático de plataformas, dependencia psicológica, estrés, y riesgo incrementado de pérdidas graves (Orujov, 2023; Oksanen et al., 2022).

²⁶ Como han advertido la Autoridad Europea de Valores y Mercados (ESMA, 2023) y la Autoridad Bancaria Europea (European Banking Authority, 2024a, 2024b) estas dinámicas erosionan la percepción del riesgo y distorsionan el juicio financiero del inversor.

²⁷ Véase, en este sentido, el estudio de CEA Institute (2022), y sobre el mecanismo social para el funcionamiento de estas plataformas el interesante estudio de Klocke et al. (2025).

un ciclo ininterrumpido de acción y recompensa que convierte la vulnerabilidad del usuario en motor de rentabilidad.

Por supuesto, también en el comercio electrónico la arquitectura del entorno digital desempeña un rol fundamental para lograr “capturar al comprador”. Los sistemas de *drip pricing*, que fragmentan la información sobre el precio total, los suplementos preseleccionados o las dificultades para cancelar una suscripción forman parte de una gramática común a la compraventa a través de internet. Conforme a un interesante estudio empírico sobre más de once mil sitios de compraventa, este tipo de patrones afectan a la mayoría de las principales plataformas globales, y su efecto agregado es un incremento sistemático del gasto no deseado (Mathur et al., 2019). La manipulación, aquí, no opera por persuasión, sino por extenuación: el consumidor termina eligiendo lo que no pretendía simplemente porque es más difícil o costoso no hacerlo.

2.3. El daño de los POD: más allá del perjuicio económico, la erosión de la autonomía individual

De cara a la determinación de la potencial responsabilidad por el diseño de los POD hay que ir más allá de la mera descripción técnica de sus mecanismos o de la identificación de los ámbitos en que se despliegan. Es necesario entender el daño que los mismos pueden producir tanto a nivel individual como colectivo y, también, aunque ello se abordará más adelante, hasta qué punto dicha lesividad es parte de la lógica económica empresarial que impulsa estas prácticas en el nuevo ecosistema digital. Y, aún sin entrar en un análisis profundo, el análisis llevado a cabo sobre los POD y sus rasgos característicos apunta a una múltiple lesividad potencial de estas técnicas, tanto en el sentido de los variados bienes jurídicos que pueden verse afectados por su uso como en el de la existencia de una multiplicidad de capas de daño que la proliferación de su uso puede conllevar. En concreto se hace evidente que hay una doble dimensión del daño en el patrón oscuro digital: la individual, en forma de pérdidas patrimoniales cuantificables, deterioro del bienestar psicológico y, sobre todo, erosión de la autonomía práctica, entendida como la capacidad del individuo para tomar decisiones informadas y libres; y la colectiva, debido al impacto social que puede tener el desarrollo de modelos de economía digital que extraen valor de la vulnerabilidad, erosionan la competencia leal y debilitan la confianza en las infraestructuras digitales y comerciales. Realizaré, a continuación, un primer análisis de estas capas potenciales de daño diferenciando entre los potenciales bienes jurídicos afectados antes de, en el siguiente punto, profundizar en cada posible afectación a los mismos a la hora de determinar la potencial responsabilidad penal

empresarial.

El primer estrato de daño, el más visible y el más fácilmente traducible a algo concreto es el daño patrimonial directo que los patrones oscuros pueden generar. Configuraciones de interfaz que dificulten cancelar una suscripción, que induzcan a compras impulsivas, que oculten información sobre costes ocultos o que fragmenten la visualización del precio total, pueden empujar a los usuarios a decisiones contrarias a su interés económico y conllevar perjuicios patrimoniales directos²⁸. Es decir, no sólo sabemos que la forma en que se presenta la información digital condiciona la capacidad real de los consumidores para comparar precios, evaluar alternativas o interpretar los costes totales, lo que reduce la calidad de la decisión económica (OECD, 2022b), sino que también hay evidencias de que las prácticas manipulativas citadas llevan en muchas ocasiones a los consumidores a tomar decisiones que producen un perjuicio directo o indirecto a sus intereses (OECD, 2022a). Por ejemplo, de que la exposición controlada a varios patrones oscuros incrementa el gasto del consumidor, afectando especialmente a la propensión a comprar y a la percepción de valor²⁹.

Más allá del perjuicio económico inmediato los patrones oscuros producen un daño más profundo, la erosión de la autonomía práctica del usuario. Este daño no se manifiesta siempre en términos monetarios, sino que afecta al núcleo de la capacidad de autodeterminación. El hecho de que las decisiones digitales se tomen en entornos de elección diseñados para modular y orientar la conducta puede no sólo distorsionar las condiciones bajo las cuales la elección tiene lugar sino, en casos extremos, impedir la deliberación reflexiva y convertir la autonomía en un ideal formal desvinculado de la realidad material de la decisión digital³⁰. Este daño a la autonomía tiene un

²⁸ Ejemplos como la manipulación del BuyBox de Amazon, el autopreferenciamiento de Google en resultados de búsqueda o la explotación algorítmica de vulnerabilidades en EOMM de videojuegos muestran cómo los POD pueden provocar un daño económico directo mediante la inducción a compras más caras o menos eficientes de un modo similar a como sucede en fraudes más tradicionales con engaño (Ziermann, 2024).

²⁹ Así, en los estudios experimentales de Koh y Seah (2023) se ha documentado con precisión cómo distintos tipos de patrones generan efectos diferenciados sobre el comportamiento de compra, confirmando, además, que la manipulación puede medirse, replicarse y producir beneficios empresariales sistemáticos. Por su parte recientes estudios empíricos en videojuegos móviles muestra, además, que la combinación de patrones monetarios y psicológicos (especialmente los basados en urgencia, recompensa variable o progresión artificial) incrementa significativamente el gasto no deseado y reduce la capacidad deliberativa, reforzando esta dimensión de daño a la autonomía (Zhang et al., 2025).

³⁰ Trabajos recientes confirmarían este vínculo entre diseño y erosión de la autonomía: cuando se eliminan elementos de interfaz destinados a la captura automática de la atención, como el

primer componente psicológico: los sesgos y atajos mentales pueden ser explotados de forma sistemática y producir consecuencias muy perjudiciales como las que se suelen incluir dentro del denominado “uso problemático de tecnologías digitales” caracterizado por pérdida de control, impulsividad y deterioro del bienestar subjetivo³¹. La literatura sobre adicción tecnológica y persuasión algorítmica confirma esta erosión de la autonomía, mostrando cómo mecanismos como el *scroll* infinito, las recompensas variables intermitentes, las notificaciones constantes o la explotación del FOMO operan sobre el sistema neural de recompensa y fomentan conductas compulsivas, especialmente en adolescentes (véase González Tapia, 2025). Como ha señalado González Tapia (2025), el desarrollo neuropsicológico en curso en menores y adolescentes aumenta la sensibilidad a la recompensa inmediata, a la validación social y a la presión del grupo, elementos que las plataformas explotan deliberadamente mediante estos diseños adictivos.

Pero más allá de los potenciales efectos económicos y psicológicos concretos que pueden producir en los usuarios³² por la afectación de los POD al proceso mismo de decidir, la erosión de la capacidad de autodeterminación sostenida a lo largo del tiempo tiene implicaciones más profundas³³. Al fin y al cabo, ese daño puede tener un alcance sistémico, pues la autonomía no es sólo un constructo psicológico sino un valor moral individual e, incluso más allá, el fundamento de la validez del consentimiento en el sistema jurídico y social que puede verse afectado profundamente si la voluntad de las personas

desplazamiento infinito, las recomendaciones algorítmicas, las notificaciones interruptivas o la reproducción automática, la distracción autorreportada desciende de manera drástica y aumenta la percepción subjetiva de control, agencia y satisfacción con la navegación (Lee et al., 2025). A mi parecer esto muestra que la pérdida de autonomía no es un fenómeno abstracto, sino un efecto conductual directo de la arquitectura digital.

³¹ Estos mecanismos han sido documentados también en la literatura clínica sobre el denominado Uso problemático de Internet, que muestra cómo determinados entornos digitales pueden generar patrones de uso caracterizados por ‘*impaired control*’, prioridad excesiva de la conducta online y persistencia pese a consecuencias negativas (Fineberg et al., 2022). Esta evidencia refuerza que las arquitecturas de decisión manipulativas no solo inducen elecciones no deseadas, sino que pueden contribuir a dinámicas de uso impulsivo o compulsivo que comprometen la autonomía y agravan las vulnerabilidades situacionales.

³² Es interesante en este sentido la revisión sistemática de Chang et al. (2024), que pone de manifiesto que la mayor parte de la literatura ha focalizado el análisis en la psicología del usuario y ha ignorado sistemáticamente la mente y los incentivos del diseñador. Obviar esto supone, como después se defenderá, olvidar que los POD no son decisiones aisladas de diseño, sino productos organizativos.

³³ Así lo entienden también Chang et al. (2024), para quienes todo análisis de la manipulación digital debe tener en cuenta a la psicología pero, también, partir de teorías de la autonomía provenientes de la filosofía para comprender las implicaciones profundas de la afectación.

es preconfigurada por la arquitectura digital y no por su deliberación interna.

Y es que el análisis del daño producido por los POD no puede detenerse en la esfera individual. Aunque el perjuicio económico y la erosión de la autonomía a nivel individual constituyen daños inmediatos, la proliferación a gran escala de estas prácticas en una suerte de institucionalización de la extracción de valor a partir de errores cognitivos previsibles de los seres humanos podría generar efectos nocivos de carácter estructural. El primero de ellos podría ser la distorsión del funcionamiento de los mercados digitales en cuanto se favorezca a las empresas que integran el engaño por diseño como estrategia de negocio y se facilite que sobrevivan las empresas más eficaces en manipular, no las más eficientes en ofrecer valor. De hecho, estas estrategias no sólo afectan a la autonomía del usuario, sino que incluso pueden constituir *engaños estructurales* que alteran deliberadamente la arquitectura de mercado, como ocurre con las jerarquías falsas, las preselecciones y el *confirmshaming* empleados por Amazon y Google en sus modelos de negocio (Ziermann, 2024).

Otros autores han advertido sobre los efectos lesivos colectivos que puede conllevar la normalización social de la manipulación³⁴. Zuboff (2015, 2019) y otros autores (véase Susser et al., 2018) hablan de un daño institucionalizable producido por una manipulación sistémica o una industrialización de la manipulación. El problema no sería únicamente la influencia indebida, sino la industrialización de la manipulación, el desarrollo de un sistema económico en el que la vulnerabilidad es el insumo productivo y el error humano la base del

³⁴ Así Julie Cohen (2012) advierte que las arquitecturas de monitorización y diseño conductual no solo condicionan elecciones concretas, sino que reconfiguran las condiciones sociales de la libertad, acostumbrando a los individuos a un entorno donde la vigilancia y la inducción permanente son elementos naturales de la vida digital. El daño aquí es institucional: una erosión lenta pero sostenida de la confianza en las infraestructuras digitales, en el valor del consentimiento y en la posibilidad de una deliberación racional sin interferencias estructurales. A mi parecer hay otro riesgo que habría que plantear, y es la desensibilización progresiva ante el engaño. Si la manipulación se convierte en la norma, si el consumidor es sistemáticamente conducido a través de itinerarios diseñados para minimizar su capacidad de decisión, la frontera entre la persuasión legítima y el abuso se difumina, la manipulación deja de ser un caso excepcional para convertirse en una infraestructura; y con ello, el Derecho se enfrenta a un desafío nuevo: cómo responder a prácticas que, siendo masivas y automatizadas, generan daños distribuidos y difíciles de imputar a actos humanos concretos.

negocio.

3. RESPONSABILIDAD PENAL EMPRESARIAL Y POD

3.1. Preámbulo: cultura organizativa empresarial y economía del patrón oscuro digital

La comprensión de la potencial responsabilidad por los daños causados por el uso de POD nos lleva a desplazar el foco desde la interacción puntual con el usuario hacia el entramado organizativo que los produce. Como se ha argumentado en el capítulo anterior, estas técnicas parecen encajar poco en la idea de un error de diseño o de una desviación puntual del actuar de las empresas digitales y acercarse más a una expresión de un modelo económico desarrollado gracias a la convergencia entre la ingeniería del comportamiento, las nuevas plataformas digitales gobernadas por algoritmos y los incentivos corporativos orientados a la maximización del valor capturado por usuario. El patrón oscuro digital no emergería de la creatividad aislada de un diseñador, sino de un ecosistema económico digital y de una cadena de decisiones estructuradas que incluyen a equipos de producto, analítica, growth, marketing, ingeniería y dirección corporativa³⁵. El patrón oscuro digital como producto organizativo, más que individual, en el que los diseñadores usan prácticas manipulativas no por intención personal, sino por las presiones organizativas de un entorno donde las decisiones se orientan sistemáticamente a favorecer los intereses corporativos (Ziermann, 2024; Gray et al., 2018a). Esto será así, al menos, en aquellos diseños persuasivos en los que además de la voluntad de captación de usuarios, el diseño se sustenta en metodologías como las que hemos analizado anteriormente: A/B testing que genere una trazabilidad precisa de qué configuración aumenta conversiones, prolonga tiempos de uso o reduce las tasas de cancelación; experimentos controlados y personalización algorítmica para una optimización continua.

Existen ya algunos estudios que muestran esta dimensión organizativa del patrón oscuro digital. Es particularmente interesante, en este sentido, el estudio empírico de Gray et al. (2018), que constituye una de

³⁵ La evidencia empírica disponible confirma esta dimensión estructural: estudios recientes sobre videojuegos *free-to-play* muestran que los patrones manipulativos aparecen integrados desde el *onboarding* inicial, repetidos sistemáticamente en títulos distintos y articulados como parte de un flujo de diseño estandarizado, lo que refuerza la idea de que no se trata de desviaciones puntuales sino de productos organizativos normalizados (Zhang et al., 2025). En la misma línea, la investigación de Rossi et al. (2024), muestra que factores macro como las estructuras de mercado, la concentración del poder informacional o el propio diseño sistémico de los servicios, son determinantes en la producción de vulnerabilidad y en la exposición diferencial a prácticas manipulativas.

las primeras demostraciones sistemáticas de la dimensión organizativa de los patrones oscuros. A partir de entrevistas en profundidad con diseñadores de diferentes empresas tecnológicas, el estudio documenta cómo los equipos de producto, marketing, analítica y dirección establecen métricas de rendimiento (como la conversión, la retención o la reducción de cancelaciones) que funcionan como incentivos directos para la adopción de patrones manipulativos; también identifica que los rasgos de oscuridad y sus metodologías derivadas antes mencionadas generan un entorno en el que la variante de diseño más eficaz, aunque sea manipulativa, tiende a normalizarse e institucionalizarse. Y cuando estas técnicas usadas por los diseñadores se acaban consolidando como “mejor práctica” interna, se incorporan al flujo estándar del diseño y de la producción de servicios digitales y se valoran las mismas por factores como el potencial aumento en la cifra de conversiones o el incremento en la monetización derivado del despliegue de estos sistemas, el patrón oscuro se debe entender como un mecanismo normalizado de optimización y como parte de una política organizativa de facto. El patrón oscuro como acto de gobierno, en cuanto que la empresa crea condiciones estructurales que hacen probable, rentable y repetible la obtención de ventajas mediante interferencias indebidas en la voluntad del usuario. De hecho, investigaciones posteriores sobre cómo las organizaciones se adaptan a la regulación para seguir manteniendo patrones oscuros probarían esa cultura organizativa. En este sentido es interesante un estudio más reciente de Gray y asociados (2021) que muestra cómo tras la entrada en vigor del RGPD las empresas en vez de eliminar los patrones oscuros los reformularon, pero manteniendo la lógica manipulativa: opciones de rechazo más ocultas, jerarquías visuales sesgadas o fricciones añadidas estratégicamente. El que los patrones oscuros funcionen, incluso, como respuestas organizativas al entorno regulatorio, muestra hasta qué punto pueden ser parte en muchos casos de la cultura empresarial, de aquella en la que, como ocurre con otros daños delictivos, lo ilícito no nace como un acto de

desviación individual sino dentro de su modelo de negocio cotidiano³⁶.

3.2. Marco regulatorio europeo administrativo y patrones oscuros: hacia un estándar de diligencia corporativa ex ante

La irrupción de los patrones oscuros en la economía digital ha obligado a Estados y a instituciones supranacionales a reconfigurar profundamente su aproximación a la protección del usuario. El marco tradicional, centrado en la información, el consentimiento y la corrección *ex post* de los abusos, parece resultar insuficiente allí donde el diseño mismo de la interfaz se convierte en un mecanismo de influencia estructural y donde, como acabamos de ver, ello parte de la propia lógica de las empresas digitales. La Unión Europea ha ido asimilando progresivamente esta constatación hasta construir un cuerpo regulatorio que ya no se limita a sancionar las prácticas manipulativas, sino que exige que sean las empresas las que la prevengan mediante una gobernanza responsable del diseño digital.

El primer movimiento en esta dirección se produjo en el ámbito de la protección de datos. Aunque el RGPD no utiliza la expresión *dark patterns*, su definición de consentimiento como un acto “libre, específico, informado e inequívoco” (art. 4. 11) RGPD) abrió el espacio para cuestionar aquellas arquitecturas de interfaz que operan precisamente en sentido contrario. Las Directrices 03/2022 del Comité Europeo de Protección de Datos reforzaron esta interpretación al considerar que los “*patrones de diseño engañosos*”, tal y como los denominaba,

³⁶ Este planteamiento encaja de lleno con la literatura sobre *corporate crime* estructural, que describe cómo determinadas organizaciones generan entornos criminógenos en los que el daño emerge de procesos ordinarios de negocio más que de actos individuales. Así lo han mostrado, en ámbitos clásicos, autores como (Shover et al., 2007; Van de Bunt, 2007). Las aportaciones más recientes refuerzan esta idea al mostrar que la empresa puede llegar a comportarse como una institución criminógena, cuya arquitectura interna, presiones de rendimiento y lógicas de mercado canalizan a sus miembros hacia resultados ilícitos *sin necesidad de desviaciones individuales conscientes*. En esta línea, el trabajo de Györy (2020) resulta especialmente relevante para el ámbito digital. Su análisis institucional de la mayor estafa financiera de la Hungría pos-transición, el *Quaestor Scandal*, demuestra cómo, en sectores altamente tecnificados, la infracción puede llegar a normalizarse hasta el punto de que la organización actúa como si ciertas conductas prohibidas fueran parte natural del modelo de negocio. Su ‘autopsia institucional’ muestra que el fraude no brota de decisiones aisladas, sino de configuraciones estructurales (escasez crónica de capital, presión competitiva, baja profesionalización, incentivos institucionales perversos) que producen sistemáticamente comportamientos lesivos. Aplicado al ámbito de los patrones oscuros, este marco permite entender que estas prácticas no son desviaciones puntuales ni malas decisiones de diseñadores concretos, sino tecnologías corporativas de influencia, surgidas de procesos rutinarios de producto, experimentación masiva (*A/B testing*) y optimización algorítmica. Su persistencia no depende de individuos malintencionados, sino de la arquitectura organizacional y económica que las hace posibles y rentables. En este sentido, la literatura sobre fraude financiero institucionalizado y la que estudia patrones oscuros convergen en un mismo diagnóstico: los daños emergen de la estructura y no de la excepción.

constituyen prácticas incompatibles con el RGPD, especialmente en el ámbito de las redes sociales. El órgano europeo destacó que los procesos de registro que dificultan rechazar opciones, los diseños que ocultan alternativas menos favorables para la plataforma o las configuraciones predeterminadas que potencian la extracción de datos reducen, en la práctica, la autonomía del usuario. Este reconocimiento supuso un cambio relevante: la infracción ya no depende únicamente del contenido de la información, sino también de la forma en que se presenta. En cualquier caso, y como reconoce la práctica sancionadora de la AEPD reciente, el uso de patrones oscuros puede afectar a principios como el de la lealtad del tratamiento de datos (art. 5.1 RGPD) y consiguientemente imponerse las sanciones correspondientes³⁷.

A partir de aquí, el concepto se extendió más allá del ámbito de la privacidad y adquirió relevancia transversal. En esto coincide el enfoque europeo con el estadounidense. De hecho, las primeras manifestaciones de la OCDE en *Dark Commercial Patterns* (2022) y la Federal Trade Commission estadounidense, en *Bringing Dark Patterns to Light* (2022), definían los patrones oscuros como prácticas de diseño basadas en el engaño jurídicamente perseguibles y parecían compartir con la UE la identificación del problema. En lo que difiere la visión europea y la americana es en el tratamiento: mientras que el sistema normativo americano parece seguir operando desde una lógica más reactiva, el modelo europeo entiende la manipulación como un riesgo sistémico derivado del propio modelo de negocio de las plataformas digitales y busca implicar a las empresas en la propia prevención (Rossi et al., 2024). Las dos expresiones más claras de esta evolución en el Derecho europeo son el Reglamento de Servicios Digitales (DSA) (Reglamento (UE) 2022/2065, 2022) y el Reglamento de Mercados Digitales (DMA) (Reglamento (UE) 2022/1925, 2022) (véase Rossi et al., 2024). Sin entrar en un análisis detallado que excedería este texto (para más detalle véase Mato Pacín (2024)), el DSA prohíbe expresamente el uso de patrones de diseño engañosos en las interfaces (art. 25 DSA) e impone obligaciones en los sistemas de recomendación

³⁷ AEPD (2023) PS/00080/2023 de 30 de junio. En el caso se detectaron los siguientes patrones oscuros: “Sobre la utilización de patrones oscuros de sobrecarga (overloading) y de ocultación (skipping), en el caso que nos ocupa, el patrón oscuro de sobrecarga (overloading) y de ocultación (skipping), se observa cuando se accede al apartado “Lista de proveedores”, una vez allí, el interesado se encuentra con una lista de unas 130 empresas (proveedores), de las cuales, más de la mitad tienen marcada por defecto la casilla de “aceptar tratamiento de datos por Interés legítimo”, lo que obliga, en el caso de querer mostrar la oposición al tratamiento a marcar una a una a lo largo de toda la lista, sin que exista la opción de poder oponerse indicándolo una sola vez o un número de veces que resulte razonable y no genere fatiga en el afectado”.

y publicidad personalizada³⁸. Asimismo, los riesgos derivados de los patrones oscuros podrían mitigarse a través del sistema de riesgos sistémicos que prevé la DSA, como expresión o parte fundamental de las obligaciones de diligencia debida que prevé el texto. Este cambio de enfoque es esencial: el Derecho deja de centrarse únicamente en el contenido y pasa a regular también el diseño, reconociendo que la arquitectura informacional actúa como un vector de influencia tan poderoso y, en ocasiones, más que la comunicación explícita. El DMA, por su parte, se orienta a la protección de la competencia, pero lo hace con una lógica semejante. Frente a las grandes plataformas que ejercen un poder estructural (porque pueden influir de manera determinante en las decisiones de los usuarios y acumular datos procedentes de múltiples servicios) el DMA introduce límites específicos: prohíbe que se favorezcan a sí mismas frente a otros operadores, restringe la combinación de datos sin consentimiento y evita la creación de barreras estratégicas que dificultan que un usuario cambie de proveedor. Aunque su lenguaje es económico, sus efectos son claros desde el punto de vista del comportamiento: reduce prácticas que disminuyen la autonomía real del usuario en los procesos de elección y contratación digitales.

Esta evolución regulatoria ha tenido su última expresión en el Reglamento de Inteligencia Artificial (RIA) (Reglamento (UE) 2024/1689, 2024), que introduce por primera vez en el Derecho europeo una arquitectura jurídica explícitamente orientada a controlar la influencia algorítmica. El Reglamento parte de un enfoque basado en el riesgo para establecer en su artículo 5 la prohibición absoluta de sistemas de IA que utilicen técnicas subliminales, o que manipulen el comportamiento de los usuarios de forma susceptible de causar un perjuicio significativo. Aunque la formulación mantiene un tono prudente y acotado, se refiere únicamente a ciertos

³⁸ En este sentido es relevante tener en cuenta el considerando 67 de la DSA que establece que: “las interfaces engañosas de las plataformas en línea son prácticas que distorsionan o merman sustancialmente, de forma deliberada o efectiva, la capacidad de los destinatarios del servicio de tomar decisiones autónomas y con conocimiento de causa. Estas prácticas pueden utilizarse para persuadir a los destinatarios del servicio de que adopten comportamientos no deseados o decisiones no deseadas que tienen consecuencias negativas para ellos. Por esta razón, debe prohibirse a los prestadores de plataformas en línea engañar o empujar en esta dirección a los destinatarios del servicio y distorsionar u obstaculizar la autonomía, la toma de decisiones o la capacidad de elección de los destinatarios del servicio a través de la estructura, el diseño o las funcionalidades de una interfaz en línea o una parte de esta. Entre estas prácticas se han de incluir, entre otras, las opciones de diseño abusivas que dirigen al destinatario hacia acciones que benefician al prestador de plataformas en línea, pero que pueden no favorecer los intereses de los destinatarios, al presentar opciones de una manera que no es neutra, por ejemplo, dando más protagonismo a determinadas opciones mediante componentes visuales, auditivos o de otro tipo, cuando se le pide al destinatario del servicio que tome una decisión.”

colectivos vulnerables (en razón de su edad, situación económica o social, discapacidad) su importancia conceptual es profunda: implica reconocer que la interferencia cognitiva —invisible, personalizada, escalable— puede constituir un daño autónomo, no dependiente de un fraude informativo clásico ni de un perjuicio económico cuantificable. Y más allá de la prohibición estricta del artículo 5 podemos plantearnos si determinados sistemas de influencia algorítmica utilizados por plataformas digitales podrían ser clasificados como sistemas de alto riesgo bajo el AI Act³⁹. Lo cierto es que el RIA recoge como alto riesgo los sistemas que inciden en ámbitos como la gestión de servicios públicos, el acceso al empleo, la evaluación crediticia, la gestión educativa o los sistemas que influyen en el acceso a servicios esenciales por lo que la utilización en estos ámbitos de algoritmos de personalización y optimización del comportamiento capaces de condicionar decisiones sin intervención consciente del usuario, obligará a quienes diseñen estas herramientas digitales a implementar las exigencias reforzadas del RIA: evaluaciones de impacto sobre derechos fundamentales, trazabilidad, documentación detallada del ciclo de vida del modelo, supervisión humana significativa, gestión continua del riesgo y mecanismos de monitorización post-despliegue (véase con más detalle Cotino Hueso y Simón Castellano, 2024).

Puede asistir algo de razón algunos autores cuando señalan que la regulación europea de los patrones oscuros a través de estas normas (DSA, el DMA, el Data Act o el AI Act) los menciona de forma fragmentaria y sin un concepto normativo unificado, y que toda esta multiplicación de categorías regulatorias y taxonomías ad hoc lo que puede conllevar es un aumento en la inseguridad jurídica por su aplicación (Isola y Esposito, 2025). Frente a ello defienden los autores que la estructura jurídica más sólida y coherente para abordar los patrones oscuros sigue siendo, en realidad, la de la Directiva 2005/29/CE relativa a las prácticas comerciales desleales (véase también Comisión Europea, 2020) conforme a las cuales las prácticas oscuras pueden clasificarse legalmente como *misleading actions*, *misleading omissions*, *harassment*, *undue influence o coercion*, lo que nos permitiría situar estas técnicas de manipulación digital dentro de categorías consolidadas. Como después argumentaré esto puede resultar muy útil para la definición de los programas de cumplimiento.

Pero lo que, a mi parecer, es importante de toda la nueva regulación administrativa centrada en lo digital y en particular de los enfoques del DSA y el AI Act, más allá de que la misma será determinante para la

³⁹ En detalle sobre la práctica prohibida Comisión Europea (2025).

potencial responsabilidad penal en muchos casos, es su entendimiento de que, desde el diseño de la interfaz hasta el diseño algorítmico, la manipulación no es un efecto colateral, sino un riesgo que debe ser identificado, anticipado y neutralizado por la propia organización empresarial. Es el estándar de diligencia *ex ante* lo que constituye el núcleo del nuevo modelo europeo de gobernanza digital y el eje de eje de la responsabilidad. Frente a un modelo tradicional centrado en la corrección *ex post* del daño, la regulación europea exige ahora anticiparlo, identificarlo y neutralizarlo antes de que se produzca. Y este enfoque impone obligaciones concretas. Las empresas deben evaluar sistemáticamente los riesgos de manipulación, analizar de qué modo configuraciones de diseño, estrategias de recomendación o patrones de interacción pueden inducir decisiones no deseadas; deben documentar las decisiones de producto, registrando qué alternativas se valoraron, qué experimentos se realizaron y qué criterios justificaron la adopción final de una interfaz; deben someter a auditoría los sistemas algorítmicos que personalizan o priorizan información, permitiendo una supervisión externa que verifique la ausencia de prácticas manipulativas; y deben mantener estructuras internas de gobernanza que garanticen la integridad de la arquitectura de elección. En este marco, el cumplimiento ya no puede reducirse a declaraciones formales de transparencia o a meros ajustes estéticos de la interfaz: debe integrarse en la estructura misma del producto digital, como elemento constitutivo de su concepción, desarrollo y despliegue (Barrio Andrés, 2023).

3.3. Posibles delitos cometidos en el seno de la empresa y umbral penal de la manipulación

La cuestión central, una vez descritos los rasgos y daños de los POD y el marco regulatorio administrativo que los aborda, es determinar en qué medida estas prácticas pueden integrar tipos penales concretos del Código Penal español y cuándo la manipulación digital deja de ser un problema de consumo o de cumplimiento administrativo para convertirse en un ilícito penal. El punto de partida es claro: el Derecho penal económico español ya dispone de figuras que protegen la libertad de decisión patrimonial, la veracidad de la información dirigida a los consumidores y la integridad de los mercados; lo que cambia con los POD no es tanto el bien jurídico protegido como el medio comisivo, que pasa a ser un “engaño por diseño” estructural e interactivo. En este contexto, voy a ceñir mi análisis a tres figuras delictivas que pueden tener especial relevancia en el ámbito empresarial: la estafa del artículo 248 CP, la publicidad engañosa del artículo 282 CP y, en entornos de inversión y trading, los delitos de manipulación de

mercado del artículo 284 CP. Como se ha dicho, hay otros daños que se pueden perpetrar por medio de POD y también podría tenerse en cuenta, en algunos casos, el descubrimiento y revelación de secretos por la responsabilidad derivada del uso ilícito de datos personales o el hacking por algún supuesto acceso no autorizado a sistemas derivado de la manipulación. Pero aquí interesa sobre todo el núcleo de injusto ligado a la manipulación de la voluntad y, por la relación con la responsabilidad de la empresa, lo que puede tener que ver con un desplazamiento patrimonial facilitado por el patrón oscuro digital.

a) La estafa y el engaño por diseño

La figura más intuitivamente conectada con los patrones oscuros es la estafa. El artículo 248 CP exige, en su formulación básica, la concurrencia de un engaño bastante que induzca a error a la víctima, un acto de disposición patrimonial y un perjuicio económicamente apreciable, todo ello presidido por el ánimo de lucro. Nada en la redacción del precepto limita el engaño a manifestaciones verbales o documentales y, de hecho, la jurisprudencia ha admitido desde hace años engaños realizados por medio de artificios técnicos que llevaban a la simulación de circunstancias, la ocultación relevante de información o mediante la creación de contextos idóneos para generar una apariencia falsa de realidad. Como ha señalado Ziermann (2024), diversos patrones oscuros pueden constituir jurídicamente una *acción de engaño* cuando presentan información falsa, sesgada o incompleta como si fuera neutral, fiable o óptima para el usuario —por ejemplo, jerarquías artificiales, autopreferencia algorítmica o preselecciones cuya finalidad real se oculta al consumidor. Desde esta perspectiva, algunas configuraciones de patrones oscuros pueden funcionar como auténticos artificios fraudulentos desde el propio diseño que cumplan con todos los elementos del tipo penal y que claramente estén diseñados dolosamente. Piénsese, por ejemplo, en herramientas que manipulen al usuario en el proceso de compra o suscripción digital para determinarle a realizar un acto de disposición patrimonial por un producto y sólo en el último tiempo, y una vez el sujeto ya lo ha llevado a cabo, condicionen la entrega del bien a un pago mucho mayor. En esos casos o en otros en los que los patrones producen un daño económico directo al usuario (precios más altos, productos menos eficientes o compras no deseadas) que encaja plenamente en el concepto penal de perjuicio patrimonial, podríamos estar ante un delito de estafa (en este mismo sentido, Ziermann, 2024). La cuestión es más compleja en el caso de patrones oscuros más habituales como los que utilizan itinerarios de cancelación opacos (*roach motels*) en los que entrar en una suscripción es fácil e inmediato, pero salir exige atravesar múltiples

pantallas, llamadas o formularios diseñados para disuadir al usuario; o en los procesos de compra con *drip pricing*, en los que el precio final sólo se revela en el último paso, cuando el consumidor ya ha invertido tiempo y esfuerzo y percibe como costoso abandonar la operación pero el sujeto aún no ha pagado nada. Es indiscutible que, en estos casos, el usuario actúa bajo una representación distorsionada de la realidad relevante y que esa distorsión no es casual, sino resultado de un diseño deliberado, pero es difícil entender que en estos casos se dan todos los elementos del tipo penal, en particular, el engaño bastante.

Efectivamente no toda manipulación digital por diseño que pueda resultar ilícita por normas administrativas como la DMA o la DSA y lleve a un acto de disposición patrimonial no querido por el sujeto puede ser considerado estafa. Como sabemos ello dependerá de que se considere que el patrón digital oscuro ha configurado o no un engaño “bastante” (STS 319/2013, de 3 de abril; STS 44/2024 de 17 enero). Y, a mi parecer el elemento esencial y configurador del engaño en este caso estribará en la “oscuridad” del patrón digital y la negación de la autonomía personal que se produce cuando el sistema digital ha determinado ex ante la imposibilidad de la comprensión de su actuar. Es decir, si el diseño se limita a aprovechar inercias de atención pero la información esencial es accesible, estaremos como mucho ante prácticas desleales o sancionables en vía administrativa, pero cuando la configuración de la interfaz genera una apariencia objetivamente falsa (por ejemplo, al presentar como gratuito un servicio que en realidad activa una suscripción periódica de difícil cancelación) y esa apariencia induce a un número relevante de usuarios a disposiciones patrimoniales no deseadas, la idoneidad del engaño es idóneo y si se ha habido disposición patrimonial estaríamos ante una estafa. La clave estará en valorar caso por caso, si el patrón oscuro crea una realidad virtual engañosa que el usuario razonablemente diligente no puede neutralizar: allí donde el diseño se apoya en la opacidad, fragmenta la información económica decisiva y dificulta artificiosamente la reversibilidad de la decisión, el salto de la mera deslealtad al engaño punible aparece⁴⁰.

b) Publicidad engañosa y engaño estructural a consumidores

Han advertido acertadamente Iñigo Corroza & Felip i Saborit (2020) que el delito de publicidad engañosa del artículo 282 del código penal ha caído en desuso y apenas ha dado lugar a resoluciones judiciales de forma complementaria a la aplicación del delito de estafa. Y quizás ello

⁴⁰ El hecho de que la normativa sectorial este sancionando estas prácticas apunta a que empiezan a considerarse desvaloradas y, consiguientemente, fuera del riesgo permitido (véase Pastor Muñoz, 2020; López Barja de Quiroga, 2012).

sea así porque los mecanismos de derecho privado y administrativo resultan suficientes para la protección de los consumidores. Sin embargo, y por lo menos potencialmente, este precepto podría resultar adecuado para responder penalmente a algunos casos de POD en plataformas y servicios de Internet que, no causando un acto de disposición patrimonial individual típico de la estafa, sí generasen un impacto económico significativo mediante la distorsión colectiva de información publicitaria relevante por medio del diseño digital.

Como es sabido, y a diferencia del artículo 248 CP, que se articula en torno al engaño bastante y a la causación de un perjuicio patrimonial concreto, el artículo 282 CP no protege directamente el patrimonio sino la confianza colectiva en la veracidad de la información publicitaria dirigida a los consumidores y usuarios. Más allá de la discusión nominal sobre el bien jurídico protegido⁴¹, existe cierto acuerdo en torno a que los intereses patrimoniales aquí sólo están tutelados de forma indirecta, al sancionarse las conductas que niegan la “expectativa conjunta de los consumidores de poderse orientar sin riesgo mediante los datos suministrados a través de la publicidad por el fabricante o comerciante” (Iñigo Corroza & Felip i Saborit, 2020, pág. 540). También existe acuerdo en que el medio comisivo es la publicidad, definida por la Ley General de publicidad en su artículo 2 como “toda forma de comunicación realizada por una persona física o jurídica, pública o privada, en el ejercicio de una actividad comercial, industrial o profesional, con el fin de promover de forma directa o indirecta la contratación de bienes muebles o inmuebles, servicios, derechos y obligaciones”. Y creo, y esto a los efectos de la potencial incardinación de algunas conductas consistentes en el diseño de POD en el art. 282 va a ser particularmente relevante, que también podría acordarse de que tal bien jurídico se puede ver afectado de múltiples formas por el diseño de plataformas digitales de venta de productos o servicios en las que la “publicidad”, no se presenta como un mensaje externo al producto, sino que se integra en la propia experiencia interactiva. La forma en que se presenta un precio, la inclusión preseleccionada de suplementos, los temporizadores que simulan escasez, las comparativas que omiten opciones más económicas, o los mensajes que exageran la urgencia de compra, constituyen auténticas técnicas publicitarias embebidas en el diseño. No estamos ante meros estímulos estéticos: se trata de decisiones de arquitectura informacional orientadas a modificar la percepción de valor y coste del producto. Así lo entienden también

⁴¹ Habla Santana Vega (2025, pp. 2161) de tres posiciones, la defensa de los derechos económicos de los consumidores; el derecho a la información versa sobre productos y servicios; y las concepciones duales.

Iñigo Corroza & Felipe Saborit (2020, pág. 542). quienes se refieren específicamente a la publicidad algorítmica personalizada por internet”, equiparándola a otras técnicas persuasivas basadas en el contacto individualizado como la venta a domicilio o telefónica, por el hecho de que “es susceptible de ser repetida en un número indefinido de ocasiones”. Efectivamente los diseños integrados en plataformas como prácticas desplegadas de forma masiva y sistemática que llevan a un número elevado de consumidores a sobreestimar las ventajas de un producto o a infravalorar sus costes, pueden considerarse publicidad en este sentido funcional.

Eso sí, para que estos patrones oscuros entren en el ámbito de la sanción penal no bastará que estemos ante una publicidad falsa o engañosa. Es necesario, en primer lugar, que la interfaz o diseño arquitectónico digital contenga algún tipo de información objetivamente falsa, gravemente incompleta o presentada de forma tal que induzca razonablemente a error sobre las características del producto o las circunstancias de su comercialización. Si el diseño lo que hace es no dar toda la información del producto o servicio, pero sin que nada de lo que se diga sea falso, y por mucho que induzca a error al consumidor al generarle una percepción del producto que no es cierta, no entraremos en el ámbito de aplicación del artículo 282. Eso sí, no será necesario que se falseen todas las condiciones del producto, sino que bastará que al hacerlo con alguna de ellas el sentido de la comunicación publicitaria pueda considerarse falso. En segundo lugar, el delito del artículo 282 CP exige que la conducta sea apta para poder causar un perjuicio grave y manifiesto a los consumidores. Aquí radica el verdadero filtro penal también aplicable a los POD: la manipulación debe poseer una entidad tal que comprometa la confianza colectiva o produzca un desvío económico relevante. Conforme a esto el delito de publicidad engañosa no exige la consumación del perjuicio patrimonial individual, pero sí la idoneidad objetiva del mensaje para causar un perjuicio grave y manifiesto a un conjunto indeterminado de destinatarios, enfatizando que la valoración debe hacerse en función del impacto potencial de la campaña, no de los efectos concretos sobre cada consumidor.

Si trasladamos conjuntamente estos criterios al ámbito de los POD, la conclusión es clara: sólo aquellos diseños manipulativos que alteren sustancialmente la comprensión de las condiciones económicas de un servicio, que afecten potencialmente a un colectivo amplio, y que formen parte de una estrategia empresarial sistemática podrán considerarse publicidad engañosa penalmente relevante. Cuando la práctica se limite a inducir incomodidad, empujar a decisiones menos óptimas

o aprovechar sesgos cognitivos sin distorsionar la información esencial, nos encontraremos como mucho ante infracciones administrativas, por más que resulten reprochables desde un punto de vista ético o regulatorio. El umbral penal se alcanza únicamente cuando la manipulación se convierte en un mecanismo estructural de distorsión del mercado que compromete efectivamente bienes jurídicos colectivos.

c) Manipulación y abuso de mercado y gamificación del riesgo financiero

Entre los tipos que pueden activarse en contextos de patrones oscuros digitales se encuentra también el delito de manipulación de precios del artículo 284 CP. La manipulación algorítmica del entorno competitivo, cuando altera precios, privilegia artificialmente ciertos productos, genera señales distorsionadas o limita la comparabilidad real entre alternativas, puede constituir una forma funcionalmente equivalente a las prácticas clásicas de manipulación de precios (véase también Ziermann, 2024). A diferencia de la estafa o la publicidad engañosa, centradas en relaciones bilaterales, este delito protege el correcto funcionamiento del mercado como institución, asegurando que las señales (precios, tendencias, volúmenes) que orientan las decisiones económicas no resulten adulteradas mediante artificios. En esta línea, la doctrina ha subrayado que el bien jurídico del artículo 284 CP no es el patrimonio individual sino la libre competencia, entendida como libre concurrencia de productos, servicios, valores o instrumentos financieros y, en concreto, como mecanismo de fijación de los precios en el mercado (Santana Vega, 2024, p. 1276)⁴².

La configuración típica del art. 284 CP se puede sistematizar en dos figuras delictivas distintas, en el apartado 1^o el clásico delito de “maquinaciones para alterar el precio de las cosas”, y en el apartado 2^o y 3^o formas de manipulaciones de mercados financieros. A mi parecer son estos últimos los que probablemente más podrían cuadrar potencialmente con alguna de las conductas que estamos analizando. Al fin y al cabo, y tal y como señala Estrada Cuadras (2023), con el tipo del apartado primero se pretende prevenir la puesta en peligro de los intereses patrimoniales de cualquier interviniente en el mercado, por lo que la conducta manipulativa debería alterar el precio del producto o servicio en el mercado correspondiente y es difícil pensar en patrones oscuros que lleven a una alteración general

⁴² Así lo establece también el propio TS en STS 369/2019 al señalar que el precepto no está “orientado a la protección del patrimonio individual, sino a garantizar la transparencia, fiabilidad y seguridad del mercado”, tal y como referencia (Estrada Cuadras, 2023, p. 264) quien, sin embargo, señala que la protección del mercado no tiene sentido sino es para “preservar los intereses patrimoniales de quienes intervienen en él”.

de los precios de un servicio digital⁴³. Por el contrario, más posible es plantearnos el uso de POD para la perpetración de manipulaciones del mercado, bien manipulaciones informativas, en las que se altera el precio de cotización de un instrumento financiero a través de la difusión en el mercado de información falsa, o bien operativas, en las que el precio se altera mediante actos de negociación ficticios o reales.

En cuanto a las manipulaciones informativas, el artículo 284.2 cita expresamente internet o las TIC como medio a través del cual se difundieren noticias o rumores o se transmitieran señales falsas o engañosas sobre personas empresas, ofreciendo a sabiendas datos económicos total o parcialmente falsos con el fin alterar o preservar el precio financiero. Desde esta perspectiva, nada impide considerar que la interfaz gamificada de una plataforma de inversión cuando utiliza patrones oscuros, señales conductuales o manipulaciones algorítmicas de rankings, buscadores o sistemas de recomendación, pueda ser utilizada para engañar y dar apariencia de mercado que no responde a la realidad económica subyacente, ya sea a través de operaciones simuladas, difusión de rumores o utilización de otros procedimientos que produzcan señales irreales. Eso sí, será necesario que el diseño simulase, exagerase o presentase información falsa o de forma engañosa, por ejemplo: cuando se ofrece como indicio de tendencia, seguridad o fiabilidad lo que, en realidad, son dinámicas internas de la propia plataforma; cuando se genera la impresión de un movimiento relevante de demanda sin soporte económico externo; o cuando se muestran señales visuales o agregadas que se presentan como datos de mercado sin serlo. Como señala Ziermann (2024), en entornos digitales la frontera entre persuasión comercial legítima y manipulación del mercado puede cruzarse cuando los incentivos económicos del proveedor dependen de la intensificación de la actividad especulativa del usuario. Eso sí, no toda mejora estética o gamificada de una interfaz integraría el tipo penal: es necesario que el patrón oscuro simule una demanda inexistente, oculte el riesgo real de un producto, sobrerrepresente el éxito o la frecuencia de operaciones ajenas, o presente como “tendencias del mercado” lo que son simples resultados del algoritmo interno.

Pero todo ello en forma de difusión de “datos económicos total o parcialmente falsos” que, además, deben ser, naturalmente, relevantes para el precio de cotización. Para que la conducta sea penalmente relevante, y no una mera infracción administrativa, la doctrina señala que las señales generadas creen apariencias no vinculadas a datos reales

⁴³ Así lo ha establecido también la jurisprudencia del STS 11/83.

o se basen en información sesgada, incompleta o incomprensible para un inversor medio, superando así los límites del riesgo permitido (Muñoz de Morales Romero, 2023). Solo en tales casos podría sostenerse que el diseño digital altera ilegítimamente las condiciones de mercado y activa la tipicidad del art. 284 CP. Como señala Estrada Cuadras (2023) esta exigencia es la parte de delimitación entre el injusto penal frente al ilícito administrativo derivado del mayor desvalor de conducta en sentido objetivo. Por eso si lo que se difunde es información falsa de otro tipo y no sobre el precio o esencial para él podríamos estar ante infracciones administrativas o no penal.

Por último, la doctrina exige la idoneidad ex ante del artificio para alterar el comportamiento del inversor, sin que sea necesario que se produzca el efectivo engaño (Feijoo Sánchez, 2019). Muñoz de Morales Romero (2023) insiste en que esta valoración debe hacerse atendiendo al tipo de inversor afectado, al contexto en que opera y a la estructura de la señal generada. En el caso de plataformas dirigidas fundamentalmente a inversores minoristas inexpertos, especialmente sensibles a estímulos visuales, refuerzos inmediatos y señales agregadas simplificadas (vulnerabilidades que han subrayado los informes recientes de la CNMV sobre inversores minoristas), un diseño gamificado que incrementa artificialmente la confianza en activos de riesgo mediante confeti digital, mensajes de “racha ganadora”, indicadores agregados del tipo “el 80 % de nuestros usuarios está comprando este producto” o ránquines de “inversores de éxito” contruidos con criterios opacos puede considerarse apto para influir de forma sistemática en decisiones de inversión.

d) La responsabilidad por estos y otros delitos de la persona jurídica

Todas estas y otras figuras delictivas pueden ser perpetradas en beneficio directo o indirecto de la empresa y mediante decisiones de diseño adoptadas por sus órganos de dirección o por empleados sometidos a su autoridad. Es en este caso cuando, conforme establece el artículo 31 bis CP, se activaría la responsabilidad penal de las personas jurídicas. Hemos visto anteriormente como un elemento característico de los patrones oscuros es su dimensión organizativa: son el resultado de cadenas de decisiones en las que se definen objetivos de negocio, se seleccionan métricas de conversión, se experimentan variantes de interfaz y se consolidan aquellas que generan mayor rentabilidad, aun cuando ello suponga incrementar el error de los usuarios. Desde la lógica del artículo 31 bis, esta forma de actuar puede interpretarse como un defecto de organización penalmente relevante cuando la empresa carece de controles efectivos para impedir

que esa búsqueda de rendimiento derive en prácticas manipulativas lesivas, o, más gravemente, cuando los incentivos internos favorecen precisamente aquellas configuraciones de diseño que incrementan el perjuicio del usuario.

En estos casos, el incumplimiento no se limita a la ausencia de protocolos formales adecuados como programas de cumplimiento normativo eficaces, sino que se manifiesta en la existencia de una cultura corporativa orientada a la explotación del error, donde el diseño engañoso se integra como práctica normalizada. Esta orientación estratégica constituye un supuesto paradigmático de defecto teleológico de organización, en el sentido de que la compañía estructura su funcionamiento hacia fines incompatibles con la legalidad. No estamos ante un fallo accidental de supervisión, sino ante una forma de gobernanza que convierte la manipulación en una herramienta de negocio. Ello supone que más allá de la identificación de quienes han participado del diseño de los patrones oscuros, la imputación a la persona jurídica requiere demostrar que la empresa sabía o podía saber (a través de la trazabilidad de experimentos A/B, análisis de datos, reportes internos y métricas de retención) que sus decisiones de diseño generaban errores previsibles y, pese a ello, decidió mantenerlas. En tales circunstancias, la estructura organizativa sustituye al dolo individual: la empresa responde penalmente porque ha creado un entorno donde el delito no sólo es posible, sino probable y rentable.

Por ello, la responsabilidad penal corporativa en materia de patrones oscuros no se fundamenta en una ampliación artificiosa del artículo 31 bis, sino en su aplicación coherente a un contexto en el que el riesgo nace del propio modelo de negocio. Allí donde la organización integra la manipulación digital en sus procesos de decisión, evaluación y recompensas internas, se configura el tipo de defecto estructural que convierte a la persona jurídica en sujeto adecuado de imputación penal.

3.4. Programas de cumplimiento y prevención de los POD lesivos

La progresiva consolidación de un estándar de diligencia ex ante en la regulación europea permite esbozar, con bastante precisión, cómo deberían configurarse los programas de cumplimiento orientados a prevenir los POD. Como se vio en el apartado anterior, el marco normativo (desde el RGPD hasta el DSA, el DMA y, sobre todo, el AI Act) ha desplazado la atención desde la corrección posterior del abuso hacia la anticipación estructural del riesgo. Este giro refleja una idea fundamental: la manipulación digital no es un accidente del diseño, sino un riesgo inherente a la propia forma de organizar la interacción en las plataformas, y por ello debe ser identificada,

evaluada y neutralizada desde el proceso mismo de concepción del producto.

Este principio de diligencia estructural tiene dos implicaciones decisivas para el Derecho penal económico. En primer lugar, fija un umbral claro de cuidado debido, un estándar normativo que permite distinguir entre errores accidentales y prácticas que, por su reiteración o previsibilidad, revelan una política organizativa orientada a aprovechar la vulnerabilidad decisional del usuario. En segundo lugar, genera un ecosistema documental que transforma el diseño digital en una actividad trazable: evaluaciones de impacto, informes internos, registros de experimentación A/B, justificaciones de decisiones de interfaz, trazabilidad de los cambios y auditorías algorítmicas. Estos materiales permiten reconstruir, en caso de controversia, si la empresa actuó de manera coherente con las obligaciones de diligencia o si, por el contrario, mantuvo o incentivó configuraciones que el derecho europeo identifica como manipulativas.

En este sentido, el cumplimiento administrativo no sólo orienta la conducta empresarial, sino que delimita el umbral penal de manera indirecta. Allí donde la normativa obliga a justificar las decisiones de diseño, evaluar el riesgo de manipulación, garantizar la transparencia de las opciones y documentar los experimentos, la persistencia en la utilización de patrones oscuros deja de ser una mera infracción regulatoria para convertirse en indicio de un defecto organizativo penalmente relevante. Si una organización, pese a advertencias de supervisores o evidencias empíricas internas, mantiene diseños que erosionan sistemáticamente la autonomía del usuario, ello puede interpretarse como la institucionalización del error ajeno como fuente de beneficio. Y para que esto suceda no es necesario que haya una intención explícita, como han mostrado algunos artículos recientes, como muestran estudios empíricos recientes, incluso sistemas de IA interactuando con interfaces gráficas pueden generar o reforzar patrones oscuros sin intención humana alguna, simplemente porque los modelos priorizan la finalización de tareas o la maximización de recompensas sobre la protección del usuario (Tang et al., 2025; véase también Sparr, 2022). Esto es clave para el entendimiento de cómo debería diseñarse los programas de cumplimiento frente a los patrones oscuros digitales: si la manipulación puede ser una tarea técnica realizada autónomamente por un *software*, pero derivada de una filosofía de empresa y no solo una decisión humana, deberán supervisar los procesos de experimentación, el *tuning* algorítmico y los test A/B como posibles fuentes autónomas de prácticas manipulativas.

Desde esta perspectiva, un programa de cumplimiento eficaz debe

partir de una premisa distinta a la del *compliance* tradicional: la prevención de la manipulación sólo es posible si se interviene en el diseño, en las métricas y en los incentivos que gobiernan la producción digital. El artículo 31 bis CP no exige sólo la existencia de protocolos, sino la adopción de modelos de organización y gestión que resulten adecuados para prevenir la comisión de delitos. En el entorno digital, esto implica repensar el *compliance* desde una lógica de “ética del diseño”, integrando en el ciclo de vida del producto criterios de respeto a la autonomía del usuario. Teniendo en cuenta los reglamentos europeos de servicios digitales y del mercado digital me atreveré a continuación a establecer las bases de un marco normativo que trate de evitar los POD atendiendo a las distintas fases del diseño digital persuasivo. El primer paso, en la fase de diseño, consistiría en incorporar dentro de los programas de cumplimiento evaluaciones de impacto sobre la autonomía del usuario, similares a las evaluaciones de impacto ambiental o de protección de datos, en las que se analice si una decisión de interfaz (una jerarquía visual, un flujo de navegación, una opción preseleccionada o una dinámica de gamificación) puede generar fricciones asimétricas, inducir decisiones impulsivas o dificultar la reversibilidad. Esta evaluación serviría para obligar a justificar por qué se elige una configuración frente a alternativas menos lesivas. En segundo lugar, en la fase de experimentación, la supervisión podría extenderse a los test A/B y a los procesos de optimización algorítmica. La experimentación es consustancial a la innovación digital, pero un sistema de *compliance* sólido debería evitar que las métricas de conversión o retención actúen como incentivos criminógenos. Para ello, la empresa debería definir límites internos claros: qué se puede probar, sobre qué colectivos, durante cuánto tiempo y con qué umbrales de transparencia. Igual que en el ámbito biomédico existe un marco ético para la experimentación con sujetos humanos, la experimentación con la vulnerabilidad cognitiva del usuario exige una análoga estructura de control. Finalmente, en la fase de control, los criterios de éxito deberían incorporar indicadores de integridad. No basta con medir consumo, permanencia o clics: el cumplimiento exige supervisar si el usuario mantiene control sobre sus decisiones, si entiende la información relevante, si la plataforma permite cancelar sin fricciones excesivas o si las dinámicas de incentivo erosionan la capacidad de deliberación. Al fin y al cabo, como recuerdan Isola y Esposito (2025), el marco normativo europeo se basa en la capacidad objetiva de la práctica para distorsionar el comportamiento del consumidor medio (Isola y

Esposito, 2025, pp. 1–2)⁴⁴. Esto implica que el cumplimiento normativo en materia de diseño digital no puede limitarse a acreditar ausencia de una supuesta intención manipuladora, sino que debe demostrar que la organización ha desplegado mecanismos preventivos capaces de evitar configuraciones que, aunque no intencionadas, produzcan efectos manipulativos.

En un modelo así el *compliance officer* adquiriría un papel más intenso que en los esquemas tradicionales. Ya no es sólo un intérprete de la norma, sino un mediador entre legalidad y diseño, una figura capaz de plantear preguntas sobre el efecto conductual de cada decisión de producto. Para ello necesita independencia funcional y competencias técnicas híbridas: conocimientos de psicología del comportamiento, diseño de interacción, analítica de datos y ética de la IA. La manipulación digital no se previene sólo con códigos de conducta, sino con capacidad real para intervenir en los procesos de producto.

No negaré que la implantación efectiva de modelos de cumplimiento de estas características en el ecosistema digital actual parece ilusorio o, al menos, naif. Al fin y al cabo, ningún programa de cumplimiento será eficaz sin un cambio cultural interno. Y previamente sería necesario desplazar la cultura organizativa de las empresas digitales que diseñan patrones oscuros hacia la transparencia, la autonomía y la integridad, de modo que estas dejen de ser requisitos regulatorios para convertirse en activos estratégicos. Pero lo cierto es que la prevención de los POD sólo puede lograrse si la empresa integra la diligencia ex ante en la arquitectura misma del diseño, sustituyendo el modelo reactivo por una ética institucional de anticipación. La literatura empírica nos está empezando a mostrar los rasgos y daños de los POD; la regulación europea ha proporcionado el marco y el enfoque de prevención necesario para evitarlo; corresponde ahora, quizás, comenzar a aplicar las sanciones y correspondería, también, a los programas de cumplimiento transformar las exigencias éticas de un mundo digital construido para la interacción autónoma de los usuarios en práctica organizativa real. Porque si la manipulación es un producto del diseño, la prevención debe empezar donde nace el diseño.

BIBLIOGRAFÍA

Barrio Andrés, M. (2023). El cumplimiento basado en el riesgo o «risk-based compliance»: Pieza cardinal del nuevo Derecho digital

⁴⁴ Señalan los citados autores que la clasificación jurídico-material basada en la UCPD (*misleading omission, misleading action, harassment, undue influence* y coerción) ofrece un mapa normativo útil para orientar la creación de métricas internas, auditorías y evaluaciones de impacto que permitan detectar ex ante patrones potencialmente ilícitos.

europeo. *Análisis del Real Instituto Elcano (ARI)*, 34, 1.

Brignull, H. (2010). Dark patterns. {<https://darkpatterns.org/>} (Último acceso: 29 de noviembre de 2025)

Büchi, M., Festic, N., & Latzer, M. (2022). The Chilling Effects of Digital Dataveillance: A Theoretical Model and an Empirical Research Agenda. *Big Data & Society*, 9(1), 20539517211065368. {<https://doi.org/10.1177/20539517211065368>}

CEA Institute. (2022). *Fun and Games: Investment Gamification and Implications for Capital Markets*. {<https://doi.org/10.56227/22.1.17>}

Chang, W. J., Seaborn, K., & Adams, A. A. (2024). Theorizing Deception: A Scoping Review of Theory in Research on Dark Patterns and Deceptive Design. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1-7. {<https://doi.org/10.1145/3613905.3650997>}

Cialdini, R. B. (1984). *Influence. The Psychology of Persuasion*. HarperCollins. {<https://www.scirp.org/reference/ReferencesPapers?ReferenceID=79870>}

Cialdini, R. B. (2007). *Influence: The Psychology of Persuasion*.

Cohen, J. E. (2012). *Configuring the Networked Self*. Yale University Press. {<https://www.jstor.org/stable/j.ctt5vm24d>}

Comisión Europea (2020). *Unfair commercial practices directive*. {https://commission.europa.eu/law/law-topic/consumer-protection-law/unfair-commercial-practices-and-price-indication/unfair-commercial-practices-directive_en}

Comisión Europea (2025). *Directrices sobre prácticas prohibidas de inteligencia artificial (IA), tal como se definen en la Ley de IA*. {<https://digital-strategy.ec.europa.eu/es/library/commission-publishes-guidelines-prohibited-artificial-intelligence-ai-practices-defined-ai-act>}

Couldry, N., & Mejias, U. A. (2019). *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*.

Davenport, T. H., & Beck, J. C. (2001). *The Attention Economy: Understanding the New Currency of Business*.

ESMA. (2023). *ESMA report on trends, risks and vulnerabilities No. 2*. {https://www.esma.europa.eu/document/esma-report-trends-risks-and-vulnerabilities-no-2-2023?utm_source=chatgpt.com}

Estrada Cuadras, A. (2023). Protección de la propiedad intelectual e industrial, del mercado y de los consumidores. *Leciones de derecho penal económico y de la empresa: Parte general y especial, 2023, ISBN 978-84-19773-45-6, págs. 587-658, 587-658*. {<https://dialnet.unirioja.es/servlet/articulo?codigo=9059793>}

European Parliament and Council. (2022). *Digital Markets Act*. {https://digital-markets-act.ec.europa.eu/index_en}

European Banking Authority. (2024a). *Risk assessment report—July 2024*. {<https://www.eba.europa.eu/publications-and-media/publications/risk-assessment-report-july-2024>}

European Banking Authority. (2024b). *Risk Assessment Report—November 2024*. {<https://www.eba.europa.eu/publications-and-media/publications/risk-assessment-report-november-2024>}

European Data Protection Board. (2022). *Dark patterns in social media platform interfaces: How to recognise and avoid them*. {https://www.edpb.europa.eu/our-work-tools/documents/public-consultations/2022/guidelines-32022-dark-patterns-social-media_en}

European Parliament and Council. (2022). *Digital Services Act*. {<https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package>}

European Parliament and Council. (2024). *EU Artificial Intelligence Act*. {<https://artificialintelligenceact.eu/>}

European Parliament and Council. (2016). *Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation)*. Official Journal of the European Union, L 119, 1–88.

Federal Trade Commission. (2022). *Bringing Dark Patterns to Light*. {<https://www.ftc.gov/reports/bringing-dark-patterns-light>}

Feijoo Sánchez. (2019). La manipulación informativa en los delitos bursátiles: Engaño, artificio e idoneidad ex ante. *Revista de Derecho Penal y Criminología*, 22, 115–148.

Ferster, C. B., & Skinner, B. F. (1957). *Schedules of reinforcement* (pp. vii, 744). Appleton-Century-Crofts. {<https://doi.org/10.1037/10627-000>}

Fineberg, N. A., Menchón, J. M., Hall, N., Dell’Osso, B., Brand, M., Potenza, M. N., Chamberlain, S. R., Cirnigliaro, G., Lochner, C., Billieux, J., Demetrovics, Z., Rumpf, H. J., Müller, A., Castro-Calvo, J., Hollander, E., Burkauskas, J., Grünblatt, E., Walitza, S., Corazza, O., ... Zohar, J. (2022). Advances in problematic usage of the internet research—A narrative review by experts from the European network for problematic usage of the internet. *Comprehensive Psychiatry*, 118, 152346. {<https://doi.org/10.1016/j.comppsy.2022.152346>}

González Tapia, M. I. (2025). Persuasión algorítmica y adicción tecnológica: Las nuevas tabaqueras. *González Tapia, M.I., «Persuasión algorítmica y adicción tecnológica: las nuevas tabaqueras», Revista de Victimología/Journal of Victimology*, n 19, 2025, pp. 43–76. {<https://doi.org/10.12827/RVJV.19.02>}. {<http://helvia.uco.es/xmlui/handle/10396/33645>}

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L. (2018). The Dark (Patterns) Side of UX Design. *Proceedings of the*

2018 CHI Conference on Human Factors in Computing Systems, 1-14. {<https://doi.org/10.1145/3173574.3174108>}

Gray, C. M., Santos, C., Bielova, N., Toth, M., & Clifford, D. (2021). Dark Patterns and the Legal Requirements of Consent Banners: An Interaction Criticism Perspective. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1-18. {<https://doi.org/10.1145/3411764.3445779>}

Gray, C. M., Santos, C. T., Bielova, N., & Mildner, T. (2024). An Ontology of Dark Patterns Knowledge: Foundations, Definitions, and a Pathway for Shared Knowledge-Building. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1-22. {<https://doi.org/10.1145/3613904.3642436>}

Györy, C. (2020). The institutional context of financial fraud in a post-transition economy: The Quaestor scandal. *European Journal of Criminology*, 17(1), 31-49. {<https://doi.org/10.1177/1477370819874436>}

Hadan, H., Sgandurra, S. A., Zhang-Kennedy, L., & Nacke, L. E. (2024). From Motivating to Manipulative: The Use of Deceptive Design in a Game's Free-to-Play Transition. *Proceedings of the ACM on Human-Computer Interaction*, 8(CHI PLAY), 1-31. {<https://doi.org/10.1145/3677074>}

Hueso, L. C., & Castellano, P. S. (2024). *Tratado sobre el reglamento de inteligencia artificial de la unión Europea*. Aranzadi. {<https://dialnet.unirioja.es/servlet/libro?codigo=993411>}

Iñigo Corroza, M. E., & Felipe Saborit, D. (2020). Protección de los consumidores. *Lecciones de derecho penal económico y de la empresa: Parte general y especial, 2023, ISBN 978-84-19773-45-6, págs. 509-551*, 509-551. {<https://dialnet.unirioja.es/servlet/articulo?codigo=9059795>}

Isola, C., & Esposito, F. (2025). A systematic literature review on dark patterns for the legal community: Definitional clarity-and a legal classification based on the Unfair Commercial Practices Directive. *Computer Law & Security Review*, 58. {<https://doi.org/10.1016/j.clsr.2025.106169>}

King, D., Delfabbro, P., Gainsbury, S., Dreier, M., Greer, N., & Billieux, J. (2019). Unfair play? Video games as exploitative monetized services: An examination of game patents from a consumer protection perspective. *Computers in Human Behavior*, 101, 131-143. {<https://doi.org/10.1016/j.chb.2019.07.017>}

Klocke, N., Müller-Okesson, D., Hasso, T., & Pelster, M. (2025). The impact of peer returns in social trading. *Journal of Behavioral and Experimental Finance*, 46, 101057. {<https://doi.org/10.1016/j.jbef.2025.101057>}

Koh, W. C., & Seah, Y. Z. (2023). Unintended consumption: The effects of four e-commerce dark patterns. *Cleaner and Responsible Consumption*, 11, 100145. {<https://doi.org/10.1016/j.clrc.2023.100145>}

Kollmer, T., & Eckhardt, A. (2022). Dark Patterns: Conceptualization and Future Research Directions. *Business & Information Systems Engineering*, 65. {<https://doi.org/10.1007/s12599-022-00783-7>}

Kollmer, T., & Eckhardt, A. (2023). Dark Patterns. *Business & Information Systems Engineering*, 65(2), 201-208. {<https://doi.org/10.1007/s12599-022-00783-7>}

Lee, H.-P. (Hank), Chiang, Y.-S., Gao, L., Yang, S., Winter, P., & Das, S. (2025). Purpose Mode: Reducing Distraction through Toggling Attention Capture Damaging Patterns on Social Media Web Sites. *ACM Trans. Comput.-Hum. Interact.*, 32(1), 10:1-10:41. {<https://doi.org/10.1145/3711841>}

López Barja de Quiroga, J. (2012). Deberes de autoprotección en el delito de estafa. *Revista Aranzadi Doctrinal*, 4 (julio 2012), 175-181.

Luguri, J., & Strahilevitz, L. J. (2021). Shining a Light on Dark Patterns. *Journal of Legal Analysis*, 13(1), 43-109. {<https://doi.org/10.1093/jla/laaa006>}

Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019). Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), 81:1-81:32. {<https://doi.org/10.1145/3359183>}

Mathur, A., Kshirsagar, M., & Mayer, J. (2021). What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1-18. {<https://doi.org/10.1145/3411764.3445610>}

Mato Pacín, M. N. (2024). ASPECTOS JURÍDICOS DEL DISEÑO DE LAS INTERFACES DIGITALES. EN ESPECIAL, LOS PATRONES OSCUROS. *DERECHO PRIVADO*. {https://www.boe.es/biblioteca_juridica/publicacion.php?id=PUB-PR-2024-330}

Miró Llinares, F., & Johnson, S. D. (2018). Cybercrime and Place: Applying Environmental Criminology to Crimes in Cyberspace. En G. J. N. Bruinsma & S. D. Johnson (Eds.), *The Oxford Handbook of Environmental Criminology* (p. o). Oxford University Press. {<https://doi.org/10.1093/oxfordhb/9780190279707.013.39>}

Muñoz de Morales Romero. (2023). *El delito de manipulación del mercado de valores* | Marcial PONS. {<http://www.marcialpons.es/libros/el-delito-de-manipulacion-del-mercado-de-valores/9788429028027/>}

OECD. (2022a). *Dark commercial patterns*. OECD Publishing. {https://www.oecd.org/en/publications/dark-commercial-patterns_-44f5e846-en.html}

OECD. (2022b). Enhancing online disclosure effectiveness. *OECD*

Digital Economy Papers. {<https://doi.org/10.1787/6d7ea79c-en>}

Oksanen, A., Mantere, E., Vuorinen, I., & Savolainen, I. (2022). Gambling and online trading: Emerging risks of real-time stock and cryptocurrency trading platforms. *Public Health*, 205, 72-78. {<https://doi.org/10.1016/j.puhe.2022.01.027>}

Orujov, A. (2023). *Trading as Gambling During Covid-19 Lockdown* (SSRN Scholarly Paper No. 4423095). Social Science Research Network. {<https://doi.org/10.2139/ssrn.4423095>}

Pastor Muñoz, N. (2020). El delito de estafa. *Lecciones de derecho penal económico y de la empresa: Parte general y especial*, 2023, ISBN 978-84-19773-45-6, págs. 271-318, 271-318. {<https://dialnet.unirioja.es/servlet/articulo?codigo=9059800>}

Rossi, A., Carli, R., Botes, M. W., Fernandez, A., Sergeeva, A., & Sánchez Chamorro, L. (2024). Who is vulnerable to deceptive design patterns? A transdisciplinary perspective on the multi-dimensional nature of digital vulnerability¹. *Computer Law & Security Review*, 55, 106031. {<https://doi.org/10.1016/j.clsr.2024.106031>}

Santana Vega, D. M. (2024). Delitos contra el mercado y los consumidores [comentarios a los arts. 278 a 289]. En *Comentarios al Código Penal: Reformas LLOO 1-2023, 3-2023 y 4-2023 / Mirentxu Corcoy Bidasolo, Santiago Mir Puig (dirs.), Guillermo Ramírez Martín, Gabriel Rogé Martín (coords.). Título XIII Delitos contra el patrimonio y contra el orden socioeconómico. Capítulo XI De los delitos relativos a la propiedad intelectual e industrial, al mercado y a los consumidores. Sección 3 de los delitos relativos al mercado y a los consumidores: Arts. 278-285. Sección 4 Delitos de Corrupción en los negocios: Arts. 286-289, p. 1247-1316*, (2024). Tirant lo Blanch (Grupo editorial Tirant lo Blanch). {<https://accedacris.ulpgc.es/jspui/handle/10553/136486>}

Shover, Neal, & Hochstetler. (2007). Theoretical Integration in the Study of Corporate Crime. En H. Pontell, & G. Geis (Eds.), *International Handbook of White-Collar and Corporate Crime* (pp. 37-58). Springer.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis* (p. 457). Appleton-Century.

Skinner, B. F. (1953). *Science and Human Behavior*.

Sparr, M. (2022). *Explicit User Manipulation in Reinforcement Learning Based Recommender Systems* (No. {arXiv:2203.10629}). arXiv. {<https://doi.org/10.48550/arXiv.2203.10629>}

Strycharz, J., Zhang, D., & Segijn, C. M. (2025). Dataveillance in Interactive Advertising: Transforming Consumer-Advertiser Interactions and Its Unintended Consequences. *Journal of Interactive Advertising*, 0(0), 1-13. {<https://doi.org/10.1080/15252019.2025.2547850>}

Susser, D., Roessler, B., & Nissenbaum, H. (2018). *Online Ma-*

nipulation: Hidden Influences in a Digital World (SSRN Scholarly Paper No. 3306006). Social Science Research Network. {<https://doi.org/10.2139/ssrn.3306006>}

Tang, J., Chen, C., Jiawen, L., Zhang, Z., Guo, B., Khalilov, I., A Gebreegziabher, S., Yao, B., Wang, D., Ye, Y., Li, T., Xiao, Z., Yao, Y., & Li, T. J.-J. (2025). *Dark patterns meet gui agents: Llm agent susceptibility to manipulative interfaces and the role of human oversight* (No. {arXiv:2509.10723}). arXiv preprint.

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving Decisions About Health, Wealth, and Happiness*.

Van de Bunt, H. (2007). Organizational Crime. En H. Pontell, & G. Geis (Eds.), *International Handbook of White-Collar and Corporate Crime* (pp. 37-58). Springer.

Yeung, K. (2017). 'Hypernudge': Big Data as a mode of regulation by design. *Information, Communication & Society*, 20(1), 118-136. {<https://doi.org/10.1080/1369118X.2016.1186713>}

Zagal, J. P., Björk, S., & Lewis, C. (2013). *Dark Patterns in the Design of Games*. Foundations of Digital Games 2013. {<https://urn.kb.se/resolve?urn=urn:nbn:se:ri:diva-24252>}

Zendle, D., & Cairns, P. (2018). Video game loot boxes are linked to problem gambling: Results of a large-scale survey. *PLOS ONE*, 13(11), e0206767. {<https://doi.org/10.1371/journal.pone.0206767>}

Zhang, G. X., Wang, Y., Nakajima, T. L., & Seaborn, K. (2025). First Contact with Dark Patterns and Deceptive Designs in Chinese and Japanese Free-to-Play Mobile Games. *Proceedings of the ACM on Human-Computer Interaction*, 9(6), 730-755. {<https://doi.org/10.1145/3748620>}

Zhang, Y., Zhou, J., Wang, F., Chen, Y., Zhou, X., Yan, Y., & Luo, J. (2024). The Impact of Cyberbullying Victimization on Internet Gaming Addiction Among College Students: The Mediating Roles of Basic Psychological Need Satisfaction and Frustration, and the Moderating Role of Parental Autonomy Support. *PSYCHOLOGY RESEARCH AND BEHAVIOR MANAGEMENT*, 17, 4105-4118. {<https://doi.org/10.2147/PRBM.S486250>}

Zhang, Z., Moradzadeh, S., Gui, X., & Kou, Y. (2025). More Than Just Microtransactions: Predatory Monetization in User-Generated Games. *Proceedings of the ACM on Human-Computer Interaction*, 9. {<https://doi.org/10.1145/3748626>}

Ziermann, F. (2024). Dark patterns: Can criminal law remedy the shortcomings of antitrust law? *New Journal of European Criminal Law*, 15(4), 388-411. {<https://doi.org/10.1177/20322844241298196>}

Zuboff, S. (2015). Big other: Surveillance Capitalism and the Prospects of an Information Civilization. *Journal of Information*

Technology, 30(1), 75-89. {<https://doi.org/10.1057/jit.2015.5>}

Zuboff, S. (2019). *The Age Of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power: Barack Obama's Books of 2019*.