Revista Penal México 16-17
Doctrina
Núms. 16 y 17, marzo de 2019-febrero de 2020

Inteligencia artificial y responsabilidad penal

Paz M. de la Cuesta Aguado *Universidad de Cantabria*

Resumen: *La aparición de agentes inteligentes artificiales que pueden realizar conductas socialmente significativas exige prever un código normativo que determine en qué condiciones es admisible la interacción social de tales agentes. El derecho penal está en condiciones de afrontar los problemas iniciales que esa cuestión exige. El código que contiene las instrucciones que regulan la actuación del ente artificial presenta más similitudes de las aparentes con los códigos de conducta que rigen el comportamiento de las personas; y cuando la actuación de un agente artificial pueda ser delictiva, el código no puede obviar los mandatos y prohibiciones penales. El grado de inteligencia que permita superar el "umbral de responsabilidad" del agente artificial inteligente no es una cuestión que corresponda determinar exclusivamente a los programadores.*

Palabras clave: *inteligencia artificial, responsabilidad penal, software, culpa in vigilando, fuentes de peligro.*

Abstract: *The emergence of artificial intelligent agents that can perform behaviors socially significant demands require a normative code that determines under what conditions the social interaction of such agents is admissible. Criminal law is able to face the initial problems that that question demands. The code that contains the instructions that regulate the performance of the artificial entity shows more similarities than the apparent ones with the codes of conduct that govern the behavior of the people; and when the action of an artificial agent can be criminal, the code cannot ignore the criminal mandates and prohibitions. The degree of intelligence that allows the intelligent artificial agent to "exceed the threshold of responsibility" is not a question that should be exclusively determined by the programmers.*

Key words: *artificial intelligent; criminal liability; software, blame on monitoring, sources of danger.*

Rec: 13-07-2019 | **Fav:** 27-07-2019

1. Inteligencia, inteligencia artificial y otros conceptos[1]

I. No es objeto de este trabajo definir qué es la inteligencia, ni siquiera posicionarse sobre ello. Entre otras razones porque el concepto de inteligencia, como característica propia del género humano, dista mucho de ser pacífico, tanto en cuanto a su contenido (¿qué es la inteligencia?) como a su extensión (¿solo son inteligentes los humanos?). No obstante, y con independencia de las dificultades para definir el concepto de *inteligencia* (*natural*), a efectos de nuestro trabajo, y con carácter preliminar, definiremos la *inteligencia* a partir de las siguientes habilidades básicas: 1) capacidad para recibir información, 2) capacidad para entender la información recibida —entendiendo por tal la capacidad para hacer utilizable por el agente, en el contexto y en función de objetivos, la información recibida—, 3) capacidad para almacenar la información recibida, 4) capacidad para utilizar la información obtenida y procesada mediante las habilidades anteriores en la resolución de problemas y 5) capacidad para tomar decisiones. En este trabajo designaremos a estas habilidades como *habilidades de la inteligencia*. En la capacidad para recibir información incluiremos la capacidad para comunicarse con otros agentes inteligentes en la medida en que primaremos la obtención de la información en la comunicación sobre otras posibles funciones.

Vaya por delante que en este trabajo no se pretende definir la inteligencia artificial, concepto que forzosamente ha de ser gradual, entendiendo, al menos por tal, la que es "mayor que cero". Ahora bien, sí interesa al jurista precisar cuándo o con qué grado de inteligencia se supera el "umbral de responsabilidad", entendiendo por tal aquel en el que el agente artificial puede ser hecho responsable de sus decisiones.

II. El concepto de *inteligencia artificial* hace referencia tanto a una rama de la ciencia de la computación que trata de diseñar agentes artificiales inteligentes como a la inteligencia de tales agentes. Aunque se ha admitido que un agente inteligente artificial es aquel que es capaz de evaluar las circunstancias y condiciones de su entorno para adoptar decisiones que maximizan sus posibilidades de éxito (Poole, Mackworth, Goebel, 1998).

En este contexto, una decisión racional será aquella no programada que se adopta utilizando la información recibida en función de objetivos o fines, teniendo en cuenta sus condicionamientos previos. Condicionamientos previos serán tanto el *software*, que podría determinar las habilidades de la inteligencia, el ámbito de actuación y los objetivos del agente, entre otros, así como el *hardware*, su cuerpo físico, que forzosamente limita su capacidad de actuación. A estos efectos, y con independencia de este último factor —*hardware*—, la inteligencia no es absoluta, sino graduable.

La calificación como *racional* de una decisión implica que es adecuada o funcional, que está orientada a objetivos o fines predeterminados. A nuestros efectos, un objetivo marcará la producción de un resultado concreto y predeterminado, de configuración fija y estable (por ejemplo: arar la finca "F" de perímetro delimitado y características físicas concretas o resolver un problema matemático concreto). Un fin será aquel que determina una valoración positiva y propone un modelo a alcanzar en un conjunto de actuaciones (por ejemplo: bajar el nivel de la temperatura del agua, evitar choques frontales en vuelos, etc.). Los objetivos están orientados a fines, pero tanto la obtención de fines como la obtención de objetivos pueden exigir el diseño de estrategias.

El fin preseleccionado orienta la actuación del agente inteligente a partir de valores, que convierten el fin en algo *valorable* como positivo o negativo. En este sentido, tiende a presumirse que los fines son buenos y calificarse de irracional la conducta que no tiende a ellos. De ahí la necesidad de determinar, de la forma más objetiva y desvalorada posible, el fin (u objetivo) que en el caso concreto persigue directamente el agente con su conducta para poder afirmar si la conducta es racional o no —de forma independiente a la valoración positiva o negativa que se realice de dicho fin u objetivo—.

Una decisión autónoma será la que se adopta por propio el agente, que es capaz de dirigir su comportamiento o el de otros con su decisión. En una primera acepción, *autonomía* hace referencia a la capacidad de actuación sin intervenciones externas o, en este

[1] Esta investigación se ha desarrollado en el seno del Proyecto de Investigación I+D DER2013-42712-P: "Inteligencia artificial: pautas de conducta y modelos normativos (sobre un código de conducta para entes inteligentes no humanos)". Agradezco las sugerencias, los debates y las aportaciones a Javier A. Díez de la Cuestadron.

caso, humanas. Sin embargo, quizá sea conveniente utilizar (o construir) un concepto normativo de autonomía que haga referencia a la posibilidad de definir las propias relaciones o la propia conducta (en paralelo con lo que en derecho sería la *autonomía de la voluntad*). Esta precisión tiene como objeto excluir del concepto de *conducta autónoma* la del autómata que para realizar su tarea no requiere apoyo externo.

La autonomía de la decisión, así entendida, hace referencia a la opción por una conducta en concreto entre una o varias alternativas de conductas posibles, pero, en principio, no hace referencia a la autonomía en la decisión de los fines o los objetivos, aunque cabe también una decisión autónoma sobre fines u objetivos, por supuesto. Pero de momento, y dado el estado de la ciencia, cuando nos refiramos a decisiones autónomas estaremos haciendo referencia a autonomía en cuanto a la conducta o comportamiento más adecuado a un fin u objetivo, excluyendo la autonomía de la decisión sobre el fin o el objetivo. Con ello estamos limitando el análisis a problemas en los que los fines u objetivos han sido fijados de forma externa al agente artificial inteligente y, como consecuencia, se implica a una persona (o a varias) en la toma de decisiones —dado que los fines u objetivos condicionan la decisión racional y autónoma del agente— y, con ello, se le hace responder también por las consecuencias de los actos del agente artificial. La medida de tal responsabilidad y los títulos en los que se fundamenta deberán ser objeto de estudio y concreción, pero no excluirá y podrá verse afectada, al menos en el ámbito teórico y de momento, la responsabilidad del agente artificial por su propia decisión. Habrá que dilucidar, en este punto, cuándo el agente artificial deja de ser un mero instrumento (en el sentido jurídico-penal del término) para dar paso a nuevas formas de autoría distintas a una, *mutatis mutandis*, autoría mediata clásica.

Por ejemplo, un robot para el control de fronteras que tiene la orden de disparar cada vez que detecte en una franja de tierra a personas en movimiento es autónomo, en el sentido de que puede actuar sin intervención externa, pero no es inteligente. La autonomía relacionada con la inteligencia es aquella que permite decidir al robot qué debe hacer según las circunstancias, con la posibilidad, incluso, de adoptar comportamientos que no han sido predeterminados en su *software*. Esta autonomía, que abriría el paso a la responsabilidad, es la derivada de la libertad de decisión como concepto contrario al de necesidad de decisión.

III. En la actualidad existen ya agentes artificiales inteligentes que adoptan decisiones de comportamiento para la consecución de objetivos, en función de fines predeterminados y en atención a las circunstancias. Estas decisiones son autónomas, lo que significa que no dependen de la autorización o control —individualizado para cada conducta— de una persona y pueden dar lugar a consecuencias que, de deberse a una conducta humana, podrían acarrear responsabilidad civil e incluso penal y que, desde luego, son admitidas como válidas y plenamente eficaces en el tráfico jurídico.

Este tipo de agentes inteligentes que adoptan decisiones y aprenden de su propia "experiencia" son comunes no solo en el ámbito industrial, sino incluso en el cotidiano de los ordenadores personales, donde cada vez más programas "aprenden" de las instrucciones y correcciones que van recibiendo —caso, por ejemplo, de los diccionarios y autocorrección en los procesadores de texto—. Bien es verdad que estos agentes inteligentes, en la actualidad, limitan su capacidad de actuación y aprendizaje a un número reducido de tareas, cuando no a una tarea única, si bien, un objetivo cercano es diseñar protocolos de comunicación que permitan dialogar a distintos agentes inteligentes, bien que sean idénticos o con la misma función, bien con distintas funciones en un proceso industrial o de otro tipo. La inteligencia artificial, en este momento tecnológico, trata de consolidar la comunicación entre agentes inteligentes y la *inteligencia artificial general* —que trata de reproducir o imitar la inteligencia humana y con ello su libertad (entendida como posibilidad de decidir entre diversas alternativas)—, es un reto aún no al alcance de la mano).

IV. La aparición de agentes artificiales inteligentes que pueden decidir autónomamente ejecutar —o hace ejecutar a otros— acciones u omisiones parcialmente predeterminadas (las decisiones) por los fines (Balke/Eimann, 2008) u objetivos marcados, que son capaces de aprender y, a partir de ese aprendizaje, también autónomo, configurar de nuevo parcialmente, nuevas estrategias y adoptar ciertas decisiones directamente derivadas de su aprendizaje o condicionadas

por este, y que, como consecuencia, pueden llegar a realizar conductas socialmente significativas, abre un complejo campo de estudio y plantea numerosas cuestiones —de toda índole pero muchas con trascendencia jurídica— que, sin duda, se irán dilucidando a medida que dichos agentes artificiales inteligentes vayan desarrollando actuaciones (comportamientos) que impliquen a terceros. Algunas novedades normativas, sin embargo, ya se han ido produciendo en los ordenamientos comparados y, específicamente, en el nuestro.

A) Ejemplo de ello es el Real Decreto-Ley 8/2014, de 4 de julio, *de aprobación de medidas urgentes para el crecimiento, la competitividad y la eficiencia* (corrección de errores publicadas en el *Boletín Oficial del Estado —boe—* 167 de 10 de julio), que entre medidas de muy diversa índole regula en España el uso de RPA (por sus siglas en inglés, *Remotely Piloted Aircaft*) o UAV (por sus siglas en inglés, *Unmanned Aerial Vehicle)*, los drones (Diez de la Cuesta, 2008).

> En general, se utiliza tanto el término dron —en español, preferido en Internet— como *drone* —en inglés literalmente "zángano"—. La RAE ha acogido ya el término en español *dron* para designar "aeronaves no tripuladas". En plural, drones, es utilizado por el citado en la Exposición de Motivos del Real Decreto-Ley 8/2014, de 4 de julio, para referirse a vehículos aéreos no tripulados —RPAS (por sus siglas en inglés, *Remotely Piloted Aircaft)* o UAVS (por sus siglas en inglés, *Unmanned Aerial Vehicle)*— con forma de aeronave.

El Real Decreto-Ley define a los drones como "aeronaves civiles pilotadas por control remoto" —excluyendo, por tanto, a aquellos drones que vuelan de forma autónoma, una vez definidos el punto de despegue y de aterrizaje— y autoriza su uso cuando su peso es inferior a los 150 Kg o, aún con peso superior, si están destinados "a la realización de actividades de lucha contra incendios y búsqueda y salvamento, dado que, en general, el resto estarían sujetas a la normativa de la Unión Europea". Con independencia de otras cuestiones, entre ellas el propio concepto de dron o *uav*, esta norma se destina a regular las condiciones de uso de las máquinas (cuya regulación más concreta remite a un reglamento) y atribuye la responsabilidad por daños causados por la operación o la aeronave, así como la eventual responsabilidad frente a la Administración Pública por incumplimiento de la normativa aplicable, al operador "que es, en todo caso, el responsable de la aeronave y de la operación" (art. 50 RDL 8/2014, de 4 de julio). Como consecuencia, se plantean tres posibles modalidades de responsabilidad penal:

- Responsabilidad personal por acciones o daños u otros resultados típicos, cuando la aeronave actúe bajo control de una persona.
- Responsabilidad personal por acciones o daños u otros resultados típicos, cuando la aeronave actúe dirigida por una persona, pero en la decisión en concreto hayan intervenido factores ajenos a la persona que dirige, que desarrolla el *software* u otras.
- Responsabilidad personal por acciones o daños u otros resultados típicos, cuando la aeronave actúe autónomamente, en este caso, con dos subvariantes:
  - como consecuencia de errores, déficits o deficiencias de cualquier tipo en el *software* que establece fines, objetivos, límites y condiciones que incidieron en la toma de decisiones;
  - subsidiariamente, cuando no concurriere el supuesto anterior, por la decisión autónoma y derivada del propio aprendizaje de la aeronave.

B) En un ámbito completamente distinto, pero con problemas no tan alejados, aunque con una trascendencia mayor, en principio, en el ámbito mercantil (pero como consecuencia eventualmente también penal, en este caso por delitos contra el patrimonio, revelación de datos protegidos, delitos contra los derechos económicos de los consumidores, etc.), agentes artificiales gestionan las páginas web de *bancos online* a través de las cuales se puede realizar todo tipo de operaciones, incluso de alto riesgo: transferencias, depósitos, contratación de deuda pública, adquisición de acciones o derivados y un largo etcétera de operaciones bancarias y bursátiles. Al margen de otras cuestiones, el agente artificial que interacciona con los clientes (consumidores) adopta decisiones con trascendencia jurídica (acepta depósitos a determinado interés, realiza transferencias, gestiona pagos o presenta declaraciones a modo de intermediario ante organismos públicos y, por supuesto, detrae comisiones, calcula y paga (o no) intereses, etc.). Aquí

también será preciso analizar, como en el supuesto anterior, las distintas modalidades posibles de responsabilidad personal como consecuencia de decisiones adoptadas autónomamente por el agente artificial.

C) Finalmente, en octubre de 2018, los periódicos internacionales se hicieron eco de algunos intentos de determinar "los comportamientos morales" que debían adoptar los coches autónomos en caso de riesgo para las personas. La cuestión, planteada desde el desconocimiento del derecho, como una cuestión moral, es un problema de estado de necesidad: en caso de que el coche autónomo tenga que decidir entre matar a A o a B, ¿cómo debe actuar? Se procedió a encuestar a un alto número de personas que dijeron lo que, efectivamente, tuvieron a bien. Y quiero pensar que con ello se trató de hacer llegar a la sociedad la certeza de que estos vehículos también pueden matar. Porque lo contrario, es decir, creer que con una encuesta van a resolver un problema tan antiguo como el de la Tabla de Carnéades, sería terrible para la especie humana.

Pues bien, con todo ello, quiero poner de manifiesto cómo la implementación de tecnología inteligente va a implicar, también, al derecho penal de una forma, además, probablemente más profunda de lo que, en un principio, podría parecer.

2. Mandatos, prohibiciones y *código*

I. El derecho penal, como rama del ordenamiento jurídico, tiene entre sus funciones principales, por un lado, fijar normativamente conductas (y en tal sentido, como ciencia, es una ciencia normativa) y, por otro, imputar delitos y exigir responsabilidad. Desde esta perspectiva, el derecho penal tiene como objetivo establecer pautas de conducta correctas o, en términos jurídicos, acordes con las previsiones normativas (desde la perspectiva de protección de bienes jurídicos y con los límites materiales y formales enunciados por la ciencia penal), y diseñar consecuencias jurídicas —sancionadoras o preventivas— a imponer cuando se infrinja la norma que establece la pauta de conducta —o dicho de otro modo: cuando la conducta individual se aparte de la pauta diseñada normativamente—.

Con todas las salvedades que tan esquemática y focalizada exposición implica, conviene recordar y resaltar precisamente estas funciones del derecho y la ciencia penal, que son de trascendental importancia para este discurso. Por un lado, porque si entes inteligentes artificiales interactúan con la sociedad o personas —o cuando lo hagan— pueden, con su comportamiento, lesionar o poner en peligro bienes jurídicos, lo que exige reflexionar sobre el segundo aspecto anteriormente enunciado: la exigencia de responsabilidad penal (en el sentido más amplio del término).

Pero, por otro lado, en cuanto que ciencia normativa, cuyo objeto son normas (jurídicas, pero normas), las aportaciones de la dogmática penal pueden muy bien ser utilizables en el ámbito *tecnológico*. Y me explico: si los entes artificiales inteligentes *actúan* o *realizan comportamientos* con significado social, su *comportamiento y actuación* habrán de adecuarse a normas. Esta adecuación normativa del comportamiento tiene aún un doble aspecto; o mejor: implica normas funcionalmente distintas.

Así, y en el ámbito estrictamente tecnológico, será preciso diseñar pautas de comportamiento que predeterminen qué tiene que hacer el ente inteligente artificial. Estas (pautas), dependiendo de las habilidades y estado de la técnica, funciones encomendadas, etc., pueden ser muy complejas y serán las que se incorporarán al *software* como instrucciones u órdenes de actuación. Entre estas pautas de conducta habrá algunas que no tendrán por objeto exclusivo la determinación de la actuación del agente para la directa y simple consecución de un objetivo, sino que junto a estas —órdenes directas— será necesario incluir otras pautas de conducta que indiquen cómo actuar ante eventuales cambios de circunstancias. La complejidad de dichas órdenes, instrucciones o comandos que contengan pautas de conducta se incrementará exponencialmente en función de la capacidad de actuación del ente, y, cuando la capacidad de actuación implique muy diversos factores, las órdenes deberán incluir excepciones.

En este sentido, el *software* contiene un conjunto de instrucciones —código de instrucciones— que posibilita y (pre)determina el comportamiento del agente (si bien, cuando se trate de agentes inteligentes tal predeterminación no puede ser absoluta). En este sentido, el *software* aparece como un sistema normativo de pautas de conducta que pueden generar comportamientos con trascendencia social.

Volviendo a uno de los ejemplos anteriores, recurramos de nuevo a un dron o UAV que tiene capacidad

para volar desde el punto A hasta el punto B de forma autónoma y sin control remoto. Esta capacidad de vuelo implica que el dron, una vez recibida la orden de vuelo e iniciado este, tiene que ejercitar su inteligencia, recurriendo a sus habilidades. Así, tomará datos de las circunstancias en que se desarrolla el vuelo —estado meteorológico, estado del territorio que se sobrevuela, etc.—, las procesará y utilizará para definir su propio vuelo (altura, velocidad, incluso trayectoria, entre otras), lo que implica que ha de adoptar decisiones que no están absolutamente predeterminadas o que no lo están con carácter necesario. Pues bien, el *software* que le permita ejecutar la orden "¡vuela de A hasta B!" incluye un complejo conjunto de normas: por ejemplo, "mide la velocidad del viento", que incluyen a su vez "abre las puertas del compartimiento donde se encuentra el anemómetro", "conecta electricidad al anemómetro", "cuenta las vueltas que da el anemómetro", y otras como "a partir de X mantente a la distancia Y del suelo", "con velocidad de viento X, viaja a velocidad Y", etc. Estas normas, que predeterminan la actuación del dron, no solo serán formuladas en sentido positivo (al estilo de los mandatos en la Teoría de las Normas), sino que en ocasiones habrán de incluir prohibiciones o excepciones (permisos). Las prohibiciones son normas que impiden la actuación y vincularán el comportamiento del UAV, el cual no podrá actuar en contra de ellas. En este sentido, los límites de velocidad, por ejemplo, son normas prohibitivas ("¡no superes los 200 km/h!", "¡no vueles por debajo de 50 km/h!"). Pero, junto a estas normas de actuación absolutas, caben también normas de actuación regidas por lógica difusa, que permitirá graduar parámetros de actuación en función de circunstancias (por ejemplo, las que determinen que el cálculo de la altura a la que volará el dron, una vez alcanzada la velocidad de crucero, dependerá de los factores X, Y, Z, en proporciones dadas).

Por otro lado, también y todavía en el ámbito estrictamente técnico —en el sentido relacionado con el vuelo del UAV—, se puede necesitar introducir excepciones a las pautas de conducta establecidas como primarias. Las excepciones permiten excluir, en función de las circunstancias, un mandato o una prohibición o introducir una variable de comportamiento según el caso. Siguiendo con nuestro ejemplo, la orden "¡vuela de A hasta B!" es la puesta en marcha de un conjunto de normas de actuación que permitirán al UAV volar desde A hasta B. Pero, si el UAV ha de tener en cuenta las circunstancias, y puesto que estas son imprevisibles, habrá de incluir normas de actuación según las circunstancias, lo que introduce márgenes de —llamémosle— discrecionalidades y riesgos que habrán de haber sido previstas. Así, por ejemplo, el código normativo incluido en el *software* deberá prever qué debe hacer el UAV cuando en pleno vuelo detecte que una paloma se cruzará en su camino. No parece difícil encontrar la solución. Es muy posible que el lector ya haya previsto que, en ese caso, el UAV debe alterar su trayectoria para evitar la colisión (por razones muy diversas que pueden ir desde las ecológicas a las meramente técnicas —evitar que las plumas entren en el motor—). De modo que, si el código normativo del UAV contenía una norma similar a "utiliza la línea directa para unir el punto de partida y el punto de destino", ahora habrá de incluir otra norma que permita excepcionar a la anterior, por ejemplo: "si en tu trayectoria hay riesgo de colisión con otro objeto que vuela, altera la trayectoria subiendo 2 metros de altura".

II. Tenemos entonces, y aún no hemos prácticamente empezado, un complejo conjunto de pautas de conducta que configura el "código" de instrucciones que posibilitan la realización de las tareas específicas que tiene encomendadas o para las que tiene competencia el agente. Traducido al lenguaje propio de los juristas, las instrucciones son normas que contienen pautas de conducta, y el conjunto de instrucciones o código es un sistema normativo.

La dogmática penal distingue, a estos efectos, entre normas primarias y normas secundarias. Normas primarias son las que contienen pautas de conducta (mandatos o prohibiciones), y las secundarias van dirigidas al operador jurídico que ha de imponer una consecuencia jurídica cuando se produzca el incumplimiento. A los efectos que estamos analizando ahora, ciertamente las normas secundarias carecen de interés; no así las normas primarias, deducidas implícitamente de aquellas (secundarias) que dirigen y ordenan el comportamiento humano. Junto a estas normas primarias que ordenan o prohíben una conducta, se sitúan los permisos —si bien existe un amplio debate acerca de la naturaleza de los mismos— que facultan para infringir un mandato o prohibición

o para decidir sobre su vigencia en caso de conflicto entre ellos.

Este esquema normativo es útil también para explicar el contenido de los códigos que conforman el *software*. Ciertamente y en abstracto, habrá instrucciones que no tengan, al menos de forma directa y aislada, trascendencia a efectos sociales. Otras, sin embargo, en la medida en que impliquen decisiones que pueden afectar a valores reconocidos y protegidos socialmente, merecen un mayor análisis desde una perspectiva jurídica. En cualquier caso, y antes de continuar, cuando se produzca una lesión o daño con trascendencia jurídica (siquiera a efectos civiles) la determinación de la responsabilidad puede venir desde la redacción de instrucciones simples aparentemente alejadas de la toma de decisión, pero que la condicionan (por ejemplo, una inadecuada instrucción sobre el conteo de vueltas de anemómetro, puede dar lugar en un error en la apreciación de la velocidad del aire y de ahí derivar una decisión de vuelo inadecuada a las circunstancias que puede dar lugar a una colisión).

Actualmente, las normas o instrucciones que se contienen en el código o *software* son normas expresas, en el sentido de que si no existe una norma expresa que habilite una actuación, en principio la máquina no realizará la conducta. Puede haber, sin embargo, varios motivos que hagan que esto no sea así: que aunque no haya una norma expresa la conducta se realice. Los motivos pueden ser una inestabilidad del sistema (derivada de un código defectuoso), un ataque externo que consiga la ejecución aleatoria o no de código no deseado o un simple *bug* (código defectuoso propio de la tarea). Todo ello remite de forma directa o indirecta al problema del error humano. A medida que se avance hacia una inteligencia artificial general, la complejidad del código (*software*) irá aumentado, lo que lo acercará a los sistemas normativos en los que no solo caben normas expresas, sino también tácitas, derivadas de aquellas o de un conjunto de ellas (de normas expresas o tácitas).

Las excepciones o permisos —limitaciones o alteraciones de las normas primarias— pueden deberse a cuestiones técnicas (para evitar calentamiento excesivo del motor, etc.) o a razones que pueden tener que ver con el objetivo (aterrizar en B) o con la interacción que la consecución del objetivo implica con terceros (evitar una zona poblada). La interacción con terceros (sociedad, personas) implica la jerarquización normativa no ya en función de razones técnicas, sino de valores que actúan como fines. Pero, con independencia del interés social en la predeterminación de los valores —cuestión en la que ahora no vamos a entrar—, la mera coordinación normativa entre normas primarias que contienen órdenes o prohibiciones es muy compleja. En primer lugar, porque las prohibiciones normalmente llevarán asociada una pauta de conducta alternativa —lo que de hecho puede tener como consecuencia su inexistencia expresa—. Y, en segundo lugar, porque la determinación exacta del momento en que una norma primaria ha de ser infringida y con qué intensidad es algo que requiere una valoración, lo que, pese a los siglos de práctica, sigue originando importantes problemas en los tribunales.

3. Principios jurídico para la inteligencia artificial

Los cambios tecnológicos no siempre implican cambios éticos; la utilización de nuevas tecnologías basadas en algoritmos inteligentes con capacidad de aprendizaje, tampoco. Pero, cuando un agente inteligente —con capacidad de aprendizaje— realiza actividades o tareas cognitivas con trascendencia social debe estar sometido a los criterios éticos y jurídicos —criterios orientados a valores— propios de la sociedad humana. Entre estos valores, Bostrom y Yudkowsky (2011) consideran imprescindibles los siguientes: transparencia en la actuación (*transparency*) —de forma que se conozca por qué criterios se rige en la toma de decisiones— y posibilidad efectiva de comprobación de la puesta en práctica de tales criterios, así como de someter a control los resultados (*auditability*); previsibilidad en las decisiones que afecten a personas (*predictability*) —en el sentido en que los penalistas entendemos el principio de seguridad jurídica—, invulnerabilidad (*incorruptibility*) —fortaleza frente a posibles manipulaciones del *software*, en el sentido de seguridad informática—, y posibilidad de identificar a "la persona responsable" (*responsibility*), y, finalmente, tendencia a no perjudicar a terceros (*a tendency to not make innocent victims* —Bostrom/Yudkowsky, 2011—).

En definitiva, todos son criterios o principios que también son exigibles *mutatis mutandis* a las personas que desarrollan su actividad en el ámbito social, y cuyas conductas pueden afectar a otras personas. De en-

tre todos ellos, y a los efectos que ahora nos interesan, conviene destacar cómo en este pequeño catálogo que hemos detallado —y que no pretende agotar todos los posibles principios de actuación ética de un agente inteligente— se traslada la responsabilidad por la decisión que adopta el agente inteligente a personas que responderán de los perjuicios por él causados. Al margen de otras cuestiones, esta responsabilidad de las personas abre dos importantes cuestiones: la primera sería la identificación de la persona responsable (¿el programador?, ¿el gestor?). En relación con los UAV o drones, nuestro ordenamiento jurídico ha optado por trasladar, a efectos administrativos, la responsabilidad al operador, en la medida en que los considera aeronaves, lo cual es lógico si se tiene en cuenta que solo regula los drones dirigidos por control remoto. Ahora bien, a efectos penales esta atribución de responsabilidad es insuficiente —incluso aunque solo se trata de aeronaves que funcionan por control remoto, pero, sobre todo, cuando no sea así—.

Es insuficiente porque el gestor (operador, en el caso del dron) recibe a su vez un producto —en términos mercantiles— construido por otra u otras personas. Como consecuencia, a efectos penales, la responsabilidad por el producto se compone de círculos de responsabilidad individual, compartida parcialmente, o no, en función de los propios ámbitos de competencia de cada uno de los intervinientes en la fabricación y uso del producto. Pero, junto a ello, será necesario reflexionar sobre la responsabilidad penal por resultados lesivos típicos que sean consecuencia de errores humanos. El problema de la relevancia penal del error humano es un problema propio de la *sociedad del riesgo* y ya se ha planteado en relación con otras materias, por ejemplo, con el uso pacífico de energía nuclear. Pues bien, en estos casos, confluyen distintas "responsabilidades" individuales (De la Cuesta Aguado, 2016):

- responsabilidad de quien comete el error en su ámbito de competencia, que puede implicar errores de distintas personas;
- responsabilidad por *culpa in vigilando*, y
- responsabilidad derivada de la creación o uso de una fuente de peligro.

La especialidad de este caso deriva de que un error humano puede ser potenciado como consecuencia de la capacidad de decisión del agente. Esto es especialmente evidente en relación con el *software* que contiene errores en el código —defectuoso o incompleto—. Todo ello exige profundizar en la dogmática del riesgo, así como en modelos (garantistas) más completos de responsabilidad por el producto.

La segunda gran cuestión en torno a la exigencia de la responsabilidad penal por los daños, lesiones o, en general, resultados —de momento— delictivos nos sitúa en el ámbito de la responsabilidad por el producto. Efectivamente, el agente artificial podría ser considerado, en el sentido antedicho, como *dependiente* de la responsabilidad de una persona, pero las exigencias que en los últimos 20 años se han venido formulando respecto de productores y agentes que, en general, colocan productos en el mercado, probablemente debieran ser revisadas para definir respuestas satisfactorias para los problemas previsiblemente planteados en estos nuevos ámbitos. Entre otras razones —aunque no solo— por la trascendencia posible de errores humanos de consecuencias previsibles o imprevisibles.

Además de lo anterior —y aquí se abre un nuevo frente en el derecho penal— respecto de algoritmos con capacidad de aprendizaje y decisión, el control de calidad ha de ser riguroso y exhaustivo, lo que exigirá reflexionar y prever los parámetros de seguridad en dicho control de calidad y, probablemente, con base en ellos, fundamentar nuevas modalidades de exigencia de responsabilidad penal —en una construcción, si no similar, sí con muchos paralelismos con la exigencia de responsabilidad penal a las personas jurídicas—.

Las cuestiones sobre las que estamos reflexionando ponen de manifiesto que, si bien con distinta terminología, los problemas que se han de plantear en relación con la interacción con agentes artificiales inteligentes han de ser muy parecidos a los que se plantean en la interacción social. Ello es así sobre todo en aquellos supuestos en que "suplantaren" a una persona; es decir, cuando el agente artificial realizare actuaciones que comprometieren a personas físicas o jurídicas, en la medida en que el agente participa como sujeto (por sí o por otro) en la sociedad. Esta situación es fácilmente comprobable en relación con los "agentes informáticos" (Balke/Torsten, 2008) que contratan servicios bancarios con particulares (*web* de bancos *online,* por ejemplo), tal como hemos visto.

Cierto es que una cosa es admitir que un agente artificial pueda participar como agente (sujeto) en el mercado y otra muy distinta que la responsabilidad de su actuación, además de trascenderle, genere responsabilidad para él mismo. Debate similar, de nuevo, ha sido el mantenido durante años en relación con la responsabilidad de las personas jurídicas, si bien las razones que llevaron a generalizar en el derecho continental su responsabilidad penal no tienen por qué ser susceptibles de extensión al ámbito de la inteligencia artificial.

4. Inteligencia artificial y lesión de bienes jurídicos

Havelly (2010) utiliza como ejemplo un caso supuestamente sucedido en Japón en 1981, en el que un operario de una cadena de montaje trabajaba junto a un agente artificial inteligente que, en un momento determinado, decidió que el trabajador era una amenaza para su trabajo y lo mató. Utilizando este ejemplo nos planteamos varias cuestiones. De ser cierto, parece evidente que el código de conducta del agente artificial no contenía norma alguna que hiciera referencia a la necesidad de, en caso de conflicto, primar un comportamiento frente a otro.

Si el agente inteligente ha de enfrentarse a una realidad cambiante —y así ha de ser, pues en caso contrario no sería necesario que lo fuera (inteligente)— podrá encontrarse ante supuestos en los que tenga la posibilidad de decidir entre distintas pautas de conducta o deba decidir entre un mal y otro. La despersonalización que implica la reflexión sobre agentes *no sujetos de derecho* (o sea, personas) puede facilitar un "movimiento de puertas giratorias" entre la dogmática "para la inteligencia artificial" y la dogmática "para las personas".

En cualquier caso, la primera tarea será la de diseñar las pautas de conducta —o explicar por qué no se debe imponer ninguna— que deben regir la toma de decisiones de estos agentes inteligentes cuando su actuación o los hechos que ejecutan tengan relevancia social.

La simplicidad de la cuestión tan solo es aparente (tanto desde la perspectiva humana como desde la perspectiva artificial) y, de hecho, en algo tan aparentemente sencillo como decidir sobre la legítima defensa (entre personas) se han vertido ríos de tinta y es más que dudoso que se pueda predeterminar la decisión judicial en un caso en concreto. Más aún, la aparentemente simple regla que permite la legítima defensa del artículo 20.4 del Código Penal español o el artículo 15.4 del Código Penal Federal de México va acompañada de un conjunto de "reglas" de aplicación de construcción doctrinal o jurisprudencial, por lo que en sí misma es de una enorme complejidad y, de hecho, prácticamente desconocida por los ciudadanos. Pero si tuviéramos que enunciar sistemáticamente los permisos —y sus graduaciones— que regulan la defensa permitida, el código de instrucciones sería sumamente complejo.

Veámoslo con ejemplos: uno de los ámbitos donde mayor es el desarrollo y el interés por la inteligencia artificial es el de la gestión de la seguridad. Un agente inteligente artificial que deba tomar decisiones que afecten a bienes jurídicos de personas —libertad, intimidad, salud o vida— necesita unas pautas de conducta claramente definidas, pero que resuelvan, también sin ningún género de dudas —o con las menores dudas posibles—, las situaciones de conflicto. Pero el diseño de estas pautas de conducta no es, en absoluto, simple; sobre todo si se observan las ambigüedades de que adolece, por ejemplo, la regulación que determina las competencias y habilitaciones a las Fuerzas y Cuerpos de Seguridad de las sucesivas leyes de seguridad ciudadana en relación con el Código Penal. Las dificultades con que se encuentran las autoridades y agentes de las Fuerzas y Cuerpos de Seguridad del Estado para determinar en cada caso concreto dónde están los límites entre sus competencias y los ámbitos de libertad del ciudadano ponen de manifiesto las dificultades para diseñar modelos que no requieran valoración —y en ese sentido "seguros"— de comportamiento.

Aunque pudiera parecer excesivamente literario, lo cierto es que algunos Estados ya están invirtiendo en la formación de "robots" inteligentes para la gestión de la seguridad o para evitar bajas personales en caso de conflicto armado (robot-soldado), sin perjuicio de los drones o aeronaves no tripuladas con fines diversos, entre otros, los de ejecutar decisiones que implican la muerte de personas (ejecuciones, en otros términos). Y, de hecho, ya existen agentes con inteligencia artificial que pueden tomar algunas decisiones y que se utilizan en casos de conflicto armado, sin que haya quedado muy claro en la información —escasí-

sima— ofrecida a la opinión pública en qué medida la decisión de algunos medios de ataque o espionaje depende totalmente de un controlador humano.

Para la resolución de estos conflictos la lógica jurídica, deóntica, difusa o cualquier otro tipo de lógica no han encontrado solución única porque de alguna forma requiere un juicio de valor que implica dos posibilidades de conducta enfrentadas y vinculadas a dos normas de conducta (mandato y excepción) y un valor (al menos).

5. Sobre la racionalidad en la construcción normativa: normas penales e instrucciones del código

He tratado, hasta aquí, de ir poniendo de manifiesto las similitudes entre los problemas que pueden surgir a la hora de definir instrucciones para un agente inteligente y la tarea propia del derecho y la ciencia penal. El derecho penal es un conjunto sistemático de normas que establecen pautas de conducta (teleológicamente orientadas, lo que implica su asociación a un sistema de valores e imperativamente impuestas). Bien es cierto que el término *norma* en derecho penal es polivalente y debe entenderse por el contexto. En esta investigación las "normas" que nos interesan son las que definen pautas de conducta: las normas primarias y las proposiciones permisivas.

La teoría de las normas —marco teórico que trata de explicar las normas primarias, dirigidas al ciudadano— es uno de los campos de la dogmática penal más difícil y que ha merecido una menor reflexión en los últimos tiempos. Sin embargo, en la configuración de la teoría general del delito ha ocupado un lugar primordial, y las aportaciones y reflexiones que han servido en el ámbito penal probablemente puedan ser aplicables no con fines estrictamente académicos, sino con la finalidad de obtener modelos de normas (en el sentido de estructuras formales para transmitir mandatos o prohibiciones) a códigos de instrucciones encargados de definir pautas de conducta para agentes con inteligencia artificial.

La dogmática penal aparece, así, como la ciencia con los instrumentos teóricos más precisos y con la mayor experiencia para abordar el problema planteado. Ahora bien, las investigaciones y publicaciones sobre delitos o delincuencia informática se han centrado mayoritariamente en la revisión de modalidades de delitos clásicos que han utilizado medios informáticos o Internet para su expansión, difusión o como medio de ejecución del delito y en el análisis de nuevas modalidades típicas relacionadas con medios informáticos. Desde una perspectiva criminológica han acaparado la atención el estudio del fenómeno *hacker* y del *ciberterrorismo* —incluyendo también aquí el terrorismo tradicional que utiliza, para su difusión o en su beneficio, Internet—. Las nuevas relaciones sociales, las nuevas formas de abuso surgidas como consecuencia de las nuevas formas de comunicación —redes sociales, Internet, etc.—, sobre todo cuando exigen conocimientos informáticos o de programación, han merecido menor atención debido a la distancia entre la ciencia informática y la ciencia jurídica. Pero la escasa comunicación entre programadores y juristas —entre tecnología informática y derecho penal— se ha debido también a que la ciencia penal es "pública", sobre todo en España. Es pública porque su financiación lo es, pero también porque lo son sus avances o aportaciones, mientras que la tecnología informática, especialmente referida al *software*, o es "privada" o, al menos, no tan pública —con excepciones significativas—. Ello es debido a que parte importante de la investigación con fines de aplicación práctica se desarrolla en el seno de empresas o de instituciones financiadas con capital privado, porque sus fines son lucrativos —luego, privados—, y porque sus avances se protegen con celo de cualquier atisbo de publicidad —recuérdese el debate sobre la ingeniería inversa—.

Finalmente, es una realidad que al derecho penal le está costando seguir el ritmo evolutivo de la tecnología informática que, cada vez a mayor velocidad, está materializando lo que hace pocos años parecía solo fruto de la imaginación. Esta última afirmación es especialmente cierta en el ámbito de la inteligencia artificial. Hasta ahora en la opinión pública inteligencia artificial es sinónimo de androides o torpes robots bípedos —ASIMO, por ejemplo, pero no solo— y, más recientemente, de drones. Sin embargo, entes con inteligencia artificial, con mayor o menor autonomía, están invadiendo nuestra vida cotidiana no solo como parte de procesos industriales, sino también en las actividades normales de millones de personas que se relacionan con "buscadores inteligentes", traductores inteligentes, "barredoras" inteligentes, tractores inteligentes o jugadores de ajedrez inteligentes. Todo

ello, sin embargo, no es más que la muestra del punto en que se encuentra la investigación sobre inteligencia artificial, que se dirige rápidamente hacia la inteligencia artificial general, a partir de un avanzado estado de desarrollo de agentes inteligentes o sistemas multiagentes. La expresión *inteligencia artificial general* se utiliza para referirse a la inteligencia real de las máquinas, que en la actualidad no han logrado el nivel de generalidad de la mente humana. En el momento en que se alcance, la velocidad de la evolución y perfeccionamiento tecnológico será altísima, puesto que, al fin y al cabo, esta inteligencia artificial general es *software*, por lo que "su capacidad de reproducción" será muy alta.

En esta situación, sin embargo, las investigaciones sobre las implicaciones sociales de la creación de entes con inteligencia artificial general son escasas en nuestro entorno. El panorama, someramente esbozado en estas líneas, abre un nuevo campo de estudio al penalista y plantea cuestiones que no solo tienen como finalidad reflexionar sobre eventuales soluciones a posibles problemas, sino que, al enfrentar la construcción dogmática a un proceso de explicitación y racionalización exento de prejuicios propios de la cultura y el lenguaje humano, puede permitir ofrecer soluciones generales a problemas que se han mantenido tradicionalmente en la indefinición o la imprecisión. Este esfuerzo, además de beneficios dogmáticos, puede facilitar la aceptación de las normas penales y del propio sistema penal democrático en sociedades multiculturales en las que confluyen modelos ideológicos diversos.

6. Bibliografía

Aibar Puentes (1993). *La imagen del conocimiento en el ámbito de los sistemas expertos y la inteligencia artificial*. Tesis doctoral, Universidad de Barcelona.

Anderson, M./Anderson, S.L. (Eds.) (2011). *Machine Ethics, New York,* Cambridge University Press.

Balke/Eymann (2008). "The Conclusion of Contracts by Software Agents in the Eyes of the Law". En Padgham, Parkes, Müller y Parsons (Eds.), *Proc. of 7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2008)*. Estoril, pp. 771-778. Disponible en línea en http://www.ifaamas.org/Proceedings/aamas08/proceedings/pdf/paper/AAMAS08_0596.pdf [citado: 12.07.2019]

Bermejo Sánchez (2003). *Desarrollo de robots basados en el comportamiento*. Ediciones UPC, Barcelona.

Bostrom/Yudkowsky (2011), "The ethics of artificial intelligence". En W. Ramsey y K. Frankish, *Cambridge Handbook of Artificial Intelligence*. Cambridge University Press, pp. 7-21.

De la Cuesta Aguado (2016). "La ambigüedad no es programable: racionalización normativa y control interno en Inteligencia Artificial". *Revista de Derecho y Proceso Penal*, 44, pp. 165-144.

Diez de la Cuesta (2008). *Análisis e integración de componentes de aviónica para UAV'S,* ejemplar dactilografiado. Sevilla, 2008. Disponible en https://docplayer.es/11679181-Analisis-e-integracion-de-componentes-de-avionica-para-uav-s-proyecto-fin-de-carrera-tutor-prof-dr-anibal-ollero.html [citado: 12.07.2019]

Goertzel/Pennachin (Eds.) (2006). *Artificial General Intelligence*. New York, Springer Verlag.

Henley (1993). "Abstract Principles, Mid level Principles, and the Rule of Law". *Law and Philosophy*, 12, pp. 121-32.

Hernández López (2004). *Heurísticas para el control deliberativo en una arquitectura de agentes inteligentes de tiempo real*. Tesis doctoral, Valencia, Universidad Politécnica de Valencia.

Hidalgo (1996). *Inteligencia artificial y sistemas expertos*. Córdoba, Servicio de Publicaciones de la Universidad de Córdoba.

González Bedia/García Carrasco (2006). "Arquitecturas emocionales en Inteligencia Artificial: una propuesta unificadora". *Teoría de la Educación: Educación y Cultura en la Sociedad de la Información*, Vol. 7, nº2, pp. 156-168.

Hirschfeld Gelman (Eds.) (1994). *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge, Cambridge University Press.

Martínez Velasco (1992). "El problema de la conciencia". *Contextos*, nº 19-20, pp. 233-262.

Moravec (1999). *Robot: Mere Machine to Transcendent Mind*. New York, Oxford University Press.

Noriega (2007). "Regulating Virtual Interactions". En Casanova, Noriega, Bourcier, *Trends in Legal Knowledge. The Semantic Web and the Regulation*

of Electronic Social Systems, Barcelona, European Press Academic Publishing, pp. 55-77.

Poole/Mackworth (2010). *Artificial Intelligence*. New York, Cambrigde University Press.

Ordóñez Pinilla (2006). "Monismo anómalo, intencionalidad, falacias mentales e inteligencia artificial". *Bajo palabra: Revista de filosofía*, II Época, nº 1, pp. 38-54.

Weiss (1999). *Multiagent Systems, A modern approach to distributed Artificial intelligence*. MIT, 1999.

Yudkowsky (2003). *Creating Friendly AI*. Disponible en http://www.singinst.org/CFAI/index.html [citado: 12.07.2019].

Zaccagnini Sancho/Adarraga Morales (1998). "Sistemas expertos y psicología cognitiva". *Estudios de psicología*, nº 36, pp. 97-130.