Prof. Dra. Maria Fernanda da Palma Pereira
Catedrática de Derecho Penal. Univ. de Lisboa, Portugal. Socia de la FICP.

~ Conflicto o contribución de la Inteligencia Artificial ante la teoría general de la infracción criminal~

I. EL DESAFÍO CENTRAL QUE PLANTEA LA INTELIGENCIA ARTIFICIAL A LA TRADICIÓN DE LA TEORÍA GENERAL DE LA RESPONSABILIDAD. DIFERENCIACIÓN DE SITUACIONES.

La inteligencia artificial, al modificar el modo de ser de la actividad humana, impone un primer gran problema a los sistemas penales contemporáneos, bien conocido en la teoría de la responsabilidad general: hasta qué punto debe mantenerse el paradigma del control personal y directo de las consecuencias de la actividad o, por el contrario, debe ser sustituido por la mera producción de riesgos, incluso en un contexto de una cierta imprevisibilidad.

Las posibilidades de producción de riesgos por parte de entes autónomos que, en determinados casos, son capaces de actualizar decisiones de comportamiento imprevistas o incluso imprevisibles ante las situaciones a las que se enfrentan –hablo por ejemplo de los vehículos autónomos o de algunos robots dotados de esta capacidad–, permitiría dos actitudes opuestas: la sustracción de las consecuencias del ámbito de la responsabilidad penal pura y simplemente con el argumento de que ni siquiera habría acción, ya que la relación entre la actividad lesiva de la máquina y cualquier momento subjetivo-final o psíquico no existiría; o, por el contrario, la aceptación de cierta responsabilidad omisiva muy amplia por la no evitación de un resultado, a pesar de que el agente que detenta la posición de garante carece de un control efectivo sobre las consecuencias o, según otro criterio, la capacidad de acción.

Si esta contraposición es bien conocida en la teoría del delito y corresponde a una opción sobre la articulación entre libertad y responsabilidad de orden constitucional, las nuevas situaciones que genera la inteligencia artificial en el caso de los entes autónomos plantean el problema de saber si el punto de partida o, dicho de otro modo, el objeto de valoración fundamental de la responsabilidad penal es la articulación directa entre la mente y la conducta física o si esta articulación no corresponderá a un mito causalista y de alguna manera a la idea naturalista que las propias neurociencias han enfatizado de

una fusión monista entre cuerpo y mente. He aquí, pues, el desafío de la inteligencia artificial al desvincular el acto físico y la decisión de actuar de los actos mentales, si consideramos la intervención humana previa. Este desafío consiste precisamente en la cuestión de saber si la base del sistema no deberá entenderse en sentido funcionalista como una función de control superior y objeto de reflexión en lugar de basarse en el modelo humano de acción en la perspectiva causal-finalista. Una visión de tipo funcionalista vendría impuesta no por elección metodológica, sino por la propia naturaleza de la nueva realidad. El dualismo mente-cuerpo sería tan sólo el resultado del desarrollo científico y tecnológico y no algo existente *ab initio* desde una perspectiva cartesiana. Lo que no resulta del modo de actuar humano históricamente conocido, el dualismo mente-cuerpo, a pesar de las interpretaciones metafísicas, se convertiría en un esquema marco de la articulación entre el punto de partida humano y la acción de entes autónomos. Pero el dualismo conduce necesariamente a la consideración de la no causalidad y de un modelo de mera coincidencia o concomitancia entre un cierto estado mental y un determinado resultado de los hechos consecuentes.

Así, en un mundo en el que la mente humana ya no puede controlar al ente autónomo, a pesar de que lejanamente haya iniciado el proceso de acción, los elementos mentalistas del agente humano únicamente justificarían una censura como coincidencia o, a lo sumo, en virtud de la previsibilidad del curso real de los acontecimientos.

Sin embargo, esta solución dualista no deja de ser una desarticulación entre libertad y responsabilidad, ya que el agente humano podría ser hecho responsable a pesar de carecer de alternativas de acción que todos podamos experimentar. De hecho, en un universo dualista, no habrá lugar para ninguna relación entre libertad y responsabilidad, sino tan sólo una posible responsabilidad por lo que se pensó de forma inconsecuente.

De mantener un esquema de libertad-responsabilidad, será necesario entender bien hasta qué punto los agentes humanos tienen control sobre las consecuencias y podrán evitarlas y, en el caso de esta imposibilidad, si aún tendrá sentido cualquier esquema de responsabilidad; es decir, si los agentes artificiales no terminarán por suscitar, cuando sean absolutamente autónomos, una especie de otra naturaleza, incontrolable para nosotros.

Intentar mantener hasta el límite una articulación entre libertad y responsabilidad parece exigible en función de la aceptabilidad de las normas por parte de los destinatarios

y, en este sentido, como ya han defendido varios autores, conviene organizar una diferenciación de situaciones típicas:

- En primer lugar, las situaciones en las que se puede hablar de conductividad en términos cibernéticos en todo el proceso de actuación del ente autónomo;
- En segundo lugar, las situaciones de mero control inicial del agente humano, pero con previsibilidad de los momentos siguientes;
- En tercer lugar, los casos de control a través de la programación y del algoritmo, pero sin control efectivo por un agente humano directo;
- En cuarto lugar, los casos de autonomía de un agente artificial como *machine learning* sin posibilidad ulterior de interferencia y sin previsibilidad con respecto a la decisión del agente autónomo artificial.

Como puede verse, en estos diferentes casos, a veces hay una rendija del control y la evitabilidad que caracterizan la función normal de la acción libre. que justifica la responsabilidad, mientras que en otros casos ha dejado de existir.

La pregunta que se plantea es saber si, aun así, en los últimos casos, no estaría justificada una responsabilidad por la propia creación de una situación impredecible para el agente, una responsabilidad por el puro riesgo asociado a lo desconocido y a la evolución algorítmica inaccesible para los agentes humanos, como sucede en las *actiones liberae in causa*.

II. PARALELISMOS ENTRE LA AUTORÍA MEDIATA Y LOS CASOS DE UTILIZACIÓN DE ENTES ARTIFICIALES AUTÓNOMOS

Este problema tiene un paralelismo con dos tipos de situaciones conocidas de la teoría del delito: la responsabilidad por autoría mediata, tanto en el caso de agentes inimputables como en el de “autores de escritorio”. También aquí puede plantearse el problema de una decisión autónoma ulterior como un mecanismo puesto en marcha y no controlable hasta el límite, en el que la realización y la dimensión del hecho delictivo deja de ser evitable. La reflexión sobre los límites de la responsabilidad de los agentes humanos ante los entes autónomos nos lleva a pensar en la enseñanza que traen estos casos y viceversa.

La cuestión central siempre es si una responsabilidad basada en la pura creación de un ente con decisiones eventualmente contrarias a los límites éticos aceptables de la

acción humana tolerable está plenamente justificada o si, razonando en términos estadísticos, las ventajas de estas tecnologías, en el ejemplo de los coches autónomos, no nos llevarían a considerar permisible el riesgo. La cuestión de la responsabilidad es, por tanto, una opción sobre el sentido de la relación entre libertad, responsabilidad y utilidad colectiva que se plantea como problema de política penal.

Se suscitan tres problemas entonces:

- El primero se relaciona con el concepto mismo de acción típica, relevante a los efectos de la determinación del inicio de la tentativa en la autoría mediata;
- El segundo tiene que ver con el contenido del dolo, su criterio y su alcance;
- El tercero se refiere a la posible relevancia de la autoría mediata en el comportamiento negligente.

En cuanto a la acción típica, se la podría admitir sin conexión física con la acción causal en situaciones específicas de puesta en marcha de un proceso que, posteriormente, resulta imparable, ya que allí también el ente artificial aparecería como una extensión del humano. Incluso tratándose de decisiones ulteriores autónomas, estaríamos ante una situación próxima de las ya conocidas y de los criterios utilizados por la doctrina.

III. LOS PROBLEMAS DEL DOLO

El contenido del dolo, empero, ha de ser repensado, pues es dudoso que la exigencia de una predicción concreta en relación al hecho se constate con objetividad y sin recurrir a la estandarización de un hombre medio en situaciones de decisión ulterior autónoma. En estos casos, si, como siempre he sostenido, no queremos ficcionalizar la predicción y respetamos las concepciones comunes de la representación del hecho por parte de los destinatarios de las normas, el problema a resolver es si una especie de mera conciencia de una posibilidad mínima, inferida por el conocimiento técnico del agente humano en una fase de programación o de utilización, será aceptable dado el alto riesgo posterior, en caso de que éste fuere previsible. ¿La idea de una especie de ceguera o indiferencia será suficiente?

Mi respuesta sería afirmativa en función de la conjugación de un alto riesgo previsto, de una aceptación del mismo en función del interés y con exclusión de una confianza basada, por ejemplo, en una experiencia previa o en una especie de creencia

tecnológica. En todo caso, el dolo, sin perder la continuidad con los criterios más exigentes de los casos tradicionales, requiere una revisión práctica de los criterios.

IV. LAS CUESTIONES DE CULPABILIDAD.

Por último, también en estos casos que empiezan a plantearse, la culpabilidad personal suscita la oposición entre una responsabilidad basada en la capacidad de motivación por la norma (una mezcla de ética de la actitud y de ética de la responsabilidad) y una responsabilidad más orientada a la denominada ética de la responsabilidad en la línea de la distinción establecida por Max WEBER.

En estas nuevas situaciones, la repercusión del principio de precaución no puede ser desestimada desde la perspectiva de lo válido para la bioética, en los términos formulados por Hans JONAS. Una mayor responsabilidad ante las incertidumbres y los riesgos condicionará el margen de disculpa, pero también es cierto, como sostuvo el inolvidable Arthur KAUFMANN, que ante las necesidades humanas de utilización de nuevas tecnologías, hay que buscar una fórmula de tolerancia, en la que se contrapesen el mínimo daño posible al mayor número de personas (a la manera del utilitarismo negativo de TAMMELO) con el máximo bien posible para el mayor número de personas. La disculpa pasará así por criterios de justicia ya conocidos y que introducirán en el principio de culpa y en la disculpa la urgencia de nuevas reflexiones.