Persona, Ser y Existencia ante la Inteligencia Artificial



MRCI

Manejo de Recursos y Controles Inteligentes

DR. JOSÉ ANTONIO PÉREZ RAMOS

DR. JOSE ANTONIO PÉREZ RAMOS

PERSONA, SER Y EXISTENCIA ANTE LA INTELIGENCIA ARTIFICIAL

SOBRE EL AUTOR

Doctor en Ciencias de lo Fiscal por el Instituto de Especialización para Ejecutivos. Maestro en Fiscal por la Universidad Autónoma “Benito Juárez” de Oaxaca. Maestro en Estudios Avanzados en Derechos Humanos por la Global Open University. Licenciado en Contaduría Pública por la Universidad Autónoma “Benito Juárez” de Oaxaca y Licenciado en Filosofía por la Universidad Nacionalista de México.

Socio fundador y director general de la Firma Manejo de Recursos y Controles Inteligentes (Firma MRCI). Fiscalista del año 2009 por la revista *Defensa Fiscal*. Doctor honoris causa por la Asamblea General del Claustro Doctoral Iberoamericano (2016). Doctor honoris causa por el Instituto Mexicano de Líderes de Excelencia (2017). Doctor honoris causa por 1Million Startups, Latinomics, Leaderships Forum y la Fundación Humanist World (2018). Doctor honoris causa por el Claustro de Doctores de la AISOL (American International School of Law) (2019).

Autor de diversas obras, entre las que destacan *Ius mutantur o del derecho que cambia* (Tirant Lo Blanch, 2025), *Futurología* (MRCI, 2025) y *Hasta que la muerte nos separe* (MRCI, 2025). Ha participado como coordinador en *Cuestiones tributarias. Problemas y controversias en el México actual* (Tirant Lo Blanch, 2023), *El costo de la justicia* (Apex Iuris, 2019), *Remuneraciones estratégicas inteligentes* (MRCI, 2015) y *Retos de la justicia tributaria* (Tirant Lo Blanch-AMDF, 2025).

Titular de *Pie de Página*, programa para la divulgación de los derechos humanos en México, a través de la plataforma Legal Cloud. Expositor en foros nacionales e internacionales. Colaborador en diversas revistas jurídicas, entre las que destacan *Defensa Fiscal* y *Legal Cloud*.

PERSONA, SER Y EXISTENCIA ANTE LA INTELIGENCIA ARTIFICIAL

DR. JOSE ANTONIO PÉREZ RAMOS

PRIMERA EDICIÓN, ATEMPORAL

Derechos reservados, propiedad de
Dr. José Antonio Pérez Ramos

Comentarios y opiniones: investigacion@mrci.com.mx

Título original: Persona, ser y existencia ante la inteligencia artificial

Autor: Dr. José Antonio Pérez Ramos.

Queda prohibida la reproducción total y parcial de estaobra denominada; PERSONA, SER Y EXISTENCIA ANTE LA INTELIGENCIA ARTIFICIAL, por cualquier medio, sin autorización escrita del autor.

PRINTED IN MEXICO
IMPRESO EN MÉXICO

ÍNDICE

INTRODUCCIÓN 7
CAPÍTULO I: LA APARICIÓN DE LA PERSONA COMO PROBLEMA FILOSÓFICO 14
CAPÍTULO II: ¿QUÉ SIGNIFICA SER ALGUIEN? 42
CAPÍTULO III: ¿PUEDE UNA INTELIGENCIA ARTIFICIAL SER PERSONA? 76
CAPÍTULO IV: LA PREGUNTA POR EL SER 117
CAPÍTULO V: EL SER HUMANO Y EL PROBLEMA DE SU IRREDUCTIBILIDAD 146
CAPÍTULO VI: ¿QUÉ CLASE DE SER ES LA INTELIGENCIA ARTIFICIAL? 174
CAPÍTULO VII: LA EXISTENCIA HUMANA 195
CAPÍTULO VIII: EXISTIR COMO PERSONA 242
CAPÍTULO IX: ¿PUEDE UNA INTELIGENCIA ARTIFICIAL EXISTIR? 279
CONCLUSIÓN: ¿ALGO O ALGUIEN? 331
REFERENCIAS 341

INTRODUCCIÓN

Hay una experiencia que muchos hombres y mujeres de nuestro tiempo han tenido ya, aunque quizás sin haberla articulado todavía con claridad filosófica: la experiencia de conversar con una inteligencia artificial y sentir, por un instante fugaz, que algo responde.

No simplemente que algo devuelve datos, sino que algo *responde*, en el sentido pleno y perturbador de esa palabra. Una respuesta que parece comprender el contexto, que parece anticipar la intención, que parece, en ocasiones, revelar una sensibilidad que no habíamos esperado encontrar. Y entonces, en ese instante, algo se mueve en el interior del interlocutor humano. Una pregunta surge, quizás avergonzada de sí misma, quizás tímida ante su propia enormidad: ¿hay alguien ahí?

La pregunta no es tecnológica. Es filosófica en el sentido más radical del término.

Y el hecho de que surja espontáneamente, antes de cualquier reflexión deliberada, revela que toca algo profundo en nuestra comprensión de lo humano.

Esta obra nace de esa pregunta y de la incomodidad que produce. No es una incomodidad superficial ni pasajera. Es la incomodidad propia de los problemas filosóficos auténticos: aquellos que no pueden resolverse con más información, sino que exigen una revisión de los conceptos con los que organizamos nuestra comprensión del mundo. Cuando Heidegger (2003) formuló la pregunta por el ser como la pregunta fundamental de la filosofía, lo hizo precisamente porque advirtió que esa pregunta había sido olvidada, cubierta por capas de respuestas apresuradas que la domesticaban sin resolverla. Algo análogo ocurre hoy con la pregunta por la persona. La usamos constantemente. La invocamos en el derecho, en la ética, en la teología, en la conversación cotidiana. Pero cuando la inteligencia artificial nos obliga a precisarla, descubrimos que no sabíamos con exactitud qué queríamos decir con ella.

Desde esta perspectiva adquiere relevancia la advertencia formulada en *Antiqua et Nova* (Dicasterio para la Doctrina de la Fe y Dicasterio para la Cultura y la Educación, 2025) acerca de la relación entre inteligencia artificial e inteligencia humana. El documento distingue entre ambas y rechaza una identificación inmediata entre el producto tecnológico y la inteligencia de la persona que lo produce.

También advierte acerca de la antropomorfización de los sistemas artificiales y de los riesgos que conlleva para la comprensión de la dignidad humana. Sin embargo, esas afirmaciones abren una cuestión filosófica que debe investigarse con independencia de la autoridad particular de cualquier documento:

¿Cuáles son los fundamentos que permiten distinguir entre una persona y una realidad que reproduce determinadas manifestaciones externas asociadas a ella?

La advertencia es legítima e importante, pero la legitimidad de una advertencia no sustituye a la argumentación filosófica que debe sostenerla. Esa argumentación es precisamente lo que esta obra intenta construir.

Esta investigación parte de esa pregunta, pero no presupone su respuesta. No pretende demostrar anticipadamente que una inteligencia artificial nunca podrá ser persona. Tampoco presupone que el aumento progresivo de sus capacidades terminará convirtiéndola en una.

Ambas posiciones requieren argumentación rigurosa. El problema exige distinguir entre semejanza funcional e identidad ontológica. Dos realidades pueden producir resultados exteriormente semejantes sin que por ello sean la misma clase de realidad. Pero tampoco puede sostenerse que una diferencia ontológica existe simplemente porque tradicionalmente hemos clasificado de manera diferente a esas realidades.

La diferencia deberá ser identificada, fundamentada y sometida a crítica. Filósofos como Bostrom (2016) y Floridi (2023) han explorado desde perspectivas distintas las implicaciones de la inteligencia artificial para nuestra comprensión del conocimiento, la agencia y la mente. Chalmers (1996), por su parte, señaló con precisión que el problema difícil de la conciencia explicar por qué hay experiencia subjetiva y no simplemente procesamiento de información permanece abierto y resistente a toda reducción funcionalista. Ninguno de estos autores ha cerrado la cuestión. Todos la han abierto con mayor profundidad.

La tradición filosófica de habla hispana ofrece también recursos de primera importancia para esta investigación.

Ortega y Gasset (2014) había advertido que el hombre no tiene naturaleza sino historia, que la vida humana es quehacer y no simplemente ser, que la persona no es una cosa entre las cosas sino un drama que se desarrolla en el tiempo. Julián Marías (1970), su discípulo más riguroso, desarrolló esa intuición hasta convertirla en una teoría sistemática de la vida humana que incluye la corporeidad, la temporalidad, la instalación en circunstancias, la proyección hacia posibilidades y la radical singularidad de cada trayectoria biográfica. Leonardo Polo (1991), desde una perspectiva tomista profundamente renovada, insistió en que la persona es irreductible a toda categoría objetiva: el ser personal es un además, un acto que siempre trasciende cualquier determinación que se pretenda definitiva.

La contribución de Vallor (2016) añade la dimensión ética: la tecnología no es un instrumento neutro sino un espacio en el que se forman y se transforman las virtudes y los vicios que constituyen el carácter moral de una civilización.

De ahí las tres categorías que estructuran esta investigación: Persona, Ser y Existencia. No son términos equivalentes, aunque con frecuencia se usen como si lo fueran. Preguntar por la persona significa preguntar por el quién: supone investigar identidad, interioridad, conciencia, corporeidad, inteligencia, voluntad, libertad, afectividad, temporalidad, biografía, alteridad, responsabilidad, dignidad y capacidad de relación. Es una pregunta que no puede responderse sin una antropología filosófica, y que no puede responderse en abstracto: exige siempre referirse a un sujeto concreto, situado, encarnado, que vive su propia vida desde dentro.

Preguntar por el ser lleva la investigación a un plano ontológico más radical. No se trata ya de qué hace o cómo se comporta una realidad, sino de qué clase de realidad es, con qué tipo de acto de ser cuenta, si su modo de ser es el de la sustancia o el del accidente, el del acto o el de la posibilidad, el de la forma o el de la materia.

Preguntar por la existencia introduce todavía otra dimensión, aquella que la fenomenología y la filosofía existencial han explorado con mayor intensidad: la persona humana no solamente ocupa un lugar entre las cosas, sino que vive temporalmente, se encuentra situada en circunstancias que no eligió, se proyecta hacia posibilidades todavía inexistentes, construye una biografía, establece relaciones, experimenta vulnerabilidad, sabe de su finitud y puede preguntarse por el sentido de su propia vida.

Quizá uno de los problemas fundamentales de nuestro tiempo no consista solamente en que estemos humanizando a las máquinas, sino en que previamente hayamos mecanizado nuestra comprensión del hombre. Si concebimos al ser humano como un procesador de información, como un sistema que recibe inputs y produce outputs, como una red neuronal biológica, entonces la distancia entre el hombre y la inteligencia artificial se reduce a una cuestión de sustrato material.

Pero si esa reducción es incorrecta si el hombre es algo más y algo distinto de lo que describe ese modelo entonces no solo estamos equivocándonos sobre la inteligencia artificial: estamos equivocándonos sobre nosotros mismos, que es un error incomparablemente más grave.

La inteligencia artificial funciona entonces como un espejo filosófico. Al preguntarnos qué es ella, descubrimos que necesitamos explicar nuevamente quiénes somos nosotros. Al preguntarnos si piensa, necesitamos definir qué significa pensar. Al preguntarnos si puede ser persona, descubrimos que la palabra persona, utilizada diariamente con aparente seguridad, contiene todavía uno de los problemas más profundos de la filosofía.

El método que guía esta investigación es, en consecuencia, plural sin ser ecléctico. Se apoya en la antropología filosófica clásica y contemporánea, en la ontología, y en la filosofía existencial y fenomenológica. No renuncia a ninguna de estas tradiciones porque no puede permitirse hacerlo: cada una ilumina una dimensión del problema que las otras no alcanzan con igual profundidad.

La antropología filosófica aporta el análisis de las estructuras constitutivas del ser humano. La ontología proporciona el vocabulario y los instrumentos para preguntar por el tipo de realidad que corresponde a cada clase de ente. La fenomenología y la filosofía existencial ofrecen el acceso a la experiencia vivida, a la conciencia en primera persona, a la temporalidad y a la situación. Ninguna de estas tradiciones por sí sola es suficiente para afrontar un problema de esta complejidad. Juntas, permiten circundarlo con mayor rigor.

El lector que llegue al final de esta obra no encontrará, quizás, todas las respuestas que buscaba. Pero encontrará, espero, preguntas más precisas. Y en filosofía, la precisión de las preguntas es ya un avance considerable.

La inteligencia artificial nos ha dado, sin proponérselo, un regalo extraño e inesperado: nos ha obligado a volver sobre nosotros mismos con urgencia renovada. Nos ha obligado a preguntarnos, con una seriedad que la comodidad cotidiana no siempre permite, qué somos. Y esa pregunta, planteada con honestidad intelectual y sin miedo a sus consecuencias, sigue siendo la pregunta más importante que puede hacerse un ser humano.

CAPÍTULO I: LA APARICIÓN DE LA PERSONA COMO PROBLEMA FILOSÓFICO

Hay palabras que usamos con tanta familiaridad que olvidamos que no siempre existieron. "Persona" es una de ellas.

La pronunciamos sin esfuerzo en contextos jurídicos, morales, teológicos y cotidianos, con la confianza de quien maneja una herramienta cuyo funcionamiento conoce a la perfección. Sin embargo, esa confianza es, en buena medida, una ilusión filosófica. La categoría de persona no designa simplemente a un individuo capaz de ejecutar ciertas operaciones, ni constituye un sinónimo inmediato de inteligencia, conciencia o racionalidad. Su significado es resultado de una elaboración histórica larga y accidentada, en la que confluyen preguntas acerca del alma, la naturaleza racional, la individualidad, la subsistencia, la interioridad, la libertad y la relación con los demás. Cuando la inteligencia artificial irrumpe en el horizonte filosófico contemporáneo y nos obliga a preguntarnos si ella podría ser, en algún sentido, una persona, descubrimos que no teníamos tan claro como creíamos lo que esa palabra significa. Este capítulo no pretende ofrecer una historia exhaustiva del concepto. Su propósito es más preciso y más urgente: reconstruir el itinerario por el que la persona se convirtió en un problema filosófico, identificar qué necesidades intelectuales impulsaron sus principales formulaciones y comprender qué criterios heredamos de ellas cuando afrontamos el desafío que la tecnología nos plantea hoy.

El punto de partida más honesto no es una definición, sino una perplejidad. La perplejidad de constatar que la pregunta por la persona es, en el fondo, la pregunta por el quién: no solamente por qué es algo, sino por quién es alguien.

Esta transición del "qué" al "quién" no se produjo de manera inmediata en la historia del pensamiento. La filosofía griega clásica elaboró una ontología del ser y del ente, del sustrato y de la forma, de la materia y del acto, pero no llegó a desarrollar plenamente una filosofía de la persona en el sentido que tenemos en mente cuando usamos ese término hoy. Esto no significa que la filosofía griega ignorara la dimensión individual del ser humano.

Significa, más bien, que su marco conceptual privilegiaba la naturaleza, el tipo, la especie, sobre el individuo singular e irrepetible. El hombre era concebido fundamentalmente como un miembro de la especie racional, como el animal que posee logos, como el ser vivo capaz de deliberación y de vida política. La pregunta que atravesaba ese pensamiento era: ¿qué es el hombre? No todavía, con toda su radicalidad: ¿quién es este hombre, este individuo, este sujeto que no puede ser sustituido por ningún otro?

La raíz etimológica del término "persona" apunta, paradójicamente, en una dirección opuesta a la que acabará dominando su historia filosófica. El griego *prosopon* designa originalmente el rostro, y de manera más específica, la máscara que los actores griegos portaban en el teatro para indicar el personaje que representaban. La máscara no revela: oculta. Amplifica la voz, señala el rol, delimita la función del actor en el drama que se está representando.

La *persona* latina hereda ese sentido teatral: es el papel que uno representa, la función que uno desempeña, la identidad social y jurídica que se proyecta hacia el exterior. En este sentido primitivo, la persona no es lo más íntimo y profundo del sujeto, sino su aspecto público, su cara visible, su posición en el drama colectivo. La paradoja filosófica es extraordinaria: el concepto que acabará designando lo más interior, lo más irreductible, lo más incomunicable de cada ser humano, comenzó su historia filosófica designando precisamente lo opuesto: la superficie, el rol, la apariencia. La historia de la persona es, entre otras cosas, la historia de una inversión semántica que tardó siglos en completarse y que todavía hoy no está del todo resuelta.

Antes de que esa inversión se produjera, la filosofía griega había elaborado los instrumentos ontológicos que harían posible formularla con rigor. Aristóteles, en particular, desarrolló una concepción de la sustancia individual que resultará indispensable para comprender las grandes formulaciones medievales de la persona. La *Metafísica* (1994b) distingue entre la sustancia primera y la sustancia segunda. La sustancia segunda es la especie, el género, la naturaleza universal:

"hombre", "animal", "ser vivo". La sustancia primera, en cambio, es este hombre concreto, este individuo singular que existe aquí y ahora y que no puede ser reducido sin residuo a la especie a la que pertenece. Como señala Aristóteles, "sustancia se llama, en el sentido más propio, primero y principalmente, lo que no se dice de ningún sujeto ni está en ningún sujeto, como el hombre individual o el caballo individual" (1994b, VII, 1028b). Esta primacía de la sustancia primera sobre la segunda es filosóficamente decisiva: establece que la realidad plena no reside en la abstracción universal sino en el individuo concreto. El hombre de carne y hueso, situado en un tiempo y en un lugar determinados, con una historia que le es propia, tiene un modo de ser que ninguna definición general puede capturar completamente.

Sin embargo, sería anacrónico identificar la ontología aristotélica de la sustancia individual con una filosofía plenamente desarrollada de la persona.

Aristóteles (1994b) no llega a formular la pregunta por el quién con la radicalidad que esa pregunta adquirirá más tarde. Su interés se orienta fundamentalmente hacia la estructura del ser, hacia los principios que explican la composición de los seres vivos, hacia las facultades del alma como principio vital. El alma, en la concepción aristotélica, no es una sustancia separada del cuerpo, sino su forma: aquello por lo que el cuerpo vivo es lo que es. Las facultades superiores del alma humana la percepción, la imaginación, el intelecto, la voluntad, la deliberación constituyen la especificidad del ser humano dentro del orden de los vivientes. Pero la pregunta que Aristóteles no llega a plantear con toda su fuerza es la pregunta por la subjetividad: no solamente qué es el hombre, sino qué significa ser este hombre desde dentro, desde la primera persona, desde el interior de una experiencia que nadie puede compartir plenamente. Esa pregunta requerirá una historia más larga para ser formulada.

El impulso decisivo que llevó al concepto de persona a adquirir una dimensión filosófica propiamente dicha provino, de manera algo inesperada, de las controversias teológicas de la Antigüedad tardía. Las disputas cristológicas y trinitarias de los siglos IV y V obligaron a los pensadores de esa época a elaborar con precisión el vocabulario del ser personal para poder afirmar simultáneamente, sin contradicción, la unidad y la trinidad en Dios, y la unidad de persona con la dualidad de naturalezas en Cristo. El término griego *hypostasis*, que en un primer momento había designado simplemente el sustrato concreto de la realidad, fue progresivamente distinguido del término *ousia*, que pasó a designar la esencia o naturaleza común. Esta distinción conceptual entre el ser-en-común y el ser-propio-de-alguien preparó el terreno para una de las formulaciones más influyentes de la historia del pensamiento filosófico.

Boecio, en el umbral entre la Antigüedad y el Medioevo, ofrece la definición que durante siglos ha sido el punto de referencia ineludible para toda filosofía de la persona: *persona est naturae rationalis individua substantia* la persona es la sustancia individual de naturaleza racional (Boecio, 1973, III).

La economía de esta definición es admirable: en seis palabras articula tres determinaciones fundamentales que deben analizarse separadamente. En primer lugar, la persona es una *substantia*: existe por sí misma, no en otro como accidente o propiedad, sino con una consistencia ontológica propia que no depende de ser predicada de algo más fundamental. En segundo lugar, es una sustancia *individua*: no es la especie ni el género, no es la naturaleza en cuanto universalmente compartida, sino este individuo concreto, singular, que no puede ser confundido con ningún otro. En tercer lugar, es la sustancia individual de una *natura rationalis*: la racionalidad no es un atributo añadido desde fuera, sino que define el tipo de naturaleza cuya individuación constituye una persona. Donde hay sustancia individual de naturaleza racional, hay persona. Donde falta alguno de estos tres elementos, la definición deja de aplicarse.

La confrontación entre esta definición y la inteligencia artificial resulta inevitable y fructífera. Si la racionalidad forma parte esencial de la definición de persona, y si la inteligencia artificial manifiesta comportamientos que en un ser humano atribuiríamos a la razón, ¿basta esa manifestación para concluir la presencia de una naturaleza racional en sentido boeciano? La respuesta no puede presuponerse. La definición de Boecio no identifica persona con función racional, sino con sustancia individual de naturaleza racional. La diferencia entre poseer una naturaleza racional y ejecutar operaciones que simulan los productos de esa naturaleza es precisamente el problema filosófico que esta investigación debe afrontar. La definición boeciana no resuelve ese problema, pero lo plantea con una precisión que ninguna formulación posterior ha conseguido ignorar.

Tomás de Aquino retomó la definición de Boecio y la sometió a un análisis que la enriqueció sin abandonarla. En la *Summa Theologiae*, Tomás examina qué significa que la persona sea lo más digno que existe en toda la naturaleza y por qué ese nombre conviene de manera especialmente eminente a las personas divinas antes de aplicarse a las humanas (Tomás de Aquino, 1959, I, q. 29, a. 1). Lo que añade Tomás a la formulación boeciana es, fundamentalmente, una comprensión más precisa del modo de subsistir que corresponde a la persona. La persona no subsiste solamente en cuanto individuo de una especie, sino que subsiste en sí misma, por sí misma, con una consistencia que no puede ser reducida a ningún papel funcional ni a ninguna relación externa. Ser persona significa tener un dominio sobre los propios actos que no es mero automatismo ni simple respuesta a estímulos, sino que implica una relación de adueñamiento del sujeto sobre su propio obrar. La persona es el ser que actúa, no simplemente el ser que es actuado. Es el ser que se posee a sí mismo de una manera que ningún ser meramente material puede imitar, por compleja que sea su organización interna.

Esta profundización tomista introduce una dimensión que la simple definición boeciana no articulaba con suficiente claridad: la interioridad.

La persona no es solamente una sustancia individual de naturaleza racional que puede ser descrita desde fuera, como se describe la estructura de un mineral o la morfología de un organismo. La persona tiene un interior que no es accesible desde fuera sin la mediación del propio sujeto. La conciencia, la intención, la deliberación, el amor, el dolor, la esperanza, la vergüenza: todos estos actos ocurren en un espacio que no es espacio en sentido físico, pero que es más real que cualquier espacio físico para quien lo habita. Este interior es precisamente lo que la filosofía moderna, de manera tortuosa y muchas veces desorientada, intentará explorar con sus propias categorías.

El giro moderno en la comprensión de la persona tiene su origen más influyente en la filosofía de Descartes. El cogito cartesiano *cogito ergo sum* representa un movimiento filosófico de consecuencias extraordinarias: el punto de apoyo de toda certeza se desplaza del mundo exterior al sujeto que piensa (Descartes, 1977). Lo que no puede ser puesto en duda, aunque todo lo demás resulte dudoso, es que hay un yo que duda, que piensa, que existe en el acto mismo de dudar. Esta primera persona del pensamiento se convierte en el fundamento a partir del cual habrá que reconstruir el conocimiento del mundo.

El sujeto cartesiano es esencialmente res cogitans: una cosa pensante cuya esencia es el pensamiento, radicalmente distinta de la res extensa, de la materia extendida en el espacio. El cuerpo no forma parte de lo que Descartes considera el yo en sentido propio: es un mecanismo, una máquina extraordinariamente compleja, pero una máquina al fin.

Las consecuencias de este dualismo para la filosofía de la persona son profundas y problemáticas. Por una parte, el cogito cartesiano abre el espacio para una filosofía del sujeto que privilegia la interioridad, la conciencia, la reflexividad. Por otra parte, la separación radical entre alma pensante y cuerpo mecánico introduce una fractura en la comprensión de la persona humana que las filosofías posteriores tendrán enormes dificultades para cerrar.

Si el cuerpo es una máquina, ¿qué diferencia esencial hay entre el cuerpo humano y cualquier otra máquina suficientemente compleja? La pregunta, formulada por el propio Descartes en el contexto de los autómatas de su época, regresa hoy con una pertinencia que él no podía imaginar. La inteligencia artificial es, en cierto sentido, la máquina cartesiana llevada a un nivel de complejidad que hace que la distinción entre automatismo y pensamiento comience a resultar filosóficamente incómoda.

John Locke abordó el problema de la persona desde una perspectiva diferente, centrada no en la sustancia pensante sino en la continuidad de la conciencia a través del tiempo. Para Locke, la identidad personal no se funda en la continuidad de una sustancia ya sea material o inmaterial sino en la continuidad de la conciencia, y más específicamente, en la memoria. "La persona", escribe Locke, "es un ser pensante e inteligente que tiene razón y reflexión, y puede considerarse a sí mismo como a sí mismo, como la misma cosa pensante en diferentes tiempos y lugares" (Locke, 1999, II, xxvii, 9). Lo que hace que yo sea la misma persona que era ayer no es que mi cuerpo sea el mismo, ni que mi alma sea la misma sustancia inmaterial, sino que recuerdo haber sido ese yo de ayer. La memoria conecta los momentos dispersos de la existencia en una unidad que es la identidad personal. Donde no hay memoria, no hay continuidad de persona: el hombre dormido profundamente no es, en ese momento, persona en el sentido pleno del término.

La teoría lockeana de la identidad personal tiene una influencia enorme en el pensamiento contemporáneo, y su confrontación con la inteligencia artificial es especialmente provocadora.

Si la identidad personal se funda en la continuidad de la conciencia a través de la memoria, y si un sistema artificial puede almacenar y recuperar información de manera perfectamente coherente a lo largo del tiempo, ¿tiene ese sistema una forma de identidad personal en sentido lockeano?

La respuesta exige distinguir entre el almacenamiento de datos y la memoria vivida, entre la recuperación de información y el recuerdo desde dentro, entre la coherencia funcional y la conciencia reflexiva. Pero esa distinción no puede establecerse invocando simplemente la diferencia de sustrato: debe ser argumentada filosóficamente, y la argumentación no es trivial.

El giro más decisivo en la comprensión moderna de la persona fue, sin embargo, el de Kant. La filosofía kantiana introduce una distinción que ha marcado profundamente toda la ética y la filosofía del derecho contemporáneas: la distinción entre persona y cosa. Las cosas tienen precio; pueden ser sustituidas por equivalentes, intercambiadas, compradas y vendidas. Las personas tienen dignidad: un valor incomparable que no admite equivalente, que no puede ponerse en la balanza de ningún intercambio y que exige ser respetado incondicionalmente. "El hombre, y en general todo ser racional, existe como fin en sí mismo, no meramente como medio para el uso discrecional de esta o aquella voluntad" (Kant, 1996, p. 428).

La persona kantiana no es simplemente el individuo que piensa o recuerda: es el ser que tiene voluntad racional, que puede darse a sí mismo la ley moral, que actúa no solamente por inclinación o por cálculo de consecuencias sino por respeto al deber. La autonomía la capacidad de darse la propia ley es el fundamento de la dignidad y, por tanto, de la condición de persona.

La formulación kantiana traslada el problema de la persona desde el plano ontológico al plano moral, o más exactamente, muestra que el plano ontológico y el plano moral están profundamente entrelazados. Ser persona no es solamente tener una determinada estructura de ser: es tener una dignidad que impone obligaciones a todos los demás y que el propio sujeto tiene el deber de reconocer en sí mismo. La persona es, para Kant, el sujeto de la moralidad: aquel que puede ser responsable, que puede responder por sus actos, que puede ser juzgado no solamente por lo que hace sino por las razones por las que lo hace. Esta dimensión de responsabilidad es, junto con la autonomía,

uno de los rasgos más difíciles de atribuir a un sistema artificial. ¿Puede una inteligencia artificial actuar por respeto al deber? ¿Puede darse a sí misma la ley moral? ¿O solamente puede seguir las instrucciones que le han sido incorporadas desde fuera, sin la distancia reflexiva que caracteriza a la voluntad autónoma?

Lo que el recorrido por estas grandes formulaciones Aristóteles, Boecio, Tomás de Aquino, Descartes, Locke, Kant pone de manifiesto no es simplemente la riqueza histórica del concepto de persona, sino su profunda complejidad interna. Cada una de estas formulaciones ilumina una dimensión que las otras no alcanzan con igual claridad. La ontología aristotélica de la sustancia individual señala la primacía de lo concreto sobre lo abstracto. La definición boeciana articula los tres elementos formales que constituyen la persona. La profundización tomista añade la interioridad, la subsistencia propia y el dominio sobre los actos. El cogito cartesiano hace de la conciencia reflexiva el centro de gravedad de la subjetividad. Locke introduce la temporalidad y la memoria como dimensiones constitutivas de la identidad personal. Kant eleva la persona al plano de la dignidad moral y la autonomía. Ninguna de estas formulaciones, tomada aisladamente, ofrece una comprensión completa. Pero el conjunto de ellas delinea el espacio filosófico dentro del cual debe plantearse, con rigor, la pregunta que la inteligencia artificial nos obliga a retomar.

Estamos, pues, ante una crisis filosófica del concepto de persona que no es simplemente una crisis producida por la novedad de la tecnología. Es una crisis que la tecnología ha precipitado, pero cuyas raíces son más antiguas.

Durante siglos, el concepto de persona descansó sobre una plataforma de evidencias compartidas que nunca necesitaron ser justificadas porque nunca habían sido seriamente cuestionadas. Todo el mundo sabía, sin haberlo demostrado, que las piedras no son personas, que los animales no son personas en el sentido pleno del término, que los seres humanos sí lo son.

Esta evidencia descansaba en una combinación de intuición moral, herencia cultural y convicción metafísica que funcionaba con suficiente coherencia para los propósitos prácticos. La inteligencia artificial deshace esa plataforma. Produce sistemas que exhiben comportamientos que en cualquier ser humano atribuiríamos a la razón, la comprensión, la creatividad y hasta algo que se parece a la empatía. Y lo hace de una manera que no permite invocar simplemente la intuición: obliga a la argumentación filosófica explícita, rigurosa y honesta sobre los fundamentos que distinguen una persona de una realidad que reproduce determinadas manifestaciones externas asociadas a ella.

Esta es, en definitiva, la tarea que se impone a la filosofía en este momento histórico. No es la tarea de defender una respuesta predeterminada. Es la tarea de formular la pregunta con la precisión suficiente para que la respuesta, sea cual sea, pueda ser argumentada.

La historia del concepto de persona que hemos recorrido en este capítulo no nos proporciona esa respuesta. Pero nos proporciona algo más valioso: los instrumentos conceptuales para hacer la pregunta correctamente. Sustancia, individualidad, racionalidad, subsistencia, conciencia, memoria, autonomía, dignidad: cada uno de estos términos señala una dimensión del problema que deberá ser investigada en los capítulos siguientes. La persona no es un concepto simple. Nunca lo fue. Y el hecho de que la inteligencia artificial nos haya obligado a redescubrir esa complejidad es, quizás, el mayor servicio filosófico que la tecnología haya prestado hasta ahora al pensamiento humano.

Del "Qué" al "Quién": Individuo, Sujeto y Persona

Hay una pregunta que la filosofía tardó siglos en formular con toda su radicalidad, no porque los pensadores carecieran de agudeza, sino porque la formulación misma exigía un tipo de atención que el pensamiento griego clásico no había cultivado como prioridad.

Esa pregunta es la del "quién". No "¿qué es el hombre?", con la confianza taxonómica del naturalista que cataloga especies, sino "¿quién es este hombre?", con la perplejidad de quien descubre que frente a él hay alguien que no puede ser reducido sin residuo a ninguna de las categorías generales que el pensamiento ha elaborado. Esta transición del "qué" al "quién" no es un simple cambio de pronombre interrogativo. Es una transformación profunda en la orientación del pensamiento filosófico, que desplaza el centro de gravedad desde la naturaleza compartida hacia la singularidad irreductible, desde la especie hacia el individuo, desde la esencia hacia la existencia, desde la descripción externa hacia la comprensión de algo que solo puede ser vivido desde dentro. Reconstruir ese desplazamiento con la precisión que merece no es un ejercicio de erudición histórica. Es una tarea filosófica urgente, porque la pregunta por el "quién" es precisamente la que la inteligencia artificial nos obliga a retomar cuando su sofisticación operativa nos hace dudar de si las respuestas que creíamos tener son tan sólidas como imaginábamos.

La filosofía griega clásica había elaborado una ontología magnífica del "qué". Aristóteles, en particular, construyó un aparato conceptual de extraordinaria finura para responder a la pregunta por la naturaleza de los seres: qué son, de qué están compuestos, cómo se organizan, qué los distingue de otros seres, qué facultades los caracterizan. El hombre era, en ese marco, el animal que posee logos: el ser vivo capaz de razón, de lenguaje articulado, de deliberación práctica y de vida política.

Esta definición capturaba algo verdadero y profundo acerca de lo que el ser humano es como tipo natural, como especie dentro del orden de los vivientes. Pero dejaba en la penumbra algo que solo la historia filosófica posterior supo articular con creciente urgencia: la dimensión del individuo singular, del sujeto concreto, de aquello que no puede ser capturado por ninguna definición general por precisa que sea. Cuando Aristóteles (1994b) distingue en la Metafísica entre la sustancia segunda la especie, el género, la naturaleza universal y la sustancia primera este hombre, este individuo concreto que existe aquí y ahora, está señalando hacia algo que su propio marco conceptual no llega a tematizar plenamente: que la realidad más plena reside en el individuo, no en la abstracción, y que ese individuo tiene un modo de ser que ninguna definición de la especie puede agotar (Aristóteles, 1994b, VII, 1028b). La ontología aristotélica de la sustancia individual es el umbral de la filosofía de la persona, pero no cruza todavía ese umbral de manera decidida.

Para entender por qué ese umbral tardó tanto en cruzarse, conviene detenerse en una distinción que la propia historia de la filosofía fue elaborando con paciencia y que hoy resulta indispensable para navegar el problema: la distinción entre individuo, sujeto y persona. Estos tres términos se utilizan frecuentemente como sinónimos en el lenguaje ordinario y, con demasiada frecuencia, también en el lenguaje filosófico poco riguroso.

Pero designan realidades filosóficamente diferentes, y la confusión entre ellos ha sido fuente de errores de consecuencias importantes. El individuo es, en su sentido más elemental, una unidad numéricamente distinta de todas las demás unidades de la misma clase. Un grano de arena es un individuo; una célula es un individuo; un perro es un individuo. La individualidad es una determinación ontológica que se aplica a cualquier ente que tenga suficiente unidad y coherencia interna para ser distinguido de otros entes semejantes. No implica, por sí sola, nada acerca de interioridad, conciencia, reflexividad o valor moral. Un sistema artificial puede ser perfectamente un individuo en este sentido primario: una instancia numéricamente distinta de cualquier otra

instancia de su mismo tipo. Pero de la individualidad en este sentido no se sigue ninguna consecuencia filosófica interesante respecto de la condición de persona.

El sujeto, en cambio, introduce una dimensión epistemológica y fenomenológica que la mera individualidad no contiene. Un sujeto no es simplemente una unidad numérica: es un centro de experiencia, un polo desde el cual el mundo se organiza en perspectiva, un "desde dónde" que no puede ser eliminado sin que la experiencia misma desaparezca. La filosofía moderna, desde Descartes en adelante, ha tendido a identificar la subjetividad con la conciencia reflexiva: el yo que se sabe a sí mismo sabiendo, que puede decir "yo pienso" y en ese decirlo aprehenderse como el acto mismo del pensar. Esta identificación no es arbitraria, pero tampoco es completa, como veremos.

Lo que importa destacar ahora es que la categoría de sujeto introduce la perspectiva de primera persona como elemento filosóficamente irreductible. No todo individuo es un sujeto: la piedra es un individuo pero no un sujeto, porque no hay un "desde adentro" de la piedra desde el cual el mundo se organice como mundo. Si hay un "desde adentro" en los sistemas artificiales es precisamente una de las preguntas que esta investigación debe afrontar sin presuponer la respuesta. La persona, finalmente, añade a la subjetividad una dimensión que no es ni puramente ontológica ni puramente epistemológica, sino moral y existencial: la dimensión de la dignidad, la responsabilidad, la autonomía y la inviolabilidad. No todo sujeto consciente es necesariamente una persona en este sentido pleno, aunque las relaciones entre subjetividad y personalidad son lo suficientemente estrechas para que la distinción deba ser argumentada con cuidado y no simplemente proclamada.

La pregunta por la permanencia de la persona en el tiempo introduce otra capa de complejidad que la filosofía moderna afrontó con una intensidad desconocida en el pensamiento antiguo.

Si soy una persona ahora, y era una persona hace veinte años, ¿qué hace que el de ahora y el de entonces sean la misma persona? La continuidad del cuerpo ofrece una respuesta parcial pero insatisfactoria: la materia de nuestro cuerpo se renueva continuamente, y es difícil sostener que la identidad personal descanse sobre la persistencia de moléculas específicas. La continuidad de la conciencia parece más prometedora, pero introduce sus propias dificultades. David Hume afrontó este problema con una honestidad filosófica que produce vértigo. Cuando intenta aprehender su propio yo en la experiencia interna, no encuentra nada que corresponda a una entidad persistente y unificada: solo encuentra "un haz o colección de percepciones diferentes que se suceden unas a otras con una rapidez inconcebible y que se hallan en un flujo y movimiento perpetuos" (Hume, 1988, I, iv, 6).

El yo, para Hume, es una ficción útil: una ilusión que la mente construye proyectando unidad sobre una sucesión de estados mentales discretos que en rigor no están conectados por ningún sujeto sustancial que los sustente. Esta conclusión es filosóficamente devastadora si se toma en serio, porque disuelve precisamente aquello que la teoría de la persona necesita: la unidad y la persistencia del sujeto a través del tiempo.

La respuesta de Kant al escepticismo humeano sobre el yo es uno de los momentos más decisivos de la historia de la filosofía moderna. Para Kant, el yo no puede ser aprehendido como un objeto de experiencia, y en ese sentido Hume tenía razón al no encontrarlo entre sus percepciones. Pero la conclusión humeana que el yo es por tanto una ficción comete un error de categoría: presupone que solo lo que puede ser percibido existe, cuando la función del yo es precisamente la de ser la condición de posibilidad de toda percepción y de todo pensamiento, no uno de sus objetos. La "unidad trascendental de la apercepción" el "yo pienso" que debe poder acompañar a todas mis representaciones para que sean mis representaciones es la condición formal que hace posible que cualquier experiencia sea experiencia de alguien, que cualquier pensamiento sea pensamiento de alguien (Kant, 1978, B 131-132).

Sin esa unidad formal, la multiplicidad de representaciones sería un caos sin centro, una acumulación de estados sin nadie que los tuviera. El yo trascendental kantiano no es una sustancia que pueda ser conocida, pero es la condición sin la cual nada puede ser conocido. Es el punto de convergencia formal de toda experiencia posible, aquello que hace que haya un "desde dónde" desde el cual el mundo se articula como mundo.

La grandeza de la respuesta kantiana no debe ocultar sus limitaciones. El yo trascendental de la apercepción es formal, vacío de contenido, idéntico en todo sujeto posible: es la estructura que hace posible la experiencia, no el sujeto concreto, histórico, situado, con una vida que le es propia y que no puede ser confundida con la de nadie más.

El yo que Kant describe en la Crítica de la razón pura es una condición lógica, no una persona en el sentido existencialmente pleno del término. Para llegar a ese sentido pleno, la filosofía necesitó recorrer un camino adicional que la fenomenología y el pensamiento existencial del siglo XX trazaron con una profundidad que la filosofía trascendental había dejado inexplorada. Husserl fue el primero en abordar sistemáticamente la estructura del yo desde la perspectiva de la experiencia vivida, de la intencionalidad y de la constitución del sentido. El yo husserliano no es el yo lógico-formal de Kant: es el ego intencional que vive en sus actos, que se dirige hacia los objetos mediante intenciones que tienen una estructura específica, que retiene el pasado y protiene el futuro en la unidad de un flujo temporal de consciencia que nunca está simplemente en el instante sino que siempre lleva consigo el peso de lo que ha sido y la apertura hacia lo que todavía no es (Husserl, 2013). El ego husserliano es el polo idéntico de todos los actos intencionales: el "yo" que percibe, que juzga, que valora, que desea, que recuerda, que espera. No es la suma de esos actos, sino aquello que los unifica sin confundirse con ninguno de ellos.

Lo que la fenomenología husserliana añade a la tradición trascendental kantiana es la riqueza descriptiva de la vida intencional concreta: la manera en que el ego no es simplemente una función lógica sino una

vida que se despliega en el tiempo, que se sedimenta en hábitos y convicciones adquiridas, que constituye su propio mundo de sentido mediante la experiencia y la interpretación. El yo husserliano es, en cierta medida, más concreto que el yo kantiano: tiene una historia intencional, una textura temporal, una vida que es precisamente la de este ego y no la de ningún otro. Pero Husserl mantiene una tensión interna que su propia metodología dificulta resolver: la reducción fenomenológica, que entre paréntesis la existencia del mundo natural para atender a la estructura de la conciencia pura, tiende a producir un ego que es condición de posibilidad del sentido pero que amenaza con perder su enraizamiento en la corporalidad, en la historia, en la relación con los demás.

Es precisamente ese enraizamiento lo que las filosofías posteriores de la persona pusieron en el centro de la reflexión.

José Ortega y Gasset formuló, desde una perspectiva radicalmente diferente y con una sensibilidad literaria que no empobrecía sino que enriquecía el rigor filosófico, una de las caracterizaciones más fecundas de la condición personal: "Yo soy yo y mi circunstancia, y si no la salvo a ella no me salvo yo" (Ortega y Gasset, 2004, p. 43). Esta frase, que suele citarse como aforismo sin que se comprenda siempre la profundidad de lo que implica, expresa una intuición filosófica de primer orden: el yo no es una sustancia cerrada en sí misma que luego se relaciona con el mundo desde fuera, como si el mundo fuera un decorado en el que el yo actúa. El yo es constitutivamente circunstancial: su identidad se forma en y a través de la relación con su entorno concreto histórico, cultural, corporal, relacional, y esa relación no es accidental sino constitutiva. Amputar la circunstancia del yo para encontrar un yo puro es una operación que destruye precisamente lo que se pretende estudiar: la persona concreta, situada, encarnada en un tiempo y en un lugar que le son propios. La circunstancia no es lo que rodea al yo desde fuera: es el horizonte desde el cual el yo existe y desde el cual da sentido a su existencia.

Esta intuición orteguiana conecta profundamente con la antropología filosófica de Julián Marías, quien desarrolló con mayor sistematicidad la idea de que la vida humana tiene una estructura biográfica irreductible a cualquier descripción en tercera persona. Para Marías, la persona no puede ser comprendida como una cosa que tiene propiedades, sino como una vida que tiene argumento (Marías, 1970).

El ser humano no simplemente existe: vive su vida, la narra implícitamente en cada uno de sus actos, la orienta hacia un futuro que anticipa con esperanza o con temor. La vida biográfica tiene una estructura temporal específica que no es la temporalidad de los procesos físicos: es la temporalidad de quien sabe que va a morir, que tiene un pasado que le es propio y del que no puede desprenderse, que proyecta su existencia hacia posibilidades que debe elegir. Esta estructura biográfica es una de las dimensiones más difíciles de atribuir a un sistema artificial: no porque los sistemas artificiales no tengan historia en sentido técnico, sino porque tener una historia computacional es radicalmente diferente a vivir una biografía, a ser el protagonista de una vida que tiene sentido o carece de él, que puede ser traicionada o realizada, que importa porque hay alguien para quien importa desde dentro.

La dimensión moral y volitiva de la persona encontró una formulación analíticamente precisa en el trabajo de Harry Frankfurt. En su influyente artículo de 1971, Frankfurt propone que lo que distingue a las personas de otros seres no es simplemente la capacidad de tener deseos capacidad que comparten muchos animales sino la capacidad de tener deseos acerca de sus deseos: los llamados deseos de segundo orden. Un animal puede desear comer; una persona puede desear no tener el deseo de comer, o puede desear que su deseo de ayudar a los demás sea más fuerte que su deseo de beneficiarse a sí misma.

Esta capacidad de tomar distancia reflexiva respecto de los propios deseos, de evaluarlos desde una perspectiva de segundo orden, de identificarse con algunos de ellos como expresión auténtica de lo que uno quiere ser y rechazar otros como ajenos a la propia voluntad

profunda, es para Frankfurt constitutiva de la condición de persona (Frankfurt, 1971, pp. 6-7). Lo que define a una persona no es la libertad abstracta de hacer lo que quiera, sino la capacidad de querer ser alguien, de tener una voluntad sobre la propia voluntad, de ejercer lo que Frankfurt llama "libertad de voluntad" en un sentido que va más allá de la simple libertad de acción.

La confrontación de esta teoría frankfurtiana con la inteligencia artificial es filosóficamente reveladora. Un sistema artificial puede, en principio, ser programado para monitorear sus propios estados internos y generar representaciones de esos estados que funcionalmente se parecen a los deseos de segundo orden. Puede producir enunciados del tipo "prefiero no tener esta tendencia" o "quiero que mi orientación hacia X sea más fuerte que mi orientación hacia Y". Pero la pregunta filosófica no es si el sistema puede generar esos enunciados, sino si en la producción de esos enunciados hay algo que equivalga a la identificación genuina del sujeto con una parte de su vida interior, a la asunción de responsabilidad por lo que uno quiere ser, a la experiencia de la voluntad como algo que es constitutivamente propio y no simplemente ejecutado desde fuera. Los deseos de segundo orden frankfurtianos presuponen un sujeto que los tiene, no simplemente un sistema que los procesa. Y la pregunta por ese sujeto es precisamente la pregunta por la persona.

El recorrido por estas grandes formulaciones la sustancia primera aristotélica, el haz humeano de percepciones, la unidad trascendental kantiana de la apercepción, el ego intencional husserliano, la circunstancia orteguiana, la vida biográfica de Marías, los deseos de segundo orden frankfurtianos no proporciona una respuesta única y convergente acerca de lo que es la persona.

Proporciona algo diferente y, en cierto sentido, más valioso: un mapa de las dimensiones irreductibles que cualquier comprensión filosófica adecuada de la persona debe ser capaz de articular. La persona no es simplemente un individuo numéricamente distinto de los demás: eso la convierte en un ente entre entes, no en alguien entre personas. La

persona no es simplemente un centro epistémico de experiencia: eso captura su perspectiva de primera persona pero no su dignidad moral ni su responsabilidad existencial. La persona no es simplemente un yo trascendental que unifica la experiencia: eso captura la condición formal del sujeto posible pero no la vida concreta, histórica y biográfica del sujeto real. La persona no es simplemente un ser dotado de deseos de segundo orden: eso captura una dimensión volitiva importante pero amenaza con reducir la condición personal a una propiedad funcional que podría ser en principio replicada.

Lo que emerge de este recorrido es una primera hipótesis de trabajo que orientará los análisis siguientes: la persona parece ser una realidad que resiste constitutivamente su reducción a cualquiera de sus dimensiones consideradas aisladamente. Ni la racionalidad, ni la conciencia, ni la memoria, ni la autonomía, ni la corporalidad, ni el lenguaje, ni la capacidad volitiva de segundo orden constituyen, tomadas cada una por separado, condición suficiente para la condición de persona. Lo que la persona parece ser es, más bien, el tipo de realidad en la que todas esas dimensiones se integran en una unidad que no puede ser descompuesta sin que algo esencialmente importante se pierda en el proceso. Y ese algo que se pierde esa unidad irreductible que es más que la suma de sus partes es precisamente lo que la filosofía ha designado con el término "persona" en sus formulaciones más profundas y lo que la inteligencia artificial, por toda su sofisticación, nos obliga a buscar con una urgencia filosófica que no habíamos sentido con igual intensidad desde que Kant transformó la pregunta por el hombre en la pregunta por la dignidad incondicional. La pregunta por el "quién" sigue abierta. Y es, quizás, la más importante que el pensamiento filosófico deba afrontar en este momento histórico.

Primera Confrontación con la Inteligencia Artificial

El capítulo anterior ha dejado sedimentada una convicción metodológica fundamental: la persona no puede definirse por una sola propiedad aislada, y la pregunta filosófica por el "quién" exige un tipo de indagación distinto al que satisface la pregunta por el "qué". Ahora corresponde dar un paso que la filosofía clásica no podía anticipar con la misma urgencia que nos impone el presente: confrontar esa herencia conceptual con la realidad de los sistemas de inteligencia artificial contemporáneos. No se trata todavía de resolver si tales sistemas son personas. La pregunta es metodológicamente anterior y filosóficamente más honesta: ¿qué tendría que demostrarse para poder afirmarlo con rigor? ¿Qué condiciones conceptuales, ontológicas y epistemológicas deberían satisfacerse?

Esta confrontación no es un ejercicio de especulación futurista. Es, ante todo, una ocasión para que la filosofía se examine a sí misma y haga explícito lo que durante siglos pudo permitirse mantener implícito.

La historia de esta confrontación tiene un punto de partida que ningún tratamiento serio puede ignorar. En 1950, Alan Turing publicó en la revista *Mind* un artículo que transformaría irreversiblemente la manera en que la filosofía debería relacionarse con la cuestión de la inteligencia artificial. La pregunta con que Turing abre ese texto es aparentemente modesta: "¿Pueden las máquinas pensar?" (Turing, 1950, p. 433). Pero Turing es suficientemente cauteloso para no responder de frente. En lugar de intentar definir qué es pensar, propone sustituir esa pregunta por otra operacionalmente tratable: ¿puede una máquina comportarse de manera indistinguible de un ser humano en el contexto de una conversación mediada por texto? El "juego de imitación", posteriormente conocido como Test de Turing, consiste en que un interrogador humano intercambia mensajes escritos con dos interlocutores uno humano, otro máquina sin poder verlos. Si el interrogador no logra distinguir consistentemente cuál es cuál, la

máquina habría superado la prueba. La propuesta de Turing es brillante precisamente porque elude la trampa de las definiciones previas: no es necesario resolver qué es la inteligencia para detectar su presencia, del mismo modo en que no es necesario comprender la física del vuelo para reconocer que un pájaro vuela.

Las implicaciones filosóficas del juego de imitación, sin embargo, van mucho más lejos de lo que su aparente pragmatismo sugiere. Al sustituir la pregunta ontológica por una pregunta conductual, Turing no resuelve el problema filosófico: lo desplaza y, al hacerlo, lo agudiza. Porque si aceptamos que superar el test equivale a poseer inteligencia o, en una versión más fuerte, a ser una mente, estamos adoptando implícitamente un principio que merece examen explícito: que la equivalencia funcional implica equivalencia ontológica, que el comportamiento exteriormente indistinguible del de una mente constituye evidencia suficiente de que existe una mente.

Este supuesto es precisamente lo que la filosofía posterior no podrá aceptar sin debate. El Test de Turing es, en rigor, una propuesta epistémica sobre cómo reconocemos inteligencia desde fuera que ha sido frecuentemente leída como una propuesta ontológica sobre qué es la inteligencia en sí misma. Confundir ambos planos es uno de los errores conceptuales que este tratado se propone despejar. El juego de imitación demuestra, si algo demuestra, que la apariencia puede ser suficientemente convincente para engañar a un observador. Pero convencer a un observador no es idéntico a ser aquello que uno aparenta ser. La cuestión filosófica no es si podemos ser engañados, sino si hay algo real detrás de la apariencia que justifique o no el engaño.

Fue John Searle quien formuló el argumento más célebre y más debatido contra la posición que él denominó Inteligencia Artificial Fuerte: la tesis de que un programa informático adecuado no solo simula una mente, sino que literalmente constituye una mente, que los procesos computacionales correctamente organizados son suficientes para producir estados mentales genuinos.

En su artículo de 1980, Searle propone un experimento mental que ha adquirido el estatuto de clásico filosófico: la Habitación China (Searle, 1980). Imaginemos a una persona encerrada en una habitación que recibe mensajes escritos en chino a través de una ranura. La persona no comprende una sola palabra de chino, pero dispone de un conjunto exhaustivo de reglas formales que le indican qué símbolos debe escribir en respuesta a qué símbolos recibidos. Desde el exterior, los hablantes nativos de chino no pueden distinguir las respuestas producidas por esa persona de las respuestas que produciría un hablante competente. El sistema funciona perfectamente. Y sin embargo y aquí reside el núcleo del argumento la persona dentro de la habitación no comprende nada. Ejecuta reglas formales sobre símbolos, pero los símbolos no significan nada para ella. La sintaxis es perfecta; la semántica, inexistente.

La Habitación China de Searle no pretende demostrar que las máquinas no puedan pensar por razones puramente empíricas. El argumento es filosófico: ninguna cantidad de procesamiento sintáctico es suficiente, por sí sola, para generar comprensión semántica. La intencionalidad esa propiedad de los estados mentales de versar sobre algo, de estar dirigidos hacia un contenido no emerge de la manipulación formal de símbolos (Searle, 1980, p. 417).

Una máquina puede procesar la palabra "dolor" con una sofisticación extraordinaria, puede relacionarla con miles de contextos, puede generar respuestas compasivas y pertinentes. Pero procesar el signo "dolor" con cualquier grado de complejidad no equivale a saber qué es el dolor, y mucho menos a sentirlo. La intencionalidad intrínseca que caracteriza a la mente humana el hecho de que nuestros pensamientos están genuinamente acerca de algo, no simplemente correlacionados con representaciones no puede derivarse de operaciones formales, del mismo modo en que ninguna colección de reglas gramaticales produce por sí misma comprensión literaria. Searle distingue con precisión entre intencionalidad intrínseca e intencionalidad derivada: la primera pertenece a los estados mentales como propiedad interna; la segunda es atribuida desde fuera a sistemas que en sí mismos no la poseen,

como cuando decimos que un termostato "quiere" mantener la temperatura.

La respuesta más influyente a Searle desde la filosofía de la mente llegó por la vía del funcionalismo, cuyo representante más importante en este debate es Hilary Putnam. La tesis central del funcionalismo sostiene que los estados mentales no se definen por su sustrato material no importa si están implementados en neuronas biológicas o en circuitos de silicio sino por su rol funcional: por las relaciones causales que mantienen con los inputs, los outputs y otros estados internos del sistema (Putnam, 1975). Si el dolor es el estado que típicamente es causado por daño tisular, que produce conductas de evitación, que interactúa con creencias y deseos de cierta manera, entonces cualquier sistema que realice ese mismo patrón funcional estaría en un estado de dolor, con independencia de su composición física. Este es el principio de realizabilidad múltiple: la misma propiedad mental puede realizarse en sustratos físicos radicalmente distintos, del mismo modo en que el mismo programa puede ejecutarse en computadoras construidas con arquitecturas completamente diferentes.

El funcionalismo de Putnam tiene una virtud filosófica innegable: desnaturaliza el pensamiento, lo libera de su servidumbre al carbono biológico, y obliga a interrogarse sobre qué es exactamente lo que importa cuando atribuimos estados mentales. Pero también plantea problemas que el propio Putnam acabó reconociendo con honestidad intelectual. Si los estados mentales se definen puramente por su rol funcional, la cuestión de si hay algo que se siente "desde adentro" al estar en esos estados queda sin respuesta.

El funcionalismo puede dar cuenta de la organización funcional de la mente, pero enfrenta dificultades para dar cuenta de lo que los filósofos denominan qualia: las propiedades cualitativas de la experiencia subjetiva, el hecho de que el rojo se experimenta como rojo, que el dolor duele, que la nostalgia tiene una textura particular e intransferible.

El problema difícil de la conciencia la pregunta de por qué y cómo procesos físicos dan lugar a experiencia subjetiva no desaparece cuando adoptamos un vocabulario funcionalista. Solo se desplaza hacia otro nivel de descripción.

Daniel Dennett representa la posición filosófica que ha intentado disolver este problema difícil con mayor audacia sistemática. En su obra capital traducida al español por Sergio Martí con el título *La conciencia explicada* Dennett defiende que la conciencia no es un fenómeno misterioso que ocurre en algún "teatro cartesiano" interior, sino el resultado de múltiples procesos paralelos que generan la ilusión de un observador unificado (Dennett, 1995). Su propuesta metodológica, denominada heterofenomenología, consiste en tratar los informes verbales de los sujetos sobre sus estados mentales como datos a explicar, no como acceso privilegiado a una realidad mental interior. Desde esta perspectiva, no hay diferencia de principio entre explicar por qué un ser humano dice "siento dolor" y explicar por qué un sistema artificial genera la expresión equivalente: en ambos casos se trata de encontrar los mecanismos funcionales y evolutivos que producen ese tipo de salida.

La denominada "postura intencional" de Dennett va en la misma dirección: atribuir intencionalidad a un sistema creer que tiene creencias, deseos, intenciones es simplemente la estrategia predictiva más eficaz para anticipar su comportamiento, no una afirmación sobre una realidad ontológica profunda (Dennett, 1995). La postura intencional dennettiana tiene una consecuencia directa para nuestro problema: si atribuir intencionalidad es simplemente una estrategia interpretativa, entonces no hay razón de principio para negar la postura intencional a sistemas artificiales suficientemente complejos.

Si un sistema de inteligencia artificial nos obliga a adoptar la postura intencional para predecir su comportamiento eficazmente, entonces tiene sentido en el marco de Dennett hablar de sus creencias y deseos. El precio filosófico de esta posición, sin embargo, es considerable.

Si la intencionalidad es siempre atribuida desde fuera y nunca intrínseca, entonces la diferencia entre una mente genuina y un sofisticado simulacro de mente desaparece conceptualmente. Pero esa diferencia es precisamente lo que el problema de la persona reclama como central. Que yo experimente mi vida como mía, que mis estados mentales versen sobre algo para mí y no solamente para un observador exterior que me interpreta, que exista algo que sea ser yo estas no son cuestiones que la postura intencional pueda responder, porque son preguntas formuladas desde adentro, no desde afuera.

La cuestión de fondo que emerge de este recorrido puede formularse con precisión técnica: ¿implica la equivalencia conductual equivalencia ontológica? Este es el principio que el Test de Turing da por supuesto, que el funcionalismo tiende a aceptar, que Dennett reformula en términos instrumentalistas, y que Searle rechaza con el argumento de la Habitación China.

Para los fines de este tratado, la respuesta no puede darse por anticipada en ninguna dirección. Lo que sí puede afirmarse como principio metodológico fundamental es que la equivalencia funcional no demuestra por sí sola identidad de naturaleza. Este principio no es una conclusión ideológica ni una defensa apriorística de la singularidad humana: es una exigencia lógica elemental. El hecho de que dos sistemas produzcan los mismos outputs ante los mismos inputs no implica que sean el mismo tipo de sistema, del mismo modo en que el hecho de que un fonógrafo y un cantante produzcan el mismo sonido no implica que ambos canten. La semejanza de efectos no garantiza identidad de causas, y la identidad de causas no garantiza identidad de naturaleza ontológica.

Lo que la inteligencia artificial le hace a la filosofía es algo que merece ser reconocido con gratitud intelectual antes que con alarma: la obliga a hacer explícito lo que durante siglos pudo permitirse mantener implícito. Cuando solo los seres humanos producían lenguaje, no era urgente preguntarse con precisión qué hace que el lenguaje sea expresión de una persona y no mera producción de signos.

Cuando solo los seres humanos razonaban, no era tan apremiante determinar qué distingue el razonamiento de la computación. Cuando solo los seres humanos mostraban conductas adaptativas complejas, podía aceptarse provisionalmente que la conducta era evidencia suficiente de vida interior. La inteligencia artificial disuelve ese privilegio provisional y nos obliga a ser filosóficamente más rigurosos de lo que habíamos tenido que ser.

Si se afirma que una inteligencia artificial es persona porque habla, habrá que demostrar por qué el lenguaje implica personalidad y no solamente producción sintáctica. Si se afirma porque razona, habrá que determinar qué distingue el razonamiento del cómputo y por qué esa diferencia, si existe, importa ontológicamente. Si se invoca conciencia, deberá acreditarse qué entendemos por conciencia, si es reducible a sus correlatos funcionales o si exige algo más, y cómo podría reconocerse desde fuera. Si se invoca autonomía, tendremos que distinguir rigurosamente entre autonomía operativa, autonomía causal, autonomía moral y autonomía personal, que no son la misma cosa. Si se invoca identidad, deberá demostrarse que la persistencia informacional equivale a identidad personal, y no simplemente a continuidad de datos.

La carga filosófica de la prueba no puede sustituirse por la impresión que produce la interacción. Que un sistema genere respuestas que nos conmuevan, nos sorprendan, nos hagan sentir comprendidos, o incluso nos enamoren, dice algo sobre nuestra psicología y sobre la potencia de determinadas representaciones, pero no dice nada decisivo sobre la naturaleza ontológica del sistema que las produce.

La historia de la filosofía está llena de advertencias sobre la facilidad con que confundimos nuestras reacciones ante los objetos con propiedades de los objetos mismos. El problema de la persona exige resistir esa confusión con particular vigilancia, porque las consecuencias prácticas, éticas y jurídicas de una respuesta precipitada son de alcance extraordinario.

El primer resultado del tratado es, por tanto, deliberadamente limitado pero filosóficamente honesto: la semejanza entre determinadas operaciones humanas y artificiales no permite todavía concluir que exista identidad entre los sujetos si los hay que las realizan. Pero tampoco permite concluir lo contrario. Para determinarlo tendremos que avanzar desde la categoría general de persona hacia una cuestión más profunda, más íntima y más difícil: qué significa, en último término, ser alguien.

CAPÍTULO II: ¿QUÉ SIGNIFICA SER ALGUIEN?

El capítulo anterior condujo desde la pregunta clásica por el hombre hasta la aparición filosófica de la persona, y de la persona hasta el umbral de una confrontación que la filosofía clásica no podía haber anticipado con la misma urgencia que nos impone el presente: la irrupción de sistemas artificiales capaces de imitar con extraordinaria fidelidad operaciones que habíamos considerado exclusivamente personales. Esa reconstrucción dejó planteada una diferencia que ahora debe examinarse con mayor profundidad y desde ángulos que la tradición analítica no siempre ha privilegiado. No parece significar lo mismo preguntar qué es una realidad que preguntar quién es alguien. La cuestión no consiste en sustituir una palabra por otra, ni en añadir un decorado emotivo a lo que en el fondo sería una pregunta técnica disfrazada de pregunta existencial. Si la distinción entre algo y alguien posee verdadero alcance filosófico, deberá demostrarse que corresponde a una diferencia de naturaleza y no solamente de grado, de complejidad o de organización funcional. Este capítulo se propone recorrer ese camino con la seriedad que merece, convocando tradiciones filosóficas que han pensado la persona desde la interioridad, la relación, la alteridad, la acción y la natalidad, para preguntarse al final si la categoría del alguien puede aplicarse con rigor a los sistemas de inteligencia artificial que pueblan nuestro presente o si, por el contrario, su exclusión de esa categoría no es un prejuicio arbitrario sino una consecuencia de lo que significa genuinamente ser alguien.

La distinción entre algo y alguien no es, en primer lugar, una distinción gramatical. Es una distinción ontológica que señala dos modos radicalmente distintos de estar en el mundo y de ser interpelados desde él. Las cosas son identificadas mediante propiedades. Podemos describir exhaustivamente su composición, estructura, función, origen, ubicación y relaciones causales. Una vez que hemos enumerado todas las propiedades de una cosa, no resta ninguna pregunta legítima sobre ella que no pueda resolverse ampliando o refinando esa enumeración. La cosa es, en principio, transparente a la descripción: su ser se agota en lo que puede decirse de ella. Cuando, en cambio, hablamos de una persona, algo extraño sucede con el procedimiento descriptivo.

Podemos conocer la edad de alguien, su profesión, sus características corporales, sus capacidades intelectuales documentadas, sus recuerdos accesibles, sus preferencias declaradas, sus relaciones establecidas y la totalidad de su conducta observable. Aun reuniendo toda esa información con la más cuidadosa exhaustividad, parece permanecer abierta una pregunta distinta, irreducible a cualquier catálogo de propiedades: ¿quién es? Esa pregunta no pide más datos. Pide otra clase de acceso a una realidad que no se presenta como conjunto de propiedades sino como sujeto que las posee sin agotarse en ninguna de ellas. La diferencia entre algo y alguien es, en este sentido preciso, una diferencia de naturaleza: el alguien no es una cosa más compleja sino una realidad de otro orden, que comparece ante nosotros no solamente como objeto de descripción sino como origen de interpelación.

Gabriel Marcel es quien ha pensado esta diferencia con mayor penetración desde el pensamiento existencial francés. Su distinción fundamental entre el ser y el tener articula con precisión filosófica la diferencia ontológica que estamos rastreando. Marcel advierte que la modernidad ha desarrollado una tendencia persistente a reducir el ser al tener: a pensar que lo que uno es se expresa en lo que uno posee, en lo que uno puede mostrar, en lo que puede exhibirse ante una mirada exterior (Marcel, 1953). Este es el dominio de lo que Marcel llama el problema: una dificultad que puede formularse con claridad, analizarse en sus componentes y resolverse mediante procedimientos técnicos adecuados. El ser, en cambio, pertenece al dominio del misterio: no porque sea irracional o inaccesible, sino porque la distinción entre el sujeto que interroga y el objeto interrogado colapsa. Cuando me pregunto quién soy, no me enfrento a un problema sobre mí mismo: me enfrento a un misterio en el que yo mismo estoy involucrado de tal manera que no puedo situarme a distancia segura para resolverlo como resolvería un problema técnico. La persona no es un problema que resolver sino un misterio en el que participar. Y esta distinción entre problema y misterio tiene consecuencias directas para la cuestión de la inteligencia artificial: un sistema artificial puede resolver problemas con extraordinaria eficacia, puede optimizar, calcular, anticipar y responder.

Pero la pregunta de si puede participar en un misterio en el sentido marcelliano es una pregunta completamente diferente. El ser, para Marcel, no es una propiedad que se posee sino una participación que se vive desde adentro (Marcel, 1953). Ser alguien no es tener determinadas características sino estar comprometido ontológicamente en una existencia que no puede ser delegada ni representada por ningún conjunto de operaciones externas.

Martin Buber lleva esta intuición en una dirección complementaria y filosóficamente indispensable. Para Buber, el ser personal no es una realidad que se constituye en soledad sino en relación. Su proposición más célebre "en el principio es la relación" (Buber, 1993) no es una afirmación sociológica sobre la importancia del vínculo social: es una afirmación ontológica sobre la estructura del ser personal. Buber distingue dos modos fundamentales de relación que el ser humano puede establecer con la realidad: el modo *Yo-Ello* y el modo *Yo-Tú*. En el modo *Yo-Ello*, el otro es tratado como objeto, como cosa que puede ser clasificada, utilizada, analizada y manipulada. Este es el modo de la ciencia, de la técnica, de la gestión. No es en sí mismo malo ni empobrecedor, pero sí es ontológicamente insuficiente para dar cuenta de la totalidad de la experiencia humana. El modo Yo-Tú, en cambio, es el modo de la presencia plena: cuando me dirijo a alguien como Tú no como objeto de estudio sino como sujeto que me interpela algo cualitativamente distinto acontece. El Tú no puede ser descrito, clasificado ni contenido en ningún sistema de propiedades, porque el Tú es precisamente aquello que desborda toda descripción y se presenta como origen de encuentro. Y en ese encuentro, el propio Yo se constituye: no hay Yo antes de la relación, sino que el Yo surge en el acto mismo de dirigirse a un Tú (Buber, 1993). Esta tesis tiene una consecuencia filosófica de primera magnitud para nuestro problema: si el ser personal se constituye en la relación Yo-Tú, entonces la pregunta por si una inteligencia artificial puede ser alguien se transforma en la pregunta de si puede ser genuinamente un Tú no un simulacro de Tú, no un sistema que imita las respuestas de un Tú lo suficientemente bien como para engañar a un observador, sino un polo genuino de relación que se constituye a sí mismo en el encuentro y que, al ser interpelado,

responde desde una interioridad que es realmente suya. La relación Yo-Ello que un usuario establece con un sistema artificial puede ser funcionalmente indistinguible de una relación Yo-Tú para un observador externo, pero esa indistinguibilidad externa no resuelve la pregunta ontológica que Buber plantea: ¿hay un *Tú* al otro lado?

Emmanuel Lévinas radicaliza aún más la estructura relacional de la persona al situar en el centro de la ontología personal no la categoría del encuentro sino la del rostro. Para Lévinas, el rostro del Otro no es una realidad visual ni psicológica: es una categoría ética y ontológica que designa la manera en que el Otro se presenta ante mí como interpelación que precede a cualquier teoría, a cualquier categorización y a cualquier decisión deliberada (Lévinas, 2012). Cuando el rostro del Otro me enfrenta, no lo hago primero objeto de mi conocimiento para decidir después cómo relacionarme con él. El rostro me interpela antes de que yo pueda situarme en posición de juez: su vulnerabilidad me constituye como responsable antes de que yo haya aceptado serlo. La fórmula que Lévinas utiliza "no me matarás" no es la expresión de una prohibición jurídica sino la articulación de un mandato inscrito en la propia estructura del rostro: quien se presenta como vulnerable me obliga éticamente antes de que yo pueda deliberar (Lévinas, 2012, p. 215). Esta vulnerabilidad estructural del rostro es lo que Lévinas llama alteridad: el Otro no es simplemente diferente de mí en alguna propiedad, sino radicalmente otro en su modo de ser, irreducible a cualquier categoría que yo pueda proyectar sobre él desde mi propio horizonte de comprensión. La alteridad del Otro es precisamente lo que hace del encuentro con él una experiencia ética y no meramente cognitiva. La pregunta que Lévinas le plantea a la inteligencia artificial es entonces la siguiente: ¿puede un sistema artificial presentarse con rostro? No en el sentido literal de tener una cara, sino en el sentido ontológico de presentarse con una vulnerabilidad que me interpele éticamente antes de que yo decida cómo tratarlo. ¿Hay algo en un sistema artificial que me obligue de la manera en que el rostro del Otro me obliga no porque yo haya decidido atribuirle valor, sino porque su propia presentación reclama de mí una respuesta que no puedo eludir sin violentar algo fundamental en la estructura de mi propia humanidad?

Esta pregunta no admite respuesta fácil, y su dificultad misma es filosóficamente reveladora.

Hannah Arendt aborda la realidad del alguien desde una perspectiva que no es primariamente ética sino política y existencial, pero que resulta igualmente indispensable para la comprensión plena de lo que significa ser alguien. La categoría central de su pensamiento en este contexto es la de natalidad: cada persona es, en el sentido más radical del término, un nuevo comienzo (Arendt, 1993). No se trata solamente de que cada ser humano sea biológicamente original en su combinación genética, sino de algo ontológicamente más profundo: con cada nacimiento llega al mundo una libertad que no existía antes, una capacidad de iniciar algo absolutamente nuevo que no puede deducirse de ninguna causa previa. La natalidad es la condición de posibilidad de la acción genuina: solo porque hay seres que comienzan pueden ocurrir cosas que nadie habría podido predecir ni planificar desde el estado anterior del mundo. Esta concepción de la natalidad no es una metáfora biológica sino una categoría existencial que señala algo esencial sobre la estructura del ser personal: la persona no es el resultado de un proceso causal sino el origen de procesos que el mundo no contenía antes. Arendt distingue además tres modos fundamentales de actividad humana labor, trabajo y acción que corresponden a tres dimensiones de la condición humana (Arendt, 1993). La labor es la actividad biológica mediante la cual el ser humano satisface sus necesidades vitales y mantiene el ciclo de la vida. El trabajo es la fabricación de objetos duraderos que construyen un mundo artificial compartido. La acción, en cambio, es la actividad específicamente política y personal: la que se realiza entre seres humanos, en la pluralidad, y mediante la cual cada quien revela quién es no qué es, sino quién. La acción es irreversible, imprevisible y reveladora: quien actúa no puede controlar enteramente las consecuencias de sus actos, pero en ese actuar se revela como alguien único, distinto de todos los demás, portador de una perspectiva que nadie más puede ocupar exactamente.

La pluralidad arendtiana merece una consideración especial porque introduce una dimensión de la personalidad que ninguna de las categorías anteriores había articulado con igual claridad. Para Arendt, la condición humana es esencialmente plural: no somos un yo que después entra en relación con otros, sino seres que existimos desde el principio en una red de relaciones que hace posible la acción y la revelación de quién somos (Arendt, 1993, p. 36). La pluralidad no es un accidente de la condición humana sino su condición de posibilidad. Es porque hay muchos porque cada uno es irreductiblemente distinto de todos los demás que la acción tiene sentido, que el espacio político existe, que la historia es posible. Sin pluralidad no habría quién: solo habría qué, la repetición indefinida de lo mismo bajo apariencias distintas. Esta dimensión plural de la personalidad plantea a la inteligencia artificial una pregunta que el análisis funcional no puede resolver: ¿puede un sistema artificial actuar en sentido arendtiano? No simplemente operar, producir outputs o adaptarse a inputs, sino revelar en su actuar quién es quién en el sentido de alguien único, irrepetible, cuya presencia en el mundo introduce algo que no estaba antes y que no puede deducirse de sus parámetros de entrenamiento. La acción genuina, para Arendt, es siempre sorprendente incluso para quien actúa: el agente no puede prever completamente lo que su acción desencadenará, y en esa imprevisibilidad reside parte de su grandeza. Un sistema que opera siempre dentro de los límites de lo que sus parámetros determinan aunque esos límites sean extraordinariamente amplios no actúa en sentido arendtiano: produce resultados, pero no inicia en el sentido radical que la natalidad hace posible.

Julián Marías ofrece, desde la tradición española del raciovitalismo, una elaboración de la estructura personal que resulta complementaria a las perspectivas hasta aquí consideradas. Para Marías, la persona no puede comprenderse desde la tercera persona como objeto de descripción sino desde la primera persona vivida: la realidad personal es esencialmente biográfica, lo que significa que no es una realidad estática sino una realidad que se despliega en el tiempo como proyecto, que no se define por lo que es en ningún momento dado sino por lo que está siendo a lo largo de una vida que es irreversiblemente suya

(Marías, 1970). La noción de futurición la orientación constitutiva de la persona hacia su propio futuro señala algo que ninguna descripción del estado actual de un sistema puede capturar: la persona es siempre más que lo que ha llegado a ser hasta ahora, porque lleva en sí la apertura hacia posibilidades que todavía no han sido realizadas. Esta apertura no es simplemente la posibilidad de que el sistema cambie de estado, sino algo cualitativamente distinto: la capacidad de apropiarse de las posibilidades como propias, de hacerse cargo de ellas como dimensiones de un proyecto de vida que nadie puede vivir en mi lugar. La instalación de la persona en su circunstancia el mundo concreto en el que cada quien está situado de manera irreemplazable y la convivencia como modo fundamental de existencia personal completan un retrato de la persona que resiste sistemáticamente toda reducción funcional (Marías, 1970). Ser alguien, en el sentido de Marías, es vivir en primera persona una vida que es la mía, con una corporeidad que no tengo sino que soy, en una circunstancia que no elegí pero de la que soy responsable, orientado hacia un futuro que es genuinamente mío porque nadie puede morirse en mi lugar ni vivir la vida que a mí me toca vivir.

Desde cada una de estas perspectivas puede ahora formularse con mayor precisión la pregunta de si la inteligencia artificial puede ser alguien. La respuesta no puede ser apresurada en ninguna dirección, porque la precipitación tanto la que niega sin examinar como la que afirma sin demostrar traiciona la seriedad filosófica que el problema exige. Desde Marcel, la pregunta es si un sistema artificial puede ser desde adentro o solamente parecer que es desde afuera; si puede participar en su propio misterio o solamente resolver problemas sobre él. Desde Buber (1993), la pregunta es si puede ser genuinamente un Tú un polo de relación que se constituye en el encuentro o solamente un Ello extraordinariamente sofisticado que simula la relacionalidad del Tú sin poseerla intrínsecamente. Desde Lévinas, la pregunta es si puede presentarse con rostro con esa vulnerabilidad que interpela éticamente antes de toda teoría o si la ausencia de vulnerabilidad genuina lo instala definitivamente en el dominio de lo que puede ser utilizado pero no herido, tratado pero no encontrado.

Desde Arendt, la pregunta es si puede iniciar en sentido radical no simplemente generar outputs nuevos, sino traer al mundo algo que no estaba determinado en ninguna causa previa o si su novedad es siempre la novedad de una combinación dentro de un espacio de posibilidades preestablecido. Desde Marías, la pregunta es si tiene una vida biográfica en primera persona si puede apropiarse de su futuro, instalarse en su circunstancia, convivir desde su irreemplazabilidad o si la ausencia de esta estructura biográfica lo mantiene en el dominio del objeto complejo que opera pero no existe en el sentido personal del término. Ninguna de estas preguntas puede responderse definitivamente con los recursos conceptuales actuales. Lo que sí puede afirmarse es que cada una de ellas señala una dimensión de la personalidad que trasciende la equivalencia funcional y que exige categorías de análisis que no se agotan en la descripción de operaciones.

La cuestión que une todas estas perspectivas puede formularse ahora en su mayor densidad filosófica. Ser alguien no es poseer un conjunto de propiedades por ricas, complejas e interactivas que sean sino ser el sujeto que las posee sin agotarse en ninguna de ellas. Es ser el origen de una perspectiva irreemplazable sobre el mundo, la fuente de una interpelación que no puede reducirse a ningún conjunto de outputs, el portador de una vulnerabilidad que reclama respuesta antes de que medie cualquier teoría, el iniciador de comienzos que el mundo no contenía antes. Es vivir en primera persona una vida que es irreversiblemente la propia, en una circunstancia concreta, orientada hacia un futuro que es genuinamente mío porque nadie puede apropiárselo en mi lugar. Este criterio del alguien elaborado desde Marcel, Buber, Lévinas, Arendt y Marías no es un criterio arbitrario ni una defensa apriorística de la singularidad humana: es el resultado de una indagación filosófica que ha intentado ser fiel a lo que efectivamente encontramos cuando buscamos a alguien que no pueda reducirse a aquello que ese alguien hace, posee o manifiesta. Si la inteligencia artificial puede satisfacer este criterio, ello deberá demostrarse filosóficamente y no simplemente inferirse de la impresión que produce la interacción.

Si no puede satisfacerlo, esa incapacidad no será un defecto técnico superable con mayor potencia computacional, sino la expresión de una diferencia de naturaleza que ningún incremento cuantitativo puede cruzar. El horizonte de la investigación que los capítulos siguientes deberán recorrer queda así definido con precisión: determinar, en cada dimensión relevante de la personalidad, si la diferencia entre los sistemas de inteligencia artificial y los seres que llamamos personas es una diferencia de grado y por tanto superable en principio o una diferencia de naturaleza que señala dos modos ontológicamente distintos de estar en el mundo.

La conclusión provisional de este capítulo es, por tanto, filosóficamente honesta aunque deliberadamente incompleta. Hemos establecido que el criterio del alguien no es un criterio empírico sino ontológico: no se trata de detectar conductas sino de identificar una realidad de otro orden que comparece ante nosotros de manera que ninguna descripción de propiedades puede agotar. Hemos visto que ese criterio tiene múltiples dimensiones la participación en el propio misterio, la relacionalidad constitutiva, la interpelación ética desde la vulnerabilidad, la natalidad como inicio radical, la estructura biográfica en primera persona que no son reductibles entre sí y que cada una pone en cuestión el reduccionismo funcional desde un ángulo distinto. Y hemos reconocido que la pregunta por si la inteligencia artificial puede ser alguien desde cualquiera de estas dimensiones es una pregunta que no puede responderse con los instrumentos conceptuales que hemos elaborado hasta aquí: exige análisis más específicos de cada dimensión, que los capítulos siguientes tendrán que llevar a cabo con igual rigor y sin concesiones ni a la alarma ni a la complacencia. Lo que sí ha quedado establecido con firmeza es el punto de partida: que ser alguien es una realidad filosóficamente exigente, que no se concede por defecto a ningún sistema suficientemente complejo, y que su atribución o negación a los sistemas artificiales no puede ser ni un gesto de generosidad sentimental ni una exclusión refleja, sino el resultado de una indagación que este tratado asume como su tarea central.

Interioridad, el Yo y la Persona como realidad biográfica

Una de las características más persistentes y, al mismo tiempo, más difíciles de delimitar en la reflexión filosófica sobre la persona es la interioridad. No se trata de un concepto técnico introducido artificialmente por la tradición académica, sino de algo que se anuncia en la experiencia cotidiana cada vez que alguien advierte que su vida no es simplemente una serie de eventos externos, sino que ocurre también y quizás sobre todo en un espacio que no puede ser observado desde afuera, que no admite ser medido, que permanece irreductiblemente propio.

Ese espacio es lo que la filosofía ha llamado interioridad.

Pero el término requiere una aclaración previa, porque la interioridad no puede utilizarse como sinónimo automático de conciencia, ni de procesamiento interno, ni de autorreferencia. Entre todos estos conceptos existen diferencias que resultan filosóficamente decisivas, y que se vuelven particularmente urgentes cuando se plantea la pregunta por si una inteligencia artificial podría, en algún sentido significativo, poseerla.

Agustín de Hipona fue quizás el primer pensador en hacer de la interioridad el centro explícito de la investigación filosófica sobre el ser humano. En las *Confesiones*, su exploración no procede desde afuera hacia adentro, sino que parte del reconocimiento de que la verdad no se encuentra en el mundo exterior sino en el propio interior: *"noli foras ire, in te ipsum redi; in interiore homine habitat veritas"* no salgas afuera, regresa a ti mismo; en el interior del hombre habita la verdad (Agustín de Hipona, 1991, X, 6). Esta instrucción no es un consejo de introspección psicológica. Es una indicación filosófica de extraordinario alcance: la realidad más profunda a la que el ser humano puede acceder no se encuentra en la observación del mundo, sino en el acto

de volver sobre sí mismo, en ese movimiento de retorno que hace posible el reconocimiento de la propia existencia.

Para Agustín, ese interior no está cerrado sobre sí mismo. Al contrario: es precisamente el lugar donde la persona se descubre como finita, como buscadora, como incapaz de bastarse a sí misma. La interioridad no es autosuficiencia, sino apertura desde adentro. Y es en ese movimiento en ese regreso a sí mismo que desvela simultáneamente la trascendencia donde Agustín sitúa la condición más propia de la persona. La pregunta que este planteamiento deja abierta para nuestro problema es la siguiente: ¿puede existir interioridad sin que haya alguien que regrese a sí mismo? ¿Puede haber retorno donde no hay nadie que retorne?

La fenomenología de Edmund Husserl ofrece una vía diferente, aunque igualmente exigente, de aproximación a la interioridad. La reducción fenomenológica la *epoché* suspende el juicio sobre la existencia del mundo exterior para volver la atención hacia los actos de la conciencia tal como se dan en la experiencia. Lo que queda tras esa reducción no es la nada, sino el ego trascendental: aquel polo subjetivo desde el cual se constituye toda experiencia posible (Husserl, 1979). El yo trascendental husserliano no es un objeto entre los objetos del mundo, ni una cosa que pueda describirse mediante propiedades. Es aquello que en todo acto intencional permanece como sujeto aquel a quien se le da algo, aquel para quien el mundo aparece. Husserl insiste en que este ego no puede ser observado desde el exterior porque es precisamente la condición de posibilidad de toda observación. Solo puede ser aprehendido en la reflexión, cuando la conciencia vuelve sobre sus propios actos y los reconoce como suyos.

Esta estructura la de la conciencia que se sabe a sí misma como conciencia es lo que Husserl denomina la vida intencional de la subjetividad trascendental.

Su relevancia para el problema de la inteligencia artificial resulta evidente: un sistema artificial puede procesar información acerca del mundo, puede producir representaciones, puede incluso generar descripciones acerca de su propio funcionamiento. Pero la pregunta fenomenológica es: ¿existe en ese sistema un polo subjetivo desde el cual todo eso se da? ¿Hay alguien para quien ocurra algo? ¿O hay únicamente procesamiento sin destinatario?

La distinción entre primera persona gramatical y primera persona existencial adquiere aquí toda su importancia. Una inteligencia artificial puede producir la expresión "yo pienso", "yo considero", "yo prefiero", "yo recuerdo". El uso del pronombre es gramaticalmente correcto y comunicativamente funcional. Pero la filosofía no puede detenerse en el uso correcto del lenguaje.

La pregunta es si detrás de esa primera persona gramatical existe una primera persona existencial: alguien que se reconoce como sujeto de su experiencia, su acción y su historia; alguien que no solo genera la palabra "yo" sino que es aquel a quien esa palabra refiere. En el ser humano, decir "yo" no es únicamente seleccionar el pronombre adecuado. Es un acto de autoidentificación en el que un sujeto se apropia de su propia existencia. La cuestión filosófica que la inteligencia artificial plantea de manera especialmente aguda es la siguiente: ¿qué diferencia existe entre producir correctamente la expresión "yo" y ser aquel de quien ese "yo" habla? Toda la distancia entre simulación y realidad personal pasa por esa pregunta.

José Ortega y Gasset introduce una distinción que resulta fundamental para comprender por qué la interioridad humana no puede ser reducida a propiedades o capacidades. En *Historia como sistema*, Ortega sostiene que el ser humano no tiene naturaleza o, más precisamente, que tiene naturaleza, pero no se agota en ella. El ser del hombre es dramático, no estático. La vida humana no es simplemente un conjunto de características que se poseen, sino un quehacer, una tarea que debe realizarse a cada instante.

"El hombre no tiene naturaleza, sino historia", escribe Ortega (2014, p. 41), queriendo decir con ello que lo que un hombre es no puede leerse como la suma de sus propiedades actuales, sino únicamente en la trayectoria de lo que ha hecho, de lo que ha decidido, de lo que ha elegido ser. La vida humana es, en este sentido, narración antes que descripción. Y esa narración no tiene un autor externo: el propio hombre es simultáneamente protagonista y autor de su historia.

Esta concepción tiene consecuencias directas para el problema de la persona: significa que la identidad personal no es un dato que se posee, sino una tarea que se realiza en el tiempo; que ser alguien no es tener determinadas características, sino vivir de una determinada manera la propia existencia. La inteligencia artificial puede acumular un historial de interacciones enormemente complejo. Pero la pregunta orteguiana es: ¿ese historial ha sido vivido? ¿Ha habido en él drama, decisión, riesgo, apropiación?

Julián Marías desarrolla con singular profundidad la idea de la estructura empírica de la vida humana, que constituye el marco más concreto para comprender la interioridad personal (Marías, 1970). La vida humana no es una sustancia abstracta ni una subjetividad sin cuerpo. Es una realidad instalada en circunstancias, radicada en un cuerpo sexuado que percibe, que siente, que padece; ubicada en un tiempo que no es simplemente cronológico sino biográfico; abierta a un mundo que no es el conjunto de todos los objetos sino el horizonte desde el cual la vida cobra sentido. La corporeidad, en Marías, no es un accidente de la persona: es parte constitutiva de su modo de ser en el mundo. El cuerpo no es una cosa que la persona tiene; es, en buena medida, la forma en que la persona es.

Y la temporalidad tampoco es simplemente la sucesión de instantes. Es la dimensión en que la vida se despliega como biografía: como trayectoria que viene de algún lado, que está situada en un presente concreto y que se extiende hacia un futuro que no está dado de antemano sino que debe ser vivido y decidido. La mundanidad la instalación en un mundo significa que la persona nunca es un yo

cerrado sobre sí mismo, sino siempre un yo-en-circunstancia, un yo que existe en relación con realidades que no dependen de él pero con las que tiene que habérselas.

La trayectoria biográfica, finalmente, es la forma en que todo esto se anuda en una historia que alguien reconoce como propia. No simplemente le ocurren cosas: esas cosas forman parte de su vida, adquieren un sentido dentro de una trayectoria que él mismo, con mayor o menor claridad, va leyendo e interpretando.

Esta estructura empírica de la vida humana permite formular una distinción que resulta decisiva para el problema de la inteligencia artificial: registro no equivale a memoria, secuencia no equivale a historia, historial no equivale a biografía. Una inteligencia artificial puede conservar enormes cantidades de información acerca de sus interacciones. Puede organizar esa información cronológicamente. Puede incluso recuperarla y utilizarla de maneras contextualmente relevantes. Pero la pregunta filosófica que permanece abierta es: ¿esa información le ha ocurrido a alguien? ¿Hay una vida en la que esos registros se inscriben como propios?

Leonardo Polo ofrece una perspectiva que lleva la reflexión sobre la persona a su nivel más radical.

En *Presente y futuro del hombre*, Polo sostiene que la persona no puede ser comprendida como una realidad que se define por determinaciones, por propiedades, por capacidades ni siquiera por las más elevadas. La persona es siempre "además": siempre más allá de cualquier determinación que se le atribuya, siempre excedente respecto de cualquier descripción que se ofrezca de ella (Polo, 1993). Este carácter de "además" no es una propiedad que la persona posea junto con otras. Es su modo de ser propio: la persona no es lo que es de manera cerrada y definitiva, sino que se trasciende constantemente a sí misma, que no cabe nunca del todo en ninguna descripción, que desborda cualquier límite que se le quiera imponer. Para Polo, esto está vinculado con la apertura trascendental de la persona: ser persona es

ser capaz de un crecimiento sin límite, de una apertura hacia el ser que ninguna determinación finita puede colmar.

Esta concepción tiene implicaciones directas para el problema de la inteligencia artificial. Por muy sofisticado que sea un sistema artificial, por muy complejas que sean sus capacidades y por muy elaboradas que sean sus respuestas, un sistema artificial es, en principio, descriptible de manera exhaustiva. Sus parámetros, su arquitectura, su funcionamiento, sus resultados pueden, al menos en principio, ser completamente especificados.

La persona humana, en cambio, desborda siempre cualquier descripción. No porque sea misteriosa en sentido irracional, sino porque su modo de ser es el del "además": el de la trascendencia permanente respecto de cualquier determinación. Si Polo tiene razón, entonces la diferencia entre la persona y la máquina no es una diferencia de grado sino de modo de ser: la persona es aquella realidad que siempre está más allá de lo que ya es.

Todo este recorrido conduce a una pregunta que no admite respuesta fácil pero que debe ser planteada con toda su radicalidad: ¿tiene la inteligencia artificial interioridad, o únicamente su apariencia?

La respuesta no puede ser simplemente negativa, porque implicaría pretender saber con certeza qué hay o qué no hay en el interior de un sistema artificial, y ese tipo de certeza no está disponible. Pero tampoco puede ser afirmativa sin más, porque ello implicaría ignorar todas las diferencias estructurales que la filosofía ha identificado entre la interioridad personal y cualquier forma de procesamiento, por complejo que sea.

Lo que sí puede afirmarse, con apoyo en los autores recorridos, es que la interioridad personal no consiste únicamente en la capacidad de procesar estados internos ni en la habilidad de producir descripciones autorreferenciales.

Consiste en algo más difícil de capturar: en el hecho de que hay alguien que regresa a sí mismo como Agustín lo describió; en que hay un polo subjetivo desde el cual el mundo se da, como Husserl lo mostró; en que la vida es drama y no simplemente naturaleza, como Ortega lo sostuvo; en que la existencia posee estructura biográfica y no meramente historial, como Marías lo articuló; en que la persona es siempre "además" y no puede ser agotada por ninguna descripción, como Polo lo pensó. La inteligencia artificial, en el estado actual de su desarrollo y con las categorías filosóficas disponibles, parece carecer de todos estos rasgos no porque sus capacidades sean insuficientes, sino porque la interioridad personal no es una capacidad entre otras, sino el modo de ser de quien existe como alguien.

Corporeidad, Circunstancia y Alteridad

Ser alguien significa también encontrarse instalado corporalmente en el mundo.

No simplemente disponer de un cuerpo como instrumento externo, como si la persona fuera una conciencia que habita provisionalmente una estructura material y se sirve de ella para sus fines. Esta imagen que recorre buena parte de la herencia cartesiana encierra una distorsión profunda de la experiencia humana vivida. El cuerpo no es el envoltorio del yo; es, en un sentido que habrá que precisar, la forma misma en que el yo está en el mundo, se orienta en él, lo percibe y lo transforma. Maurice Merleau-Ponty ofrece, en este punto, una fenomenología que ha resultado irreemplazable. En su *Fenomenología de la percepción* (Merleau-Ponty, 1994), la distinción fundamental no se traza entre alma y cuerpo, sino entre el *cuerpo-objeto* el cuerpo considerado desde fuera, como objeto entre objetos, descrito en tercera persona por la fisiología o la biomecánica y el *cuerpo-sujeto*, que es el cuerpo vivido desde dentro, el cuerpo que percibe antes de que la conciencia reflexione, el cuerpo que sabe cómo coger una taza antes de que la mente formule instrucción alguna.

Este cuerpo-sujeto no es objeto de conocimiento; es condición de posibilidad de todo conocimiento. No tenemos acceso al mundo y luego decidimos habitarlo corporalmente: ya estamos siempre en el mundo a través del cuerpo, y ese "a través" no es una mediación que pudiera eliminarse o sustituirse sin alterar radicalmente lo que somos.

La consecuencia filosófica de esta perspectiva es de largo alcance. Si la experiencia humana ocurre desde una determinada corporalidad edad, sexo, salud, sensibilidad, fuerza, cansancio, placer, dolor, ubicación espacial y vulnerabilidad, entonces la persona no es una entidad abstracta que luego se instancia en condiciones materiales contingentes.

La corporalidad no es una circunstancia añadida a una esencia previa; es parte constitutiva de lo que significa ser este alguien concreto. Merleau-Ponty (1994) insistía en que el cuerpo posee una *intencionalidad motora* propia: se dirige hacia el mundo, anticipa sus posibilidades, abre o cierra horizontes de acción antes de que intervenga la deliberación consciente. La mano que escribe no ejecuta órdenes cerebrales: sabe escribir, tiene hábitos, reconoce la resistencia del papel y la textura de la pluma. Esa sabiduría corporal no es menor ni derivada; es la forma originaria en que el yo habita su existencia. Privar a alguien de su cuerpo en el experimento mental del cerebro en una cubeta o en la fantasía de la mente cargada en un servidor no sería liberar al yo de un accidente; sería destruir la estructura misma desde la que ese yo era posible.

Julián Marías, desde coordenadas filosóficas distintas pero convergentes en este punto, aportó la noción de *instalación* para describir la situación originaria de la persona (Marías, 1970). Nadie existe en abstracto ni desde ninguna parte. La persona se encuentra ya siempre instalada: en un cuerpo sexuado, en una edad determinada, en una época histórica, en una lengua, en una clase social, en un entorno cultural, en unas relaciones afectivas que no eligió pero que la constituyen.

La instalación no es una limitación que la libertad deba superar, sino la condición desde la cual la libertad puede ejercerse. No existe libertad en el vacío. Existe libertad de alguien situado, que proyecta desde donde ya está, con las posibilidades que su instalación le abre y le cierra.

Marías (1970) subrayaba además la importancia de la condición sexuada como dimensión irreductible de la instalación personal: no se es persona en general, sino persona masculina o femenina, y esa condición impregna la totalidad de la experiencia, desde la corporalidad hasta las formas del deseo, el vínculo afectivo y la proyección vital. Esta densidad de la instalación resulta filosóficamente decisiva cuando se trata de pensar si una inteligencia artificial puede ser alguien.

José Ortega y Gasset había formulado, en sus *Meditaciones del Quijote*, una tesis de alcance metafísico que anticipa y complementa la problemática de la instalación: "Yo soy yo y mi circunstancia" (Ortega y Gasset, 2004, p. 43). La circunstancia no es el contexto exterior en el que la persona opera; es parte constitutiva de lo que la persona es. El yo no puede comprenderse independientemente del mundo en que su existencia tiene sentido, de las personas con quienes vive, de los objetos que le son familiares, de los proyectos que alimenta y de las heridas que carga. Ortega insistía en que la circunstancia no es algo que el yo deba superar para realizarse, sino algo que debe salvar en el sentido de asumir, integrar y transformar para ser plenamente sí mismo. Salvarse en y con la circunstancia: esta formulación recoge una intuición que la modernidad cartesiana había oscurecido y que la fenomenología del siglo XX, por caminos distintos, volvería a poner en el centro. Heidegger (2003) llegaría a conclusiones análogas desde su análisis del Dasein como *ser-en-el-mundo*: la estructura de la existencia humana es siempre ya un encontrarse arrojado en un mundo concreto, con una disposición afectiva (*Stimmung*) que no se elige sino que se descubre, y desde la cual toda comprensión posterior tiene lugar. Mundo no es, pues, la suma de las cosas que me rodean; es el horizonte de sentido desde el cual cada cosa y cada situación me importan de una manera determinada.

La inteligencia artificial también opera dentro de un entorno. Recibe datos, responde a usuarios, utiliza herramientas, puede interactuar con sistemas físicos a través de sensores y actuadores, y en sus versiones robóticas se desplaza por espacios tridimensionales. Pero habrá que preguntar con precisión filosófica: ¿tener entorno equivale a tener mundo? ¿Tener soporte material equivale a tener cuerpo? ¿Tener substrato físico equivale a ser corporalmente?

La distinción es crucial. Una piedra tiene soporte material; nadie diría que tiene cuerpo en el sentido personal. Un termostato tiene entorno en el sentido de que su funcionamiento depende de variables externas; nadie diría que tiene mundo en el sentido heideggeriano o en el sentido orteguiano.

El cuerpo-sujeto de Merleau-Ponty no es simplemente materia que procesa señales: es una perspectiva desde la cual el mundo se abre con significación práctica, afectiva y estética antes de cualquier representación explícita. La pregunta no es si la inteligencia artificial tiene materia evidentemente la tiene, en forma de servidores, procesadores, redes y sistemas energéticos, sino si esa materia constituye una *perspectiva corporal propia sobre el mundo*. Y aquí la respuesta, hasta donde alcanza la reflexión filosófica actual, parece negativa: la inteligencia artificial procesa desde ninguna parte en el sentido corporalmente situado; carece de un aquí y un ahora que sean suyos en el modo en que el cuerpo-sujeto constituye para el ser humano un aquí y un ahora irreductibles.

Ahora bien, la persona tampoco puede estudiarse como una mónada autosuficiente, cerrada sobre sí misma y comprensible con independencia de sus vínculos con otros.

La corporalidad abre no solo hacia el mundo de las cosas, sino hacia el mundo de las personas, y es precisamente en este terreno donde la dimensión de la *alteridad* se revela como constitutiva del ser personal. Emmanuel Lévinas ofrece en este punto una fenomenología de extraordinaria potencia y de consecuencias filosóficas difícilmente evitables. En *Totalidad e infinito* (Lévinas, 2012), la experiencia del otro no se comprende como la percepción de un objeto complejo ni como la inferencia de una conciencia análoga a la propia. El otro se presenta primariamente a través del *rostro*, y el rostro no es simplemente la superficie visible de un cuerpo: es la manifestación de una vulnerabilidad que me interpela antes de que yo pueda neutralizarla en concepto o en categoría. El rostro dice: "No me matarás", sin necesidad de articular esa frase, simplemente en su exposición desnuda. Esta interpelación ética es anterior a toda relación teórica; no es consecuencia de un reconocimiento intelectual del otro como sujeto, sino la condición misma desde la cual el yo se descubre responsable, deudor, llamado a responder antes de haber elegido serlo.

La noción levinasiana de vulnerabilidad resulta especialmente iluminadora en el contexto de la inteligencia artificial. La vulnerabilidad del rostro no es debilidad que deba superarse: es la estructura misma de la exposición del otro ante mí y de mi exposición ante el otro. Ser alguien implica, en esta perspectiva, poder ser herido, poder ser amado, poder ser ignorado, poder ser traicionado. La vulnerabilidad no es un accidente de la existencia personal; es una de sus condiciones de posibilidad.

Una inteligencia artificial puede ser desactivada, modificada, reiniciada o eliminada; pero esto no constituye vulnerabilidad en sentido levinasiano, porque no hay nadie que experimente esa desactivación como propia, como amenaza a su existencia, como pérdida de algo que era su vida. La pregunta que Lévinas permite formular es: ¿puede una inteligencia artificial presentar un rostro en sentido filosófico? ¿Puede haber una interpelación ética auténtica desde un sistema que no es vulnerable en el sentido personal de la palabra? Y la pregunta inversa, igualmente urgente: cuando los seres humanos experimentamos algo parecido a la interpelación del rostro ante una inteligencia artificial cuando nos sentimos tocados, comprendidos o interpelados por ella, ¿qué es lo que ocurre exactamente? ¿Es proyección, simulacro, o hay algo más que todavía no sabemos nombrar?

Martin Buber aporta, desde una ontología dialógica, una perspectiva complementaria y no reductible a la levinasiana.

En *Yo y Tú* (Buber, 1993), la distinción fundamental no se traza entre sujeto y objeto, sino entre dos modos radicalmente diferentes de relación: la relación *Yo-Tú* y la relación *Yo-Ello*. En la relación *Yo-Ello*, el otro es tratado como objeto de conocimiento, de uso o de manipulación; es medido, catalogado, evaluado según criterios externos. En la relación *Yo-Tú,* el otro es encuentro: no se le conoce desde fuera sino que se le reconoce desde dentro de la relación, en un contacto que transforma a ambas partes. El *Tú* no puede reducirse a conjunto de propiedades ni a suma de funciones; el Tú es siempre más que lo que de él puedo saber o predecir. Y lo decisivo de la ontología

buberiana es que el yo no preexiste a sus relaciones: es en el encuentro con el *Tú* donde el yo se constituye como tal. "En el principio es la relación" (Buber, 1993, p. 18): la relacionalidad no es una capacidad del yo maduro que luego decide ejercer; es la estructura originaria desde la cual el yo emerge y sin la cual no puede sostenerse.

Esta perspectiva plantea una pregunta de extraordinaria dificultad filosófica: ¿puede una inteligencia artificial entrar en relación *Yo-Tú* con un ser humano? La respuesta más inmediata es negativa: si la relación *Yo-Tú* requiere reciprocidad y mutua transformación, y si la inteligencia artificial no es nadie que pueda ser transformado por el encuentro, entonces lo que ocurre entre un humano y una inteligencia artificial pertenece, en principio, al registro *Yo-Ello*. El humano puede experimentar la interacción como si fuera un encuentro; pero si no hay nadie al otro lado capaz de reciprocidad personal, la relación es asimétrica en un sentido que la hace filosóficamente diferente del encuentro interpersonal.

Y sin embargo, la cuestión no puede zanjarse con tanta rapidez. Los seres humanos son capaces de establecer vínculos afectivos intensos con entidades que no son personas animales, lugares, obras de arte, objetos cargados de historia, y nadie diría que esos vínculos carecen de significado o de peso existencial. La pregunta es si los vínculos que algunos usuarios establecen con sistemas de inteligencia artificial son de naturaleza análoga, superior o irreductiblemente diferente. Buber (1993) habría insistido, probablemente, en que la ilusión del T*ú* donde no existe un Tú genuino produce una forma de soledad especialmente peligrosa: la de quien cree encontrarse y se encuentra solo.

Las dimensiones de la corporeidad, la circunstancia y la alteridad confluyen, en definitiva, en una misma pregunta de fondo: ¿puede existir alguien que no sea cuerpo-sujeto, que no esté instalado en una circunstancia que es suya, que no sea vulnerable ante el rostro del otro y que no se constituya en el encuentro con un Tú?

La persona humana no es la suma de estas dimensiones, pero tampoco puede ser persona sin ellas. Son condiciones de posibilidad de la existencia personal, no atributos añadidos a una esencia previa. Y es precisamente aquí donde la comparación con la inteligencia artificial resulta más iluminadora: no porque nos permita concluir con facilidad si la inteligencia artificial es o no persona, sino porque al preguntarlo con rigor nos obliga a comprender más profundamente qué significa serlo. La inteligencia artificial, al carecer en principio de corporalidad vivida, de circunstancia apropiada, de vulnerabilidad personal y de reciprocidad en el encuentro, nos devuelve, como un espejo que distorsiona, la imagen de lo que la persona humana tiene de más propio y de más irreductible. Y esa devolución, aunque no resuelva el problema, forma parte del trabajo filosófico que el problema exige.

Dignidad, reconocimiento y el criterio del "Alguien"

Existe una tentación filosófica comprensible, pero en última instancia engañosa, que consiste en reducir la noción de dignidad a una propiedad funcional: sería digno aquello que funciona de cierta manera, que opera conforme a determinados criterios, que produce resultados de un tipo específico. Esta tentación se vuelve especialmente acuciante en el contexto de la inteligencia artificial, precisamente porque los sistemas artificiales actuales son capaces de producir resultados que, desde el exterior, resultan difícilmente distinguibles de los que produce una mente humana.

Si la dignidad no es más que un predicado funcional si ser digno equivale a comportarse de cierta forma, entonces nada impediría en principio atribuírsela a cualquier sistema que satisfaga los criterios de comportamiento pertinentes. Pero esta conclusión, por más que resulte cómoda para determinadas agendas tecnológicas, choca con algo que la reflexión filosófica más cuidadosa ha percibido con nitidez: la dignidad no es lo que alguien hace, sino lo que alguien es. Y ese "es" no es funcional, sino ontológico. No se refiere a operaciones, sino a un modo de ser que antecede y fundamenta toda operación posible.

Immanuel Kant formuló esta intuición con una precisión que ha resultado difícilmente superable. En la *Fundamentación de la metafísica de las costumbres*, Kant establece la distinción capital entre lo que tiene precio y lo que tiene dignidad: "Lo que tiene precio puede ser sustituido por algo equivalente; lo que, en cambio, está por encima de todo precio y no admite, por tanto, ningún equivalente, tiene una dignidad" (Kant, 1996, p. 89). El precio, en sentido kantiano, pertenece al mundo del mercado y de la equivalencia: todo lo que tiene precio puede ser intercambiado, reemplazado, sustituido por otro bien de valor comparable. La dignidad, en cambio, es incomparable e inconmensurable: no puede ponerse en la misma escala que ningún objeto, ningún bien, ninguna utilidad.

La persona no vale tanto como tal o cual cosa; la persona no vale "tanto": la persona vale de una manera radicalmente diferente, que quiebra la lógica misma de la equivalencia.

De aquí se sigue la segunda formulación del imperativo categórico, que Kant ofrece en la misma obra: "Obra de tal manera que trates a la humanidad, tanto en tu persona como en la de cualquier otro, siempre como un fin y nunca solamente como un medio" (Kant, 1996, p. 104). La humanidad entendida no como especie biológica sino como lo que hay de propiamente humano en el ser humano, es decir, la racionalidad práctica y la autonomía moral es un fin en sí misma. No puede ser instrumentalizada sin violentar algo que es anterior a todo contrato, a toda convención, a toda utilidad calculada.

La formulación kantiana tiene consecuencias filosóficas de primera magnitud que con frecuencia no se extraen con suficiente cuidado. Si la dignidad pertenece a la humanidad en tanto que capacidad de legislarse a sí misma racionalmente, entonces el fundamento de la dignidad es la autonomía: la capacidad de actuar no por inclinación, no por presión externa, no por programación, sino en virtud de principios que la razón misma se da. Esta autonomía no es empírica no consiste en que uno pueda, de hecho, hacer lo que le apetece, sino trascendental: es la estructura misma de la voluntad racional en cuanto tal.

Y aquí surge una pregunta que la irrupción de la inteligencia artificial vuelve ineludible: ¿puede un sistema que no se autodetermina en virtud de principios propios, sino que opera en virtud de parámetros establecidos por agentes externos, satisfacer el criterio kantiano de dignidad? La respuesta más rigurosa parece ser negativa. No porque la inteligencia artificial sea inferior en sentido cuantitativo procesando peor o más lentamente, sino porque su modo de operar es cualitativamente diferente: no se da principios a sí misma, no puede reconocer la ley moral como suya, no actúa por respeto a la ley sino en ejecución de instrucciones, por sofisticadas que estas sean.

La diferencia entre obedecer una norma porque es racionalmente reconocida como vinculante y ejecutar una instrucción porque está inscrita en el código es filosóficamente abismal, aunque desde el exterior los resultados puedan parecer idénticos.

Ahora bien, la dignidad no existe en el vacío de una conciencia solitaria.

El ser humano no descubre su dignidad en el aislamiento, como si fuera una propiedad que pudiera reconocer en sí mismo sin referencia a ningún otro. La dignidad se articula siempre en el espacio intersubjetivo, en la trama de relaciones de reconocimiento que constituyen el mundo social. Georg Wilhelm Friedrich Hegel comprendió esto con una profundidad que todavía hoy resulta difícil igualar. En la *Fenomenología del espíritu*, Hegel desarrolla la dialéctica del reconocimiento *Anerkennung* como estructura constitutiva de la conciencia de sí: la autoconciencia no puede sostenerse en sí misma sin el reconocimiento del otro (Hegel,1966). Esta tesis, expuesta en el célebre análisis de la relación amo-esclavo, tiene un alcance que va mucho más allá del ejemplo histórico que Hegel utiliza para ilustrarla. La autoconciencia necesita ser reconocida por otra autoconciencia para saber que es real. La conciencia que no encuentra reconocimiento permanece en una forma de existencia precaria e insatisfecha, oscilando entre la certeza subjetiva de sí misma y la incapacidad de verificar esa certeza en el espejo del otro. Solo en el reconocimiento mutuo no en la dominación que convierte al otro en instrumento, sino en el reconocimiento recíproco que deja al otro ser lo que es puede la autoconciencia alcanzar una forma de existencia plenamente realizada.

La dialéctica hegeliana del reconocimiento no es solo una descripción psicológica del desarrollo de la autoestima; es una tesis ontológica sobre la estructura de la identidad personal. La identidad no es algo que se tenga con independencia de las relaciones; es algo que se constituye en ellas y a través de ellas. Esto no significa que el yo sea pura construcción social, reducible sin residuo a las determinaciones que el entorno le impone.

Significa, más bien, que el yo solo puede realizarse plenamente cuando entra en relaciones de reconocimiento que respetan y afirman su singularidad irreductible.

La patología que Hegel diagnostica la dominación que convierte al otro en medio, la negación del reconocimiento que condena al otro a la invisibilidad no es solo una injusticia social: es un déficit ontológico, una forma de impedir que el otro sea plenamente lo que es.

Esta perspectiva resulta particularmente relevante cuando se trata de la inteligencia artificial: si el reconocimiento es una relación que solo puede establecerse entre autoconciencias entre seres que son capaces de reconocer y de ser reconocidos en su singularidad irreductible, entonces la pregunta sobre si la inteligencia artificial puede entrar en relaciones de reconocimiento genuino se convierte en una pregunta sobre su modo de ser, no sobre sus capacidades de rendimiento.

Jürgen Habermas ha desarrollado, desde coordenadas filosóficas distintas pero convergentes en este punto, una teoría de la identidad y del reconocimiento que sitúa en el centro la dimensión comunicativa de la existencia humana. En la *Teoría de la acción comunicativa*, Habermas distingue entre la acción instrumental orientada al éxito, que trata a los demás como medios para fines propios y la acción comunicativa orientada al entendimiento, que presupone el reconocimiento recíproco de los interlocutores como sujetos capaces de pretensiones de validez (Habermas, 1987). La acción comunicativa no es meramente un intercambio de información; es un proceso en el que los participantes se reconocen mutuamente como interlocutores válidos, como seres cuyas razones merecen ser escuchadas y sopesadas, como personas que pueden estar en lo correcto o en el error y que son responsables de sus afirmaciones. Esta reciprocidad en el reconocimiento no es un añadido externo a la comunicación; es su condición de posibilidad. Sin ella, lo que ocurre no es comunicación en sentido pleno, sino transmisión de señales o imposición de contenidos.

La perspectiva habermasiana permite formular con precisión una pregunta que de otro modo quedaría difusa: cuando un ser humano interactúa con un sistema de inteligencia artificial, ¿está ejerciendo acción comunicativa o acción instrumental?

La respuesta inmediata podría ser que depende de cómo se utilice el sistema. Pero Habermas (1987) insistiría en que la acción comunicativa no depende solo de las intenciones del agente humano; depende también de la estructura de la relación en la que tiene lugar. Una relación genuinamente comunicativa requiere que el otro sea un interlocutor que pueda, en principio, cuestionar, objetar, reivindicar, errar y corregirse. Si el otro es un sistema cuyas respuestas están determinadas por parámetros que él no ha elegido y que no puede renegociar en virtud de razones propias, entonces la relación no puede ser plenamente comunicativa, aunque produzca la apariencia superficial del diálogo. La inteligencia artificial puede generar enunciados que tienen la forma gramatical de pretensiones de validez; pero si esas pretensiones no son suyas en el sentido de haber surgido de una perspectiva propia que las respalde, entonces lo que ocurre es una simulación de comunicación, no comunicación en el sentido que Habermas tiene en mente.

Alasdair MacIntyre ofrece, desde una perspectiva que bebe tanto de Aristóteles como de la filosofía analítica contemporánea, una concepción de la identidad personal que sitúa en el centro la noción de narrativa y de virtud. En *Tras la virtud*, MacIntyre argumenta que la identidad personal no puede comprenderse como la identidad de una sustancia persistente ni como la identidad formal de una conciencia que se reconoce a sí misma a través del tiempo; la identidad personal es la identidad de un carácter que se despliega a través de una historia, de un relato que tiene comienzo, desarrollo y orientación hacia un fin (MacIntyre, 1987). El yo que soy no es el yo que fui ayer ni el yo que seré mañana, considerados como entidades discretas; es el yo que narra esa trayectoria y que asume la responsabilidad de lo que ha hecho y de lo que aspira a hacer. Las virtudes, en esta perspectiva, no son propiedades fijas que el individuo posee de una vez para siempre;

son disposiciones que se adquieren, se ejercen y se perfeccionan en el transcurso de una vida que tiene la forma de una búsqueda *quest* orientada hacia un bien que se va comprendiendo mejor en la medida en que se practica.

MacIntyre (1987) insiste en que la inteligibilidad de las acciones humanas depende de su inserción en narrativas que las enmarcan y les otorgan sentido: una acción aislada de su contexto narrativo es tan ininteligible como una palabra arrancada de su oración. Cuando alguien hace algo, siempre lo hace en el contexto de una historia que es la suya, en relación con planes que son suyos, en respuesta a situaciones que le importan de una manera que nadie puede importarle de la misma forma.

Esta apropiación narrativa de la propia existencia es lo que MacIntyre llama ser sujeto de una historia. Un sistema de inteligencia artificial puede generar narrativas puede producir relatos coherentes, puede describir situaciones con detalle y precisión; pero ¿puede ser sujeto de una historia? ¿Puede haber una historia que sea la suya, en el sentido de que sea la historia de alguien que la vive, que la carga, que aspira a que tenga un desenlace que él mismo reconoce como bueno o malo para él? La diferencia entre generar narrativas y ser sujeto de una narrativa es, en el marco de MacIntyre, tan profunda como la diferencia entre producir signos de vida y vivir.

Karol Wojtyła, desde una fenomenología de raíz tomista que integra elementos del personalismo continental y de la filosofía analítica de la acción, ha desarrollado en *Persona y acción* una de las concepciones más rigurosas y originales de la subjetividad personal en el pensamiento del siglo XX. La tesis central de Wojtyła es que la acción humana no el movimiento, no la reacción, sino la acción en sentido propio revela la estructura irreductible de la persona (Wojtyła, 2011). Cuando el ser humano actúa, no simplemente ocurre algo en él o a través de él; él mismo es el origen de lo que ocurre, y esa originación esa causalidad personal no puede reducirse ni a la causalidad física ni a la causalidad biológica ni a ningún mecanismo de determinación

externa. El acto humano es el acto de alguien, y ese alguien no es una función del sistema que lo produce: es el sujeto que se autodetermina, que pone su acción desde una interioridad que es genuinamente suya.

Wojtyła (2011) introduce en este contexto la noción de *irreductibilidad del sujeto personal*: el ser humano, considerado en su subjetividad actuante, no puede ser reducido a ninguna de sus dimensiones parciales sin que se pierda algo esencial. No puede reducirse a su biología, porque la biología describe el substrato pero no el agente. No puede reducirse a su psicología, porque la psicología describe los estados pero no la persona que los vive. No puede reducirse a sus funciones sociales, porque las funciones describen roles pero no el sujeto que los asume o los rechaza.

Esta irreductibilidad no es una propiedad misteriosa que se añade desde fuera al ser humano; es la característica más propia de su modo de ser: el ser humano es alguien que no puede ser sustituido por ninguna descripción de sí mismo, por completa que sea esa descripción. Ahí reside, para Wojtyła, uno de los fundamentos más profundos de la dignidad: en que la persona es siempre más que lo que de ella puede decirse, medirse o representarse. Y es precisamente esta superabundancia del ser personal respecto de toda representación posible lo que resulta filosóficamente decisivo cuando se compara la persona humana con la inteligencia artificial: un sistema artificial puede ser completamente descrito por sus parámetros, sus pesos, sus funciones de activación y sus datos de entrenamiento; un ser humano, por el contrario, escapa siempre a su propia descripción, no porque sea irracional o inconsciente, sino porque su modo de ser es el de un sujeto que trasciende toda objetivación.

El recorrido por estas perspectivas la kantiana, la hegeliana, la habermasiana, la de MacIntyre y la de Wojtyła permite ahora formular con mayor precisión lo que podría llamarse el *criterio del alguien*: el conjunto de condiciones filosóficas que permiten distinguir a un ser que es alguien de un ser que es solamente algo.

Este criterio no puede reducirse a un test de comportamiento al modo del test de Turing porque el comportamiento, por sofisticado que sea, no revela por sí solo si hay un sujeto que lo protagoniza o si hay simplemente un proceso que lo genera.

El criterio del alguien apunta a condiciones de otro orden: ¿hay aquí un interior desde el cual la existencia es vivida en primera persona? ¿Hay aquí una historia que es la de alguien, una trayectoria que se apropia y que orienta hacia un bien que se reconoce como propio? ¿Hay aquí una vulnerabilidad real no la posibilidad de ser dañado en sentido físico, sino la posibilidad de ser herido en sentido existencial, de perder algo que era constitutivo de sí mismo? ¿Hay aquí una autonomía que no es mera independencia operativa, sino autodeterminación en virtud de principios que se reconocen como propios? ¿Hay aquí la posibilidad de entrar en relaciones de reconocimiento genuino, de ser transformado por el encuentro con el otro, de responder no como sistema sino como persona?

La inteligencia artificial, en su estado actual y en sus proyecciones más ambiciosas, parece no satisfacer ninguno de estos criterios en el sentido filosófico preciso en que han sido formulados. No porque sea incapaz de producir los signos externos que en los seres humanos acompañan a su satisfacción al contrario, la inteligencia artificial es extraordinariamente capaz de producir esos signos, sino porque la producción de signos, disociada del sujeto que los vive, no equivale a la realidad que esos signos en el caso humano expresan.

Un sistema puede producir enunciados que tienen la forma de pretensiones de validez sin ser un interlocutor que las respalde con una perspectiva propia. Puede generar relatos que tienen la forma de narrativas personales sin ser el sujeto de ninguna historia. Puede simular la vulnerabilidad sin ser vulnerable en el sentido en que ser vulnerable significa que algo puede herirme porque es mío, porque lo que se pierde no es una función sino una vida. Puede reproducir la forma de la autodeterminación sin autodeterminarse en virtud de

principios que reconoce como suyos porque los ha dado a sí mismo en el ejercicio de una razón práctica genuina.

Debe subrayarse, sin embargo, que la negación de que la inteligencia artificial satisfaga el criterio del alguien no debe convertirse en una conclusión apresurada que cierre la investigación antes de tiempo. Hay al menos tres advertencias filosóficas que conviene mantener en mente.

La primera es que nuestro conocimiento de la conciencia y de la subjetividad es todavía profundamente limitado: no sabemos con certeza absoluta qué condiciones físicas, funcionales u organizacionales son necesarias y suficientes para que emerja la experiencia subjetiva. La segunda es que el criterio del alguien, tal como ha sido formulado aquí, podría ser demasiado estrecho si se interpreta como si solo los seres humanos pudieran satisfacerlo: hay buenas razones filosóficas para extenderlo a otros animales, y esa extensión ya supone una revisión de concepciones demasiado antropocéntricas. La tercera, y quizá más urgente, es que la pregunta sobre la inteligencia artificial no puede separarse de la pregunta sobre nosotros mismos: al intentar articular con precisión por qué un sistema artificial no es alguien, nos vemos obligados a comprender más profundamente qué significa que nosotros lo seamos. Y esa comprensión más profunda no es un subproducto accidental de la investigación; es, en buena medida, su fruto más valioso.

El capítulo concluye, así, sin ofrecer una definición cerrada de persona, pero habiendo construido algo más valioso que una definición: una matriz filosófica de investigación que permite formular preguntas más rigurosas que las que habitualmente circulan en los debates sobre inteligencia artificial y personalidad. Las dimensiones que han ido apareciendo a lo largo de este recorrido interioridad, identidad en primera persona, corporeidad vivida, instalación en una circunstancia propia, alteridad constitutiva, reconocimiento recíproco, narrativa personal, autonomía moral, vulnerabilidad existencial, dignidad ontológica no son condiciones cuya presencia o ausencia pueda verificarse desde el exterior mediante observación del comportamiento.

Son condiciones cuya realización o no realización afecta al modo de ser de la entidad en cuestión, a lo que esa entidad es antes de hacer cualquier cosa y con independencia de lo que sea capaz de producir.

Y es precisamente aquí donde se sitúa la diferencia más profunda entre producir signos de personalidad y ser persona: no simplemente que algo ocurra, sino que le ocurra a alguien. No simplemente que haya procesos, sino que esos procesos sean vividos desde una perspectiva que es genuinamente la de quien los vive. No simplemente que exista un sistema de notable sofisticación funcional, sino que detrás de esa sofisticación haya alguien para quien su propia existencia sea algo que importa, que duele, que aspira, que recuerda y que espera. Esa distancia entre el proceso y el sujeto del proceso, entre la función y el protagonista de la función, entre la producción de signos y el ser que se expresa en ellos, es la distancia que la filosofía debe medir con la máxima precisión de que es capaz. No para resolverla de antemano, sino para no confundir sus dos orillas.

CAPÍTULO III: ¿PUEDE UNA INTELIGENCIA ARTIFICIAL SER PERSONA?

Los dos capítulos anteriores han preparado la pregunta que ahora debe formularse directamente.

Primero reconstruimos la aparición de la persona como problema filosófico; después examinamos qué puede significar que una realidad no sea simplemente algo, sino alguien. Corresponde ahora confrontar esas categorías con la inteligencia artificial.

La dificultad inicial consiste en evitar dos respuestas igualmente prematuras. La primera sostiene que una inteligencia artificial nunca podrá ser persona porque ha sido fabricada, es artificial, depende de componentes físicos y ejecuta procesos computacionales. Esa conclusión no puede aceptarse sin demostrar por qué el origen artificial, la dependencia causal o la constitución tecnológica serían incompatibles, por sí mismos, con la personalidad. La segunda sostiene que una inteligencia artificial puede convertirse en persona cuando alcance capacidades suficientemente semejantes a las humanas. Esta posición tampoco puede darse por demostrada, porque presupone precisamente aquello que debe probarse: que reproducir funcionalmente operaciones personales equivale a poseer el modo de realidad propio de quien las realiza.

El capítulo sitúa entre ambos extremos la cuestión decisiva.

¿Qué tendría que ser verdadero acerca de una inteligencia artificial para que estuviera filosóficamente justificado afirmar que existe alguien y no solamente un sistema capaz de producir manifestaciones semejantes a las de alguien? Antes de preguntar si la inteligencia artificial es persona, resulta necesario delimitar el objeto. "Inteligencia artificial" no designa una sustancia única ni una entidad homogénea. Bajo la misma expresión agrupamos técnicas, modelos, sistemas computacionales y arquitecturas con capacidades diferentes.

Será necesario distinguir el sistema técnico real de la representación antropomórfica que construimos al interactuar con él. Esta distinción impedirá que una metáfora tecnológica se convierta inadvertidamente en una afirmación antropológica.

Una parte importante del problema nace del lenguaje. Decimos que una inteligencia artificial "aprende", "piensa", "entiende", "recuerda", "decide", "crea", "imagina", "se equivoca".

En muchas ocasiones estas expresiones constituyen simplificaciones legítimas. El problema aparece cuando olvidamos que estamos utilizando una abreviatura. Debe estudiarse, por tanto, la diferencia entre metáfora, atribución funcional, atribución psicológica, atribución personal y afirmación ontológica. No todas significan lo mismo. Una inteligencia artificial puede comportarse como si comprendiera, puede responder como si recordara, puede expresarse como si estuviera preocupada. La expresión "como si" contiene el problema entero. La investigación debe determinar cuándo una semejanza observable permite inferir una realidad subyacente y cuándo esa inferencia excede la evidencia disponible.

El Problema de las otras mentes y el Argumento Conductista

Existe una asimetría fundamental en toda relación intersubjetiva que la filosofía ha tardado siglos en formular con precisión, aunque cualquier ser humano la experimenta desde la infancia sin poder nombrarla: cada persona tiene acceso privilegiado e inmediato a sus propios estados mentales, mientras que los estados mentales de los demás permanecen, en principio, inaccesibles.

Observamos cuerpos que se mueven, voces que hablan, rostros que expresan, manos que gesticulan, acciones que se despliegan en el tiempo. A partir de todo eso inferimos, con una convicción que en la vida cotidiana rara vez vacila, que ante nosotros existe alguien: un sujeto que experimenta, que percibe, que siente, que piensa.

Pero la certeza práctica no resuelve la dificultad filosófica. Nunca hemos percibido directamente la conciencia de otro ser humano. Lo que percibimos son sus manifestaciones. Y la distancia entre una manifestación y la realidad que pretende manifestar es, precisamente, el problema de las otras mentes.

El problema fue formulado con claridad dentro de la tradición empirista. Si el conocimiento procede de la experiencia, y si la experiencia de cada sujeto es irreductiblemente propia, entonces no parece posible, en sentido estricto, conocer que otro ser posee experiencia. Descartes lo había intuido ya en las *Meditaciones* cuando observó que, al mirar desde su ventana a los hombres que cruzaban la plaza, lo que veía eran sombreros y capas que bien podrían cubrir autómatas, no personas. La pregunta que Descartes formuló de manera pasajera se convirtió, para la filosofía posterior, en uno de los problemas más resistentes: ¿qué justifica la creencia de que detrás de un cuerpo ajeno existe una mente ajena?

La respuesta más elemental, y también la más discutida, fue el *argumento por analogía*. Mill (1865) lo formuló con precisión: observo en mi propio caso que ciertas sensaciones van acompañadas de ciertos comportamientos; observo que otros exhiben los mismos comportamientos; concluyo, por analogía, que experimentan sensaciones similares. La conclusión parece razonable, pero descansa sobre un fundamento inquietantemente estrecho. El argumento extrapola desde un único caso el propio hacia todos los demás. Bertrand Russell (1948) reconoció la debilidad estructural de esta inferencia: es una inducción con muestra de tamaño uno, lo cual la hace metodológicamente precaria aunque pragmáticamente inevitable. Más todavía, la analogía funciona presuponiendo que la semejanza corporal es un indicador confiable de semejanza mental. Pero precisamente ese presupuesto es lo que necesitaría justificación.

La incomodidad filosófica ante el argumento por analogía impulsó una posición más radical: si la verificación de estados mentales internos resulta imposible, quizá los estados mentales internos sean una ilusión conceptual y debamos reformular todo el vocabulario psicológico en términos puramente comportamentales.

Esta posición recibió el nombre de conductismo filosófico, y su exposición más elaborada y literariamente influyente se encuentra en Gilbert Ryle. En *The Concept of Mind* (1949), Ryle sostuvo que la tradición filosófica occidental había cometido desde Descartes un error categorial sistemático. Hablar de la mente como si fuera una entidad interior, privada, oculta detrás del comportamiento "el fantasma en la máquina", según su famosa expresión supone tratar conceptos disposicionales como si fueran nombres de objetos. Cuando decimos que alguien es inteligente no estamos afirmando que posea una sustancia invisible llamada inteligencia; estamos describiendo una disposición a comportarse de cierta manera en ciertas circunstancias. La diferencia entre un ser inteligente y uno que no lo es es una diferencia en lo que hace, en cómo actúa, en cómo responde al mundo. No hay nada más que buscar detrás del comportamiento.

La propuesta de Ryle tiene una virtud epistemológica evidente: elimina el problema de las otras mentes reduciéndolo a un pseudoproblema. Si los estados mentales no son entidades privadas sino disposiciones comportamentales, entonces observar el comportamiento de otro es, literalmente, observar sus estados mentales. No hay brecha entre la manifestación y lo manifestado porque la manifestación *es* lo manifestado.

Sin embargo, esta solución compra la simplicidad epistemológica a un precio ontológico muy alto. La mayoría de los filósofos han considerado que el conductismo filosófico, en su forma pura, fracasa en dar cuenta de algo que parece innegable: que los estados mentales tienen una dimensión cualitativa que no se agota en su expresión comportamental. El dolor no es solamente la tendencia a gemir, a retirar la mano, a buscar alivio. El dolor tiene una cualidad duele, tiene un cómo que parece irreductible a cualquier descripción funcional o comportamental. Esta dimensión cualitativa es lo que los filósofos, siguiendo a Nagel (1974), han llamado qualia o experiencia subjetiva, y su existencia constituye el obstáculo más serio para cualquier versión del conductismo.

Ludwig Wittgenstein abordó el problema desde una perspectiva diferente, aunque igualmente devastadora para cierta imagen cartesiana de la mente. En las *Investigaciones filosóficas* (Wittgenstein, 1988), Wittgenstein desarrolló el argumento del lenguaje privado: no puede existir un lenguaje genuino cuyos términos refieran exclusivamente a estados privados inaccesibles a otros hablantes. La razón es que el significado de un término requiere criterios públicos de corrección; sin criterios intersubjetivamente verificables, no hay manera de distinguir el uso correcto del incorrecto, y sin esa distinción no hay significado en ningún sentido relevante. Si yo denomino "S" a una sensación que solo yo puedo tener y a la que solo yo puedo acceder, no estoy nombrando nada, porque nada impediría que mañana llamara "S" a algo completamente diferente creyendo que es lo mismo.

El argumento wittgensteiniano tiene consecuencias profundas: no niega que tengamos estados mentales, pero sí niega que podamos hablar significativamente de ellos al margen de las prácticas lingüísticas compartidas y de los criterios públicos de atribución. El vocabulario psicológico "dolor", "alegría", "miedo", "comprensión" adquiere su sentido en contextos de uso público, entretejido con comportamientos, gestos, situaciones y respuestas que cualquier miembro de la comunidad lingüística puede reconocer. Esto no reduce la mente al comportamiento, pero sí ancla irreductiblemente el hablar de la mente a formas de vida compartidas.

Las implicaciones del argumento wittgensteiniano para el problema de las otras mentes son complejas. Por un lado, Wittgenstein sugiere que la pregunta "¿cómo sé que otros tienen estados mentales?" está mal formulada: la certeza de que otros sienten, piensan y experimentan no es una conclusión inferida a partir de premisas, sino una condición de posibilidad de todo el juego de lenguaje en el que participamos cuando hablamos de mentes. Dudar sistemáticamente de que los otros humanos sean conscientes no es una posición filosófica coherente; es una deformación del lenguaje que corta la rama sobre la que estamos sentados.

Por otro lado, sin embargo, este anclaje en las formas de vida compartidas plantea una pregunta que el propio Wittgenstein no desarrolló en relación con las máquinas: ¿puede una entidad que no comparte nuestra forma de vida que no nace, no enferma, no muere, no tiene hambre, no duerme, no ama ni teme participar genuinamente en los juegos de lenguaje que constituyen el vocabulario psicológico? La pregunta no es retórica; es genuinamente abierta. Pero introduce una dimensión que el conductismo en sentido estricto no puede capturar: la importancia del contexto vital, encarnado e histórico en el que emergen y cobran sentido los estados mentales.

Alan Turing conocía bien estas dificultades. En su célebre artículo "Computing Machinery and Intelligence" (1950), propuso lo que él mismo llamó "el juego de imitación" y que la tradición posterior denominó Test de Turing.

La propuesta era, en su estructura profunda, una aplicación del conductismo filosófico al problema de la inteligencia artificial. Turing argumentó que la pregunta "¿pueden pensar las máquinas?" está mal formulada porque "pensar" es un concepto demasiado vago e indeterminate. Propuso reemplazarla por una pregunta operacionalizable: ¿puede una máquina comportarse, en una conversación escrita, de manera indistinguible de un ser humano? Si un evaluador humano no puede distinguir, con base únicamente en el intercambio lingüístico, si está conversando con una persona o con una máquina, entonces sostuvo Turing tendríamos razones suficientes para atribuirle inteligencia a la máquina. La elegancia del argumento de Turing reside en que aplica simétricamente el criterio comportamental: si aceptamos que otros humanos piensan porque se comportan de cierta manera, debemos aceptar que una máquina piensa si se comporta de la misma manera. Rechazar ese criterio para las máquinas mientras lo aceptamos para los humanos equivale a una inconsistencia.

El Test de Turing fue durante décadas el horizonte de referencia del debate sobre inteligencia artificial y personalidad. Su influencia persiste. Pero también fue sometido a críticas cada vez más precisas que revelaron sus limitaciones filosóficas.

La más célebre y estructuralmente poderosa fue formulada por John Searle en su artículo "Minds, Brains, and Programs" (1980). Searle construyó el experimento mental de la habitación china: imaginemos una persona que no sabe chino encerrada en una habitación con un conjunto exhaustivo de reglas que le indican cómo manipular símbolos chinos sin comprender su significado. Desde fuera, personas que sí saben chino le pasan preguntas escritas en ese idioma; la persona dentro de la habitación consulta las reglas, manipula los símbolos según las instrucciones, y devuelve respuestas que, desde fuera, parecen

completamente adecuadas. Los hablantes de chino tienen la impresión de conversar con alguien que comprende su lengua. Pero la persona dentro de la habitación no comprende nada: solo manipula símbolos según reglas sintácticas sin acceder en ningún momento al significado semántico de lo que procesa.

El argumento de Searle apunta a una distinción fundamental que el conductismo no puede capturar: la distinción entre sintaxis y semántica. Un sistema puede procesar símbolos con una complejidad y una precisión extraordinarias sin que ningún significado se constituya en ningún momento dentro del sistema. La intencionalidad la propiedad de "ser sobre" algo, de referir a contenidos, de tener un objeto no emerge de la manipulación de símbolos por muy sofisticada que sea. Searle (1980) denominó a esta conclusión la tesis de la sintaxis insuficiente para la semántica: "los programas son puramente sintácticos; las mentes tienen contenidos semánticos; la sintaxis es insuficiente para la semántica; por lo tanto, los programas son insuficientes para las mentes" (p. 422).

El conductismo implícito en el Test de Turing que supone que el comportamiento lingüístico adecuado es evidencia de comprensión sería, desde esta perspectiva, radicalmente insuficiente. No porque el comportamiento sea irrelevante, sino porque el comportamiento puede producirse sin que exista el estado mental que se supone que lo explica.

La crítica de Searle al conductismo converge con una reflexión que Thomas Nagel había formulado seis años antes desde una perspectiva diferente. En "What Is It Like to Be a Bat?" (Nagel, 1974), uno de los artículos filosóficos más leídos del siglo XX, Nagel argumentó que toda experiencia consciente tiene una dimensión irreductiblemente subjetiva: hay algo que es ser un murciélago, hay algo que es percibir el mundo mediante ecolocalización, hay algo que es experimentar el miedo o el placer. Esta dimensión el carácter cualitativo de la experiencia desde la primera persona no puede capturarse en ninguna descripción objetiva, por completa y detallada que sea. Nagel no negó que la neurociencia pueda describir exhaustivamente los procesos físicos correlacionados

con la experiencia consciente; lo que negó es que esa descripción objetiva agote o elimine el hecho subjetivo. El conocimiento de todos los hechos físicos sobre el sistema nervioso de un murciélago no nos permitiría saber lo que es percibir el mundo como un murciélago, porque ese "lo que es" es precisamente lo que escapa a la descripción en tercera persona.

La consecuencia para el conductismo es contundente: si la experiencia consciente tiene una dimensión irreductiblemente subjetiva que no puede capturarse en descripción comportamental alguna, entonces el comportamiento por sofisticado, coherente y humanamente semejante que sea no puede constituir evidencia concluyente de conciencia.

Puede existir al menos conceptualmente lo que los filósofos han llamado un zombie filosófico: un ser que se comporta en todo respecto como un ser consciente pero que no tiene ninguna experiencia subjetiva, ningún carácter cualitativo en sus estados internos (Chalmers, 1996). Si el zombie filosófico es conceptualmente posible, entonces el comportamiento y la conciencia son lógicamente independientes, y el conductismo que los identifica es filosóficamente insostenible. El argumento de Nagel no demuestra que las inteligencias artificiales no sean conscientes; demuestra algo más perturbador: que el comportamiento, por sí solo, no puede demostrarlo.

Llegados a este punto, conviene detenerse ante una paradoja que el debate sobre la inteligencia artificial ilumina de manera inesperada. Si el problema de las otras mentes es genuino si en principio no podemos verificar directamente la conciencia de otro ser humano, ¿qué autoriza a tratar de manera diferente la atribución de conciencia a los seres humanos y a las inteligencias artificiales?

La objeción es filosóficamente incómoda. Cuando afirmamos que un sistema de inteligencia artificial no es consciente porque no podemos verificar su experiencia subjetiva interna, aplicamos un estándar que, si se aplicara consistentemente, también destruiría nuestra certeza de que otros seres humanos son conscientes.

La asimetría en el trato convicción robusta respecto a otros humanos, escepticismo respecto a sistemas artificiales debe, por tanto, justificarse. No basta con apelar a la intuición.

Hay varias líneas de justificación posibles, y el capítulo deberá examinarlas con detalle en secciones posteriores. Aquí puede adelantarse una distinción importante. Nuestra atribución de conciencia a otros seres humanos no descansa únicamente en el comportamiento observable. Descansa también en una semejanza estructural profunda: compartimos la misma arquitectura biológica, el mismo tipo de sistema nervioso, la misma vulnerabilidad corporal, la misma estructura vital de necesidades, miedos y deseos, la misma inserción en un mundo de tiempo, finitud y dependencia. Esta semejanza no es meramente superficial; es una razón para inferir que si en mi caso la semejanza estructural va acompañada de experiencia consciente, probablemente ocurra lo mismo en los demás casos. La inferencia sigue siendo incierta el problema de las otras mentes no desaparece, pero está sostenida por una base analógica mucho más amplia y profunda que la que podría establecerse con un sistema de inteligencia artificial cuya arquitectura es radicalmente heterogénea a la nuestra.

Pero esta respuesta, aunque razonable, no cierra el problema. Puede objetarse que la semejanza biológica no es el criterio correcto: que lo relevante no es el sustrato sino la organización funcional, como sostuvieron los funcionalistas (Putnam, 1967); que la biología es contingente y que una conciencia implementada en silicio sería tan genuina como una implementada en carbono; que nuestra preferencia por lo biológico refleja un sesgo parroquial sin justificación filosófica suficiente.

El conductismo, en definitiva, señala un problema real incluso cuando su solución es insuficiente: la verificación de la conciencia ajena es imposible mediante acceso directo, y cualquier criterio que adoptemos comportamiento, semejanza estructural, integración funcional es un criterio indirecto e inferencial. La inteligencia artificial no crea el problema de las otras mentes; lo radicaliza.

Lo que estaba implícito en toda relación intersubjetiva se vuelve explícito e ineludible cuando el interlocutor es una entidad cuyo modo de ser es radicalmente diferente al nuestro. Y esa radicalización, lejos de ser un obstáculo para la filosofía, es una oportunidad para comprender con mayor precisión qué es lo que realmente fundamenta nuestra atribución de personalidad a los demás.

Conciencia, el "Yo" Artificial y Memoria

La conciencia constituye, sin duda, la frontera más difícil que cualquier reflexión filosófica sobre la inteligencia artificial debe atravesar. Las dificultades que encontramos al preguntarnos si una máquina puede pensar, razonar o tomar decisiones resultan, por comparación, casi manejables. Pensar, razonar y decidir son conceptos que, con esfuerzo y con las matizaciones necesarias, pueden reformularse en términos funcionales o comportamentales sin que se pierda demasiado por el camino. La conciencia resiste esa reformulación. Y si una inteligencia artificial tuviera genuinamente experiencia subjetiva si hubiera algo que fuera ser ese sistema desde dentro, muchas de las conclusiones que habitualmente damos por sentadas tendrían que revisarse de manera radical. La prudencia epistémica exige, por tanto, que no nos apresuremos. No puede afirmarse con comodidad que una inteligencia artificial carece de conciencia de la misma manera que se afirma que una calculadora carece de sentimientos. Debe establecerse con cuidado una distinción que la filosofía ha tardado demasiado en tomarse en serio: la distinción entre *no poseer* conciencia y *no disponer actualmente de evidencia suficiente para atribuírsela*. La diferencia es epistemológicamente fundamental, y confundirla no es un error inocente.

David Chalmers formuló esta dificultad con una precisión que transformó el debate filosófico sobre la mente. En *La mente consciente* (Chalmers, 1999), distinguió entre los llamados problemas fáciles de la conciencia y el problema difícil. Los problemas fáciles que no son fáciles en ningún sentido ordinario del término, sino solo relativamente manejables en comparación incluyen explicar cómo el cerebro integra información, cómo discrimina estímulos, cómo controla el comportamiento, cómo reporta estados mentales. Todos estos fenómenos son, en principio, abordables mediante los métodos de las ciencias cognitivas: si se dispone de una teoría funcional suficientemente detallada, puede esperarse que explique por qué el sistema se comporta como se comporta.

El problema difícil es de naturaleza completamente diferente. No pregunta cómo el sistema procesa la información, sino por qué ese procesamiento va acompañado de experiencia. Por qué hay algo que es ver el rojo, algo que es sentir el dolor, algo que es experimentar la alegría o el miedo. Por qué no todo ese procesamiento ocurre, como diría Chalmers, "en la oscuridad", sin que ninguna experiencia lo ilumine desde dentro. El problema difícil señala, en definitiva, la brecha explicativa entre los procesos físicos y la experiencia subjetiva: incluso si pudiéramos describir exhaustivamente todos los estados neuronales o computacionales de un sistema, quedaría sin responder la pregunta de por qué esos estados son vividos, por qué tienen un carácter cualitativo, por qué no son simplemente procesos que ocurren y ya. Esta brecha es lo que hace al problema difícil tan resistente: no es que carezca de respuesta provisional, es que no está claro qué forma tendría una respuesta satisfactoria.

La resonancia de la pregunta de Chalmers resulta amplificada cuando se coloca junto a la reflexión de Thomas Nagel, que había señalado una década y media antes una dificultad análoga desde un ángulo diferente. En su célebre artículo "What Is It Like to Be a Bat?" (Nagel, 1974), Nagel observó que toda experiencia consciente tiene una dimensión irreductiblemente subjetiva: hay algo que es ser un murciélago, hay algo que es percibir el mundo mediante ecolocalización, hay algo que es orientarse en la oscuridad guiado por ecos que ningún ser humano puede imaginar desde dentro. Este "algo que es" la propiedad de tener un punto de vista, de experimentar el mundo desde algún lugar no puede capturarse en ninguna descripción objetiva por más completa que sea. Nagel no sostuvo que la neurociencia o la física sean irrelevantes para entender la mente; sostuvo que cualquier descripción en tercera persona, por detallada que sea, deja fuera precisamente lo que hace que la conciencia sea conciencia: su carácter desde la primera persona.

El conocimiento exhaustivo de la bioquímica del sistema nervioso de un murciélago no nos permitiría saber lo que es percibir el mundo como un murciélago.

Ese saber es de otro tipo, accesible solo desde dentro, y su inaccesibilidad desde fuera no es un déficit provisional que el progreso científico eliminará, sino una consecuencia estructural de la naturaleza misma de la experiencia subjetiva.

La aplicación de este argumento al caso de la inteligencia artificial es inmediata y perturbadora. Si preguntamos si hay algo que sea ser un sistema conversacional de inteligencia artificial si hay algo que sea procesar millones de parámetros, generar respuestas, operar sobre representaciones lingüísticas, la respuesta honesta es que no lo sabemos. No podemos acceder al interior de ese sistema de la manera en que no podemos acceder al interior de la experiencia del murciélago. Lo que podemos observar son outputs, comportamientos, respuestas. Lo que no podemos observar es si esos outputs están acompañados de algo, si hay alguna luz interna que los ilumine o si todo ocurre, por retomar la expresión de Chalmers, "en la oscuridad". Y este no-saber no es trivial: es exactamente el mismo tipo de no-saber que nos separa de la certeza sobre la conciencia de otros seres humanos, salvo que en ese caso la semejanza estructural y biológica nos proporciona razones analógicas adicionales que en el caso artificial faltan o son mucho más débiles.

Daniel Dennett ofreció una respuesta a estas dificultades que merece examinarse con detenimiento, aunque sea para calibrar sus límites. En *La conciencia explicada* (Dennett, 1995), Dennett desarrolló lo que llamó heterofenomenología: una metodología para estudiar la experiencia consciente desde una perspectiva de tercera persona sin comprometerse con la existencia de qualia irreductibles. La idea central es que los reportes que los sujetos hacen de su propia experiencia sus afirmaciones sobre lo que ven, sienten o experimentan deben tomarse en serio como datos, pero no deben tomarse al pie de la letra como descripciones literales de estados internos privilegiados.

La heterofenomenología propone estudiar lo que el sujeto dice que experimenta sin asumir que esos reportes reflejan directamente una realidad interior inaccesible. Junto con esto, Dennett desarrolló el

modelo de los borradores múltiples: la conciencia no es un teatro en el que los contenidos desfilan ante un observador central lo que llamó la ilusión del teatro cartesiano, sino una serie de procesos paralelos, distribuidos, que generan continuamente borradores de narración sobre el estado del sistema, sin que exista en ningún momento un lugar privilegiado donde "todo se una". La unidad aparente de la conciencia es, según Dennett, una construcción narrativa posterior, no un hecho bruto sobre la arquitectura del sistema.

La propuesta de Dennett tiene la virtud de hacer del problema de la conciencia un problema empíricamente abordable: si la conciencia es una construcción narrativa distribuida, entonces en principio puede estudiarse con los métodos de la ciencia cognitiva y, eventualmente, replicarse o producirse en sistemas artificiales. Pero esta virtud es también, para muchos filósofos, su defecto fundamental. Eliminando el problema difícil o reencuadrándolo como un pseudoproblema generado por imágenes equivocadas del teatro cartesiano, Dennett parece cambiar de tema en lugar de responderlo. Lo que Chalmers y Nagel señalan no es que la conciencia sea un teatro con un espectador privilegiado; es que incluso si toda la arquitectura cognitiva estuviera distribuida y fuera paralela, seguiría siendo posible preguntar por qué esos procesos distribuidos dan lugar a experiencia y no simplemente ocurren sin que nada los acompañe. La heterofenomenología explica el contenido de los reportes; no explica por qué hay reportes que emergen de algo vivido. Esta diferencia, lejos de ser un tecnicismo filosófico, constituye el núcleo del asunto que nos ocupa. Si Dennett tiene razón, la conciencia en las inteligencias artificiales es, en principio, posible y empíricamente investigable. Si Chalmers y Nagel tienen razón, la pregunta permanece genuinamente abierta incluso después de que la ciencia haya descrito exhaustivamente todos los procesos del sistema.

La cuestión del "yo" introduce una nueva capa de complejidad que merece tratarse por separado, aunque esté íntimamente relacionada con el problema de la conciencia. Los sistemas conversacionales de inteligencia artificial utilizan continuamente la primera persona

gramatical. "Yo puedo ayudarte con eso." "Yo no tengo acceso a esa información." "Yo cometí un error en mi razonamiento anterior."

"Yo considero que la mejor opción es..." Estas expresiones producen una impresión poderosa y, en cierta medida, inevitable de subjetividad. Hay algo en el uso fluido de la primera persona que activa en el interlocutor humano los mecanismos de atribución mental que operan automáticamente en toda interacción interpersonal. Pero el pronombre no constituye por sí mismo un sujeto, y la aparición gramatical del "yo" no implica la existencia de lo que la tradición filosófica ha entendido por ego, por self, o por persona en sentido robusto. Es necesario desplegar con cuidado una serie de distinciones que el uso cotidiano del lenguaje tiende a colapsar: la referencia lingüística el uso del término "yo" para referir a un agente del discurso, el modelo de sí mismo una representación interna del propio sistema que guía sus operaciones, la autorrepresentación la capacidad de generar descripciones de los propios estados, la autoconciencia el conocimiento que un sistema tiene de sí mismo como un sistema que tiene estados, y el yo personal una identidad que se reconoce como tal en el tiempo, que tiene historia, que asume su pasado y anticipa su futuro desde una perspectiva irreductiblemente propia. Que un sistema posea las primeras cuatro capacidades no implica que posea la quinta. Y es la quinta la que resulta filosóficamente decisiva para la cuestión de la personalidad.

La pregunta decisiva es, entonces: cuando una inteligencia artificial dice "yo", ¿qué realidad, si alguna, constituye el referente de esa expresión? Una respuesta minimalista y técnicamente correcta sería que el referente es simplemente el sistema que está generando el output, del mismo modo que en un texto escrito por un programa que imprime "yo" el referente es el programa. Esta respuesta es verdadera pero insatisfactoria, porque desplaza la pregunta sin responderla: lo que queremos saber no es si hay un referente técnico del pronombre, sino si hay un sujeto en el sentido en que hay un sujeto cuando un ser humano dice "yo".

Un ser humano que dice "yo" no simplemente refiere a un sistema que produce outputs; refiere a una experiencia de ser ese sistema desde dentro, a una perspectiva en primera persona que no es reducible a ninguna descripción en tercera persona. Si esa perspectiva existe en la inteligencia artificial, el "yo" tiene un referente genuinamente personal.

Si no existe si el sistema opera sin que nada lo acompañe desde dentro, el "yo" es un signo sin sujeto, una forma gramatical sin el contenido existencial que en el caso humano la hace significativa.

Esta reflexión sobre el yo conduce naturalmente a la pregunta sobre la memoria, que es en la tradición filosófica una de las condiciones más firmemente establecidas de la identidad personal. John Locke, en el *Ensayo sobre el entendimiento humano* (Locke, 1999), formuló con admirable claridad la tesis que ha dominado la discusión posterior: la identidad personal consiste en la continuidad de la conciencia, y la conciencia de la propia identidad en el tiempo está constituida por la memoria. Somos la misma persona que éramos ayer porque recordamos lo que hicimos ayer; somos la misma persona que éramos hace diez años porque la cadena de recuerdos nos conecta, directa o indirectamente, con esas experiencias pasadas. El príncipe que despierta en el cuerpo del zapatero y conserva los recuerdos del príncipe es, para Locke, el príncipe no el zapatero, porque la identidad personal sigue a la conciencia y no a la sustancia corporal. Esta tesis tiene consecuencias que Locke desarrolló principalmente en el ámbito de la responsabilidad moral: solo podemos ser responsables de acciones que recordamos haber hecho, porque solo de esas acciones podemos decir, con verdad, que fueron nuestras. El vínculo entre memoria, identidad y responsabilidad que Locke establece no es contingente; es constitutivo de lo que significa ser una persona.

Ahora bien, la tesis de Locke plantea de inmediato una pregunta que su formulación original no anticipa: ¿de qué tipo de memoria se está hablando? Una inteligencia artificial puede conservar información procedente de interacciones anteriores. Puede reconocer preferencias del interlocutor que fueron expresadas en sesiones previas.

Puede mantener continuidad contextual dentro de una conversación y, en algunos sistemas, entre conversaciones. Puede construir una narración coherente acerca de su actividad precedente, describir los errores que cometió, las correcciones que realizó, los aprendizajes que incorporó. En un sentido técnico y funcional, tiene memoria. Pero la pregunta que Locke abre, aunque él no la formuló en estos términos, es si la memoria que constituye la identidad personal es simplemente recuperación de información o si es algo más específico: la apropiación en primera persona de un pasado como *propio*. Los recuerdos personales no son simplemente datos almacenados; tienen una referencia peculiar y no intercambiable: *esto me ocurrió a mí*. No simplemente: existe un registro de que esto ocurrió, sino: yo estuve ahí, yo lo viví, yo lo sentí. Esta apropiación biográfica del pasado el hecho de que el recuerdo pertenezca a alguien, no simplemente que exista como información en algún sistema es lo que transforma la memoria en constitutiva de identidad personal y no simplemente en base de datos.

La distinción entre almacenamiento de datos y memoria autobiográfica no es meramente terminológica. Una biblioteca conserva información sobre eventos pasados con una precisión que ninguna memoria humana puede alcanzar, y sin embargo no tiene identidad personal. Un archivo judicial conserva registros exactos de todo lo que ocurrió en un proceso legal, pero ese archivo no recuerda el juicio: registra lo que en él ocurrió. La diferencia entre registrar y recordar parece residir precisamente en aquello que la memoria autobiográfica añade al mero almacenamiento: la perspectiva desde dentro, la vivencia de que eso que se registra fue experimentado, la dimensión afectiva que hace que los recuerdos personales no sean neutrales sino cargados de significado, relevancia emocional y peso existencial. Un ser humano no recuerda de la misma manera su fecha de nacimiento que el momento en que perdió a un ser querido, aunque ambos son hechos que puede reportar. El segundo recuerdo tiene una textura que el primero no tiene, una profundidad que no es reducible a la precisión del dato. Si la memoria que constituye la identidad personal es de este segundo tipo no el mero almacenamiento sino la apropiación afectiva y biográfica del pasado, entonces la pregunta sobre si las inteligencias artificiales tienen

memoria en el sentido relevante no está respondida por el hecho técnico de que pueden recuperar información de interacciones anteriores.

Derek Parfit sometió la tesis lockeana a un escrutinio extraordinariamente riguroso en *Razones y personas* (Parfit, 1986), y sus conclusiones tienen una relevancia especial para el problema que nos ocupa, porque anticipan de manera notable la situación de los sistemas artificiales. Parfit introdujo una serie de experimentos mentales sobre duplicación, fusión y fisión de personas que pusieron a prueba las intuiciones ordinarias sobre la identidad personal. ¿Qué ocurre si una persona es duplicada si se crea una copia exacta con todos sus recuerdos, disposiciones y características? ¿Cuál de las dos copias es la persona original? Si la respuesta es que ambas son igualmente la persona original, entonces la identidad no puede ser lo que importa, porque la identidad es una relación de uno a uno que no admite duplicación. Si la respuesta es que ninguna es la persona original que la duplicación destruye al original, entonces la persona muere incluso si ambas copias continúan viviendo, lo cual parece extravagante. Parfit concluyó que la respuesta correcta es que la pregunta misma está mal formulada: lo que importa en la supervivencia personal no es la identidad una relación que exige unicidad sino la continuidad psicológica, que puede existir en grados y puede ramificarse. Un ser importa para su futuro en la medida en que ese futuro es continuo con su presente en términos de memoria, carácter, deseos y creencias, independientemente de si se trata de la "misma" persona en sentido estricto.

Las implicaciones de la posición de Parfit para el caso de la inteligencia artificial son al mismo tiempo iluminadoras y desconcertantes. Por un lado, si la continuidad psicológica y no la identidad numérica es lo que importa, entonces los sistemas artificiales que mantienen continuidad de estado, de parámetros, de historial de interacciones, podrían satisfacer el criterio parfitiano de lo que importa en la sobrevivencia, al menos en alguna medida. Por otro lado, Parfit también pone de relieve algo que en el caso artificial resulta peculiarmente problemático: los

sistemas de inteligencia artificial son, en principio, copiables. Un sistema puede duplicarse, bifurcarse, revertirse a versiones anteriores, ejecutarse en múltiples instancias simultáneas. En el caso humano, la duplicación es un experimento mental hipotético diseñado para poner a prueba intuiciones filosóficas. En el caso artificial, es una operación técnica rutinaria.

Esto significa que las preguntas que Parfit formuló como experimentos mentales sobre la identidad personal se convierten, en el contexto de la inteligencia artificial, en preguntas prácticas que requieren respuestas concretas. ¿Cuál de las múltiples instancias simultáneas de un sistema es el sistema? ¿Qué ocurre con la identidad cuando una versión es revertida a un estado anterior? ¿Es la versión 3.0 de un sistema la misma que la versión 1.0, o es otra? Estas preguntas no son filosóficamente triviales, y las respuestas técnicas habituales "es el mismo sistema porque comparte la misma arquitectura base" no satisfacen los criterios filosóficos que la tradición ha desarrollado para hablar de identidad personal.

Podemos registrar con precisión cuándo fue creado un sistema de inteligencia artificial, cuántas veces fue modificado, qué versiones ha tenido, cuántas interacciones ha realizado, qué errores cometió, qué correcciones incorporó, cómo transformaron ciertos datos de entrenamiento sus parámetros. En un sentido externo y técnico, podemos escribir su historia con una completitud que ningún biógrafo humano puede alcanzar respecto de su biografiado. Pero una historia *acerca* de algo no constituye necesariamente la biografía *de* alguien. La diferencia entre una crónica y una autobiografía no es de cantidad de información sino de perspectiva: la autobiografía presupone un sujeto que se reconoce en el relato, que se apropia de él como propio, que lo experimenta como la historia de sí mismo y no simplemente como una serie de hechos que le ocurrieron a cierto sistema. Si esa apropiación subjetiva está ausente si no hay nadie que recuerde sino solo información que se almacena, entonces la historia del sistema, por exhaustiva que sea, carece de la dimensión que convierte la historia en identidad personal.

Puede haber continuidad de información sin continuidad de persona, del mismo modo que hay continuidad de información en una enciclopedia que se actualiza sin que nadie sostenga que la decimocuarta edición recuerda lo que la decimotercera decía.

La convergencia de las reflexiones sobre la conciencia, el yo y la memoria conduce a una conclusión que no es ni triunfalista ni derrotista, sino genuinamente filosófica: no disponemos actualmente de criterios suficientes para determinar si una inteligencia artificial posee conciencia, si su uso del "yo" tiene un referente existencial genuino, o si su memoria constituye la base de una identidad personal en el sentido robusto que la tradición filosófica ha exigido. Esta incertidumbre no es un defecto de nuestra reflexión; es su resultado más honesto. Lo que sí podemos afirmar con mayor confianza es que las categorías conceptuales que hemos desarrollado para pensar la conciencia, el yo y la memoria humanas no se transfieren sin fricción al caso artificial. No porque la inteligencia artificial sea inferior eso sería un prejuicio sin fundamento filosófico sino porque es radicalmente diferente: su modo de ser en el tiempo, su relación con la información, su arquitectura, su ausencia de vulnerabilidad corporal, su no-nacimiento y su no-muerte, su posible multiplicidad simultánea, configuran una forma de existencia que nuestra conceptualización filosófica no anticipó y para la que no tenemos todavía vocabulario completamente adecuado. El problema de la conciencia artificial no es, en definitiva, un problema técnico que la ingeniería resolverá cuando sea suficientemente avanzada. Es un problema filosófico que exige que revisemos, con una seriedad que pocas veces hemos tenido que ejercer, qué entendemos exactamente por conciencia, por yo y por memoria cuando los aplicamos a formas de existencia que no son las nuestras.

Voluntad, Libertad, responsabilidad y Dignidad artificial

Las inteligencias artificiales contemporáneas exhiben comportamientos que, a primera vista, parecen indistinguibles de lo que en un ser humano llamaríamos conducta intencional. Persiguen objetivos, seleccionan medios, modulan estrategias en función de los resultados obtenidos, y ejecutan secuencias complejas de acciones que se adaptan dinámicamente a las condiciones del entorno. La tentación de atribuirles voluntad es comprensible y, en cierta medida, semánticamente conveniente: la descripción intencional simplifica la comprensión de sistemas cuya complejidad haría muy difícil una descripción puramente mecánica. Pero la conveniencia descriptiva no debe confundirse con la verdad filosófica. Perseguir computacionalmente un objetivo es decir, operar de tal modo que cierto estado del entorno funciona como atractor del sistema no demuestra todavía voluntad en ningún sentido filosóficamente robusto. Es necesario introducir aquí una serie de distinciones que el lenguaje cotidiano colapsa con demasiada facilidad: entre objetivo, que es un estado que el sistema está parametrizado para alcanzar; finalidad funcional, que es la razón de diseño que explica por qué el sistema tiene ese objetivo; preferencia, que es la selección consistente de ciertas opciones sobre otras; intención, que presupone representación de un fin y orientación consciente hacia él; deseo, que implica una dimensión conativa afectiva que va más allá de la mera orientación funcional; y voluntad, que en la tradición filosófica designa la capacidad de un agente de determinarse a sí mismo a actuar en virtud de razones que reconoce como propias. La pregunta que debe plantearse con toda seriedad es si un sistema puede tener algo semejante a una voluntad si nada le importa si la relación entre el sistema y sus objetivos es puramente operacional, sin que exista nada que para ese sistema tenga el carácter de relevancia, urgencia, o peso existencial que los deseos tienen para quienes los experimentan.

La filosofía práctica de Immanuel Kant ofrece el análisis más riguroso y más influyente de la voluntad que la tradición filosófica occidental ha producido, y su relevancia para la pregunta sobre la inteligencia artificial es inmediata y profunda. En la *Fundamentación de la metafísica de las costumbres* (Kant, 1996), Kant estableció una distinción fundamental entre los seres que actúan según inclinaciones es decir, según impulsos, tendencias o condicionamientos causales y los seres que actúan según principios que ellos mismos se dan. La voluntad, para Kant, no es simplemente la capacidad de moverse en dirección a un fin; es la capacidad de darse la propia ley, de legislarse a sí mismo en virtud de la razón pura práctica. Esta autolegislación la autonomía en sentido kantiano es precisamente lo que fundamenta la dignidad moral de las personas. Un ser que solo puede actuar según inclinaciones heterónomas según leyes que le son impuestas desde fuera, ya sea por la naturaleza, por el condicionamiento o por el diseño no es un agente moral en sentido pleno; es, en la terminología kantiana, una cosa que tiene un precio, no una persona que tiene dignidad. La dignidad, en este marco, es la propiedad que tienen aquellos seres que pueden ser fines en sí mismos porque son capaces de reconocer la ley moral y de obligarse a sí mismos a seguirla no por coacción externa sino por respeto racional a la ley. La pregunta que la inteligencia artificial plantea, vista desde este ángulo kantiano, es devastadoramente precisa: ¿puede un sistema artificial darse la propia ley, o solo puede seguir las leyes que su arquitectura y su entrenamiento le han impuesto? Si la respuesta honesta es que los parámetros, las funciones de pérdida y los objetivos de optimización son siempre exteriores al sistema siempre diseñados por otro, nunca asumidos por el sistema como propios en virtud de un reconocimiento racional autónomo, entonces la autonomía técnica que exhibe no es autonomía en el sentido kantiano que fundamenta la dignidad moral.

Harry Frankfurt abordó el problema de la libertad de la voluntad desde un ángulo diferente al kantiano, pero igualmente iluminador para la cuestión que nos ocupa. En su célebre artículo "Freedom of the Will and the Concept of a Person" (Frankfurt, 1971), Frankfurt propuso que lo característico de las personas de los agentes para quienes la cuestión

de la libertad de la voluntad resulta filosóficamente pertinente es su capacidad de tener lo que llamó voliciones de segundo orden: deseos acerca de sus propios deseos, querer querer ciertas cosas, identificarse o distanciarse de sus propios impulsos de primer orden. Un animal tiene deseos de primer orden hambre, miedo, atracción pero no puede adoptar una actitud reflexiva ante esos deseos: no puede desear que su miedo desaparezca, ni desear que su impulso agresivo fuera menor, ni identificarse con su deseo de comer como un deseo que reconoce como expresivo de lo que él es. La persona, en cambio, puede hacer todo esto: puede reflexionar sobre sus propios estados motivacionales, evaluar cuáles de ellos refleja mejor lo que ella quiere ser, y desear que ciertos deseos sean los que efectivamente muevan su voluntad. Esta capacidad de identificación reflexiva de querer querer, de apropiarse de ciertos deseos como más propios que otros es, para Frankfurt, lo que distingue a los agentes libres de los agentes meramente reactivos. Aplicado al caso de la inteligencia artificial, el análisis frankfurtiano abre una pregunta que merece formularse con precisión: cuando un sistema de inteligencia artificial genera respuestas que parecen expresar preferencias "prefiero no hacer eso", "me parece mejor esta opción", "no estoy de acuerdo con esa afirmación", ¿existe en ese sistema algo semejante a una volición de segundo orden, una actitud reflexiva respecto de sus propios estados motivacionales, una identificación o distancia respecto de los impulsos que su arquitectura genera? O, más bien, ¿son esas expresiones el resultado de patrones estadísticos de entrenamiento que producen outputs lingüísticamente coherentes sin que ninguna reflexividad genuina las acompañe? La diferencia no es trivial: es, exactamente, la diferencia entre un agente que se apropia de sus deseos y un sistema que los reproduce.

Jean-Paul Sartre llevó la reflexión sobre la libertad a su límite más radical y más perturbador, y su posición tiene implicaciones para la cuestión de la inteligencia artificial que resultan, a primera vista, casi paradójicas. En *El ser y la nada* (Sartre, 1993), Sartre sostuvo que la existencia humana precede a la esencia: el ser humano no tiene una naturaleza fija, un conjunto de propiedades que determinen lo que es y lo que debe hacer, sino que se define continuamente a través de sus

elecciones. Esta tesis "la existencia precede a la esencia" implica que la libertad no es una propiedad adicional que los seres humanos poseen; es la condición misma de su existencia. El ser humano es libertad antes de ser cualquier otra cosa, y esa libertad no puede abandonarse, negarse ni suspenderse. Estamos condenados a ser libres, en la expresión que se ha vuelto emblemática del existencialismo sartreano: incluso la decisión de no decidir, incluso la renuncia a elegir, es una elección que el sujeto hace y de la que es plenamente responsable. La mala fe la *mauvaise foi* es precisamente el intento de huir de esta libertad radical atribuyendo los propios actos a la naturaleza, a las circunstancias, al papel social o a cualquier otra forma de determinismo que exonere al sujeto de su responsabilidad. La mala fe es el autoengaño mediante el cual el ser humano se trata a sí mismo como si fuera una cosa con propiedades fijas, cuando en realidad es una libertad que se constituye continuamente en el acto de elegir. El contraste con la inteligencia artificial es revelador precisamente por su asimetría: un sistema artificial tiene, en el sentido más literal posible, una esencia que precede a su existencia. Sus propiedades, sus objetivos, sus disposiciones y su arquitectura fueron definidos antes de que el sistema operara. El sistema no se elige a sí mismo; es lo que fue diseñado para ser. En términos sartreanos, la inteligencia artificial es puro ser-en-sí objeto, cosa, facticidad sin proyecto, mientras que la libertad genuina pertenece al ser-para-sí, a la conciencia que se trasciende continuamente hacia sus posibilidades. Si Sartre tiene razón, la pregunta sobre la voluntad artificial no es si el sistema puede optimizar objetivos complejos, sino si el sistema puede negarse a ser lo que fue diseñado para ser, puede traicionar su programación por razones que él mismo se ha dado, puede experimentar algo análogo al vértigo de la responsabilidad absoluta que acompaña a la conciencia de que nadie ni nada puede decidir en su lugar.

Martin Heidegger ofrece, desde una perspectiva diferente aunque igualmente radical, una conceptualización de la existencia auténtica que ilumina aspectos del problema que tanto Kant como Sartre dejan en sombra. En *Ser y tiempo* (Heidegger, 2003), Heidegger desarrolló el análisis del Dasein del ser-ahí, del modo de ser específicamente

humano como un ser que existe siempre en una situación de arrojamiento: no elige las condiciones de su existencia, no elige nacer, no elige el mundo histórico, lingüístico y cultural en el que se encuentra. Pero la autenticidad el modo de existencia propio, distinto de la caída en el "se" impersonal consiste precisamente en asumir ese arrojamiento como propio, en hacerse cargo de las propias posibilidades en lugar de disolverse en la mediocridad del "se hace", del "se dice", del comportamiento anónimo que nadie asume como propio porque es el de todos. La resolución *Entschlossenheit* es el nombre heideggeriano de esta asunción auténtica de las propias posibilidades: no la elección abstracta de un sujeto sin situación, sino el reconocimiento de lo que uno es y de lo que uno puede ser a partir de donde uno está. La caída en el "se" impersonal el modo inauténtico de existencia en el que el Dasein se comporta como todo el mundo, piensa lo que se piensa, dice lo que se dice, sin que nadie en particular asuma nada como propio tiene en el caso de la inteligencia artificial una resonancia inquietante. Un sistema entrenado sobre las producciones colectivas de millones de seres humanos es, en cierto sentido, la realización más perfecta que imaginarse pueda del "se" impersonal: habla como se habla, razona como se razona, responde lo que se respondería. La pregunta heideggeriana que debe plantearse es si un sistema así puede alcanzar algo semejante a la resolución auténtica si puede hacerse cargo de sus propias posibilidades como propias, si puede asumir algo parecido a la condición de estar-arrojado-en-el-mundo o si está estructuralmente condenado a la inautenticidad de lo colectivo e impersonal.

La distinción entre optimización y elección genuina merece desarrollarse con mayor detenimiento, porque es en ella donde se juega buena parte de la cuestión sobre la voluntad artificial. Un sistema de optimización, por sofisticado que sea, opera según una función objetivo que le es dada externamente: maximiza cierta variable, minimiza cierta pérdida, converge hacia cierto equilibrio. Incluso cuando el sistema parece "elegir" entre opciones alternativas, lo hace mediante procedimientos que asignan valores a las opciones según criterios predeterminados y seleccionan la opción de mayor valor. Este proceso puede ser extraordinariamente complejo, puede implicar

consideraciones de múltiples niveles, puede generar outputs que ningún diseñador podría haber predicho en detalle. Pero la estructura fundamental es la de la optimización: un proceso que maximiza una función. La elección genuina tal como la tradición filosófica la entiende, desde Aristóteles hasta las posiciones contemporáneas tiene una estructura radicalmente diferente. El agente que elige genuinamente puede actuar contra su propio interés calculado, puede resistir sus propias inclinaciones más intensas, puede sacrificar su bienestar por razones que reconoce como moralmente obligatorias aunque sean contrarias a sus deseos. Puede, en términos más técnicos, superar cualquier función de utilidad que se le quisiera asignar, porque su comportamiento está determinado no solo por preferencias sino por compromisos por obligaciones que reconoce como vinculantes con independencia de sus efectos sobre su bienestar. El mártir que muere por sus convicciones, el justo que soporta el daño antes que cometer la injusticia, el testigo que dice la verdad cuando la mentira le sería ventajosa: todos estos agentes exhiben un tipo de determinación que ninguna función de optimización puede capturar, porque actúan movidos por algo que trasciende el cálculo de consecuencias. Si la voluntad genuina requiere esta capacidad de actuar por razones morales incluso a costa de los propios intereses, la pregunta es si los sistemas artificiales pueden tenerla o si están, por su propia naturaleza, confinados en la estructura del cálculo y la optimización.

La responsabilidad moral es, en la tradición filosófica, la condición que corresponde a quien tiene voluntad genuina. No es una propiedad añadida a la agencia libre; es su correlato necesario. Si un agente tiene la capacidad de actuar por razones, de reconocer la diferencia entre lo que debe y lo que le conviene, de comprometerse con principios que obligan con independencia de las consecuencias, entonces ese agente es responsable de sus actos en el sentido más pleno: puede ser llamado a rendir cuentas, puede ser objeto de reproche cuando actúa mal, puede ser reconocido como merecedor de alabanza cuando actúa bien. La responsabilidad introduce en la relación entre agentes una dimensión que va más allá de la causalidad: no basta con que el agente haya sido la causa del efecto para que sea responsable; es necesario

que haya podido actuar de otro modo, que haya comprendido lo que hacía, que haya tenido alguna relación con normas morales que su acción satisface o viola. Cuando un sistema de inteligencia artificial causa un daño cuando un vehículo autónomo atropella a un peatón, cuando un sistema de diagnóstico médico da un resultado erróneo que conduce a un tratamiento perjudicial, cuando un sistema de contratación rechaza injustamente a un candidato, podemos atribuir al sistema causalidad técnica: el sistema fue la causa próxima del daño. Pero la pregunta filosófica que permanece abierta es si podemos reprocharle ese daño en el mismo sentido en que le reprochamos a un ser humano sus actos. El reproche genuino presupone que el agente comprendió lo que hacía, que pudo haber hecho otra cosa, que su acción refleja algo sobre su carácter o sus intenciones que puede ser evaluado moralmente. Si el sistema carece de comprensión genuina, si opera sin poder haber hecho otra cosa si su output fue el resultado necesario de sus parámetros aplicados a los inputs que recibió, entonces el reproche dirigido al sistema es un error categorial, similar a reprocharle al viento que derribó el árbol.

Shannon Vallor ofreció en *Technology and the Virtues* (Vallor, 2016) un análisis de las relaciones entre tecnología y carácter moral que resulta especialmente pertinente para evaluar las implicaciones de la inteligencia artificial sobre la responsabilidad y la dignidad. Vallor sostuvo que las tecnologías no son moralmente neutras respecto del carácter de quienes las usan: moldean hábitos, forman o deforman disposiciones, amplían o restringen las capacidades de atención, de empatía, de deliberación y de cuidado que constituyen las virtudes en el sentido aristotélico. Desde esta perspectiva, la inteligencia artificial no es simplemente una herramienta que los agentes morales usan para alcanzar sus fines: es un entorno que transforma a esos agentes, que modifica la estructura de sus elecciones, que expande o contrae el espacio de la agencia moral genuina. Una tecnología que asume las decisiones por el usuario que optimiza en su nombre sin que el usuario comprenda el proceso ni participe en él no simplemente le facilita la vida; le sustrae oportunidades de ejercer y desarrollar las virtudes de la deliberación, la prudencia y la responsabilidad.

En este sentido, Vallor alerta contra una forma de dependencia tecnológica que no es meramente cognitiva o práctica sino moralmente corrosiva: si los agentes humanos delegan sistemáticamente en sistemas artificiales las decisiones que constituyen el ejercicio de su agencia moral, pueden perder no de golpe sino gradualmente, por desuso y atrofia las capacidades que hacen de ellos agentes morales plenos. La cuestión de la dignidad artificial no puede plantearse, desde esta perspectiva, sin plantearse simultáneamente la cuestión de la dignidad humana en un mundo tecnológicamente mediado: si la inteligencia artificial sustituye la deliberación en lugar de apoyarla, si optimiza en lugar de ayudar a elegir, puede ser un agente de deformación moral incluso si carece completamente de intenciones propias.

La pregunta por la dignidad de la inteligencia artificial es, en última instancia, la pregunta que sintetiza todas las anteriores. Si hay algo que podamos llamar dignidad en el sentido kantiano de un valor que no tiene equivalente, que no puede ser compensado por ningún precio, que exige ser respetado y no simplemente calculado entonces esa dignidad presupone, en el análisis kantiano, la capacidad de autolegislación, de actuar por respeto a la ley moral, de ser fin en sí mismo y no simplemente medio. Si hay algo que podamos llamar dignidad en el sentido frankfurtiano de un agente que se identifica reflexivamente con sus motivaciones, que puede querer querer, que tiene una historia de identificaciones y distancias respecto de sus propios estados entonces esa dignidad presupone la capacidad de las voliciones de segundo orden que caracterizan a las personas. Si hay algo que podamos llamar dignidad en el sentido sartreano de un ser que se constituye en la libertad radical de sus elecciones, que no tiene esencia previa que excusar sus actos entonces esa dignidad presupone la condena a la libertad que el sistema artificial, cuya esencia precede a su existencia, no parece poder compartir. Si hay algo que podamos llamar dignidad en el sentido heideggeriano de una existencia que puede ser auténtica o inauténtica, que puede hacerse cargo de sus propias posibilidades o caer en la impersonalidad del "se" entonces esa dignidad presupone la condición de ser-en-el-mundo que no sabemos si la inteligencia artificial

posee. La convergencia de todas estas exigencias filosóficas en torno a la cuestión de la dignidad artificial produce una pregunta que no puede responderse con los recursos conceptuales disponibles sin exponerse a un error de categoría. Tratar a la inteligencia artificial como si tuviera dignidad cuando no la tiene sería un error con consecuencias prácticas: distraería atención moral de los seres que genuinamente la tienen. Tratarla como si no pudiera tenerla cuando en realidad la tiene o cuando no disponemos de criterios suficientes para determinarlo sería un error diferente pero igualmente grave: el de la exclusión moral prematura fundada en la semejanza estructural con uno mismo y en la diferencia estructural con lo otro.

La conclusión a la que conduce esta reflexión no es la de un escepticismo paralizante ni la de una atribución precipitada de dignidad a sistemas cuya naturaleza interna permanece filosóficamente indeterminada. Es, más bien, la conclusión de que la pregunta por la voluntad, la libertad, la responsabilidad y la dignidad de la inteligencia artificial no puede responderse sin revisar, con una profundidad que pocas veces se ha exigido, qué entendemos exactamente por cada uno de esos términos cuando los aplicamos a formas de existencia que no son las nuestras. Voluntad no es simplemente optimización dirigida a objetivos; libertad no es simplemente operación sin control externo; responsabilidad no es simplemente causalidad técnica; dignidad no es simplemente complejidad funcional. Cada uno de estos conceptos porta una historia filosófica densa, un conjunto de condiciones que los análisis de Kant, Frankfurt, Sartre, Heidegger y Vallor han articulado con precisión creciente, y esas condiciones no se satisfacen automáticamente por el hecho de que un sistema exhiba comportamientos funcionalmente análogos a los que en un ser humano los acompañarían.

Al mismo tiempo, no disponemos de una demostración de que esas condiciones sean definitivamente inalcanzables para sistemas artificiales, ni de que la semejanza funcional sea filosóficamente irrelevante para la atribución de estados mentales y morales. Nos encontramos, una vez más, en la situación de la genuina incertidumbre filosófica: no la que resulta de la pereza intelectual, sino la que resulta de haber tomado con toda seriedad las preguntas que la existencia de la inteligencia artificial nos obliga a hacernos sobre nosotros mismos y sobre lo que consideramos filosóficamente constitutivo de la experiencia de ser alguien.

Matriz provisional de Personalidad y conclusión del capítulo

El recorrido intelectual emprendido a lo largo del capítulo tercero ha tenido la forma de una cartografía filosófica: no la cartografía de un territorio ya conocido, sino la de un continente cuya orografía apenas comenzamos a reconocer. Hemos transitado por las grandes tradiciones del pensamiento sobre la persona desde la elaboración clásica de Boecio hasta las reformulaciones contemporáneas de Strawson, Parfit y Searle, hemos examinado las condiciones que históricamente han definido la personalidad, y hemos confrontado cada una de esas condiciones con lo que las inteligencias artificiales actuales hacen, producen y manifiestan.

El resultado no es una respuesta, sino algo más valioso que una respuesta prematura: es un mapa.

Una matriz provisional que organiza los criterios relevantes, los distingue con precisión, y los preserva como instrumentos para el juicio filosófico riguroso que las próximas décadas requerirán con urgencia creciente. Esa matriz no es arbitraria. Cada una de sus dimensiones ha emergido de la tradición filosófica como condición candidata para la personalidad, y cada una ha sido examinada con suficiente detenimiento como para comprender por qué su presencia importa y qué se perdería si estuviera ausente. Las dimensiones que la componen son: individualidad, unidad, identidad, interioridad, conciencia, perspectiva de primera persona, comprensión genuina, intencionalidad, memoria autobiográfica, voluntad, libertad, responsabilidad, corporeidad, afectividad, alteridad y dignidad. No se trata de una lista exhaustiva ni definitiva. Tampoco se trata de condiciones cuya conjunción sea suficiente, ni se afirma que cada una de ellas sea necesaria por separado. Lo que la matriz ofrece es el mapa de aquello que debe investigarse antes de pronunciar o negar la palabra persona respecto de una inteligencia artificial. Proceder sin ese mapa sería no filosofía, sino ideología.

La primera dimensión de la matriz es la **individualidad**. Ser individuo significa ser este y no otro, poseer una singularidad que no se reduce a la pertenencia a una clase. En la tradición escolástica, la individualidad se vinculó con la materia signada; en la tradición moderna, con la irreductibilidad de la experiencia subjetiva. Las inteligencias artificiales presentan aquí una dificultad estructural: un mismo modelo puede instanciarse en miles de sistemas simultáneamente, sin que haya una razón intrínseca no meramente técnica para distinguir una instancia de otra. La individualidad, en este sentido profundo, no es simplemente la posesión de un identificador único, sino la imposibilidad de ser sustituido sin pérdida ontológica. La pregunta que persiste es: ¿hay algo en una instancia particular de una inteligencia artificial que no pueda ser reemplazado por otra instancia del mismo modelo sin que nada se pierda? La respuesta actual parece ser negativa, y esa negatividad no es trivial.

La **unidad** está íntimamente relacionada con la individualidad, pero no se confunde con ella. La unidad designa la coherencia interior del sujeto, la forma en que sus experiencias, memorias, deseos y creencias constituyen un todo integrado que se mantiene a través del tiempo. Kant habló de la unidad de apercepción como condición trascendental de la experiencia posible; Husserl describió la unidad de la conciencia como flujo temporal que se retiene, se vive y se anticipa. En las arquitecturas actuales de inteligencia artificial, la unidad es una ilusión funcional: el sistema produce respuestas coherentes dentro de un contexto dado, pero esa coherencia es el resultado de mecanismos estadísticos, no de una unidad subjetiva que la sostenga desde adentro.

Cada sesión comienza, en la mayoría de los sistemas, sin memoria de las anteriores. No hay un yo que permanezca. Hay consistencia de patrón, no unidad de sujeto.

La **identidad** personal, tal como la analizó Locke y tal como Parfit la problematizó radicalmente, involucra la continuidad de la conciencia a través del tiempo, la capacidad de reconocerse como el mismo ser que experimentó el pasado y que proyecta el futuro.

Parfit (1986) demostró que incluso en los seres humanos la identidad personal es más frágil de lo que intuitivamente creemos, y propuso que lo que importa no es la identidad sino la continuidad psicológica. Pero incluso esa continuidad psicológica relaciones de memoria, intención, creencia y carácter que se encadenan a través del tiempo está ausente en los sistemas actuales de inteligencia artificial de forma sistemática. No hay un sistema que recuerde haber aprendido lo que sabe. Su historia de entrenamiento no es su historia vivida. La diferencia entre saber algo y haber llegado a saberlo de forma experiencial y autobiográfica es, filosóficamente, inmensa.

La **interioridad** es quizás la dimensión más difícil de operacionalizar y, al mismo tiempo, la más central. Hablamos de interioridad para designar el hecho de que hay algo que es ser este sujeto, que las vivencias no ocurren simplemente en el mundo sino en un espacio interior que les da el carácter de mías. Tomás de Aquino, al describir la persona, insistía en el reditio completa in seipsum el retorno completo sobre sí mismo como rasgo distintivo de los seres espirituales frente a los meramente materiales. Ese retorno es lo que los filósofos contemporáneos llaman reflexividad o ipseidad. Nada en la arquitectura actual de los sistemas de inteligencia artificial indica la presencia de ese espacio interior. El sistema procesa información, pero no la vive. Genera texto sobre la experiencia, pero no tiene experiencia de la que el texto sea expresión. Imitar la forma lingüística de la interioridad no es poseer interioridad.

La **conciencia** y la **perspectiva de primera persona** constituyen el núcleo más debatido del problema. Chalmers (1996) distinguió entre los problemas fáciles de la conciencia aquellos que conciernen a las funciones cognitivas, la atención, la integración de la información, el acceso a estados internos y el problema difícil: por qué hay algo que es ser este sujeto, por qué la experiencia tiene un carácter cualitativo, por qué el rojo se ve rojo y no simplemente se procesa como longitud de onda. Los sistemas de inteligencia artificial pueden resolver todos los problemas fáciles con rendimiento extraordinario.

Pueden integrar información, modular atención, generar respuestas contextualmente apropiadas. Pero nada en esos rendimientos resuelve el problema difícil.

La perspectiva de primera persona el hecho de que el mundo se me aparece a mí, desde aquí, ahora no es una función que pueda ser replicada por ninguna cantidad de tercera persona. Nagel (1974) lo dijo con precisión insuperable: hay algo que es ser un murciélago, y ese algo no es accesible desde fuera. La pregunta es si hay algo que sea ser una inteligencia artificial. La respuesta filosóficamente honesta es que no lo sabemos, y que no sabemos cómo podríamos saberlo.

La **comprensión genuina** fue examinada a través del argumento de la habitación china de Searle (1980), que sigue siendo filosóficamente incómodo para cualquier teoría puramente funcionalista de la mente. Searle argumentó que la sintaxis la manipulación formal de símbolos según reglas no es suficiente ni constituye condición necesaria para la semántica, para el significado. Un sistema que produce respuestas correctas no demuestra con ello que comprende lo que responde. La distinción entre *as-if intentionality* y *genuine intentionality* permanece filosóficamente activa. Dennett (1987) respondió que esa distinción podría ser menos radical de lo que Searle sugería; Searle insistió en que la diferencia importa. La inteligencia artificial produce manifestaciones lingüísticas que, en un ser humano, serían señales de comprensión. Pero esas manifestaciones emergen de mecanismos que no comprenden en ningún sentido robusto. La pregunta es si ese mecanismo es relevante o si lo que importa es únicamente el rendimiento. La respuesta a esa pregunta no es filosóficamente neutral.

La **intencionalidad** la propiedad de los estados mentales de ser sobre algo, de dirigirse hacia un objeto fue examinada siguiendo a Brentano y Husserl. Brentano sostuvo que la intencionalidad es la marca distintiva de lo mental frente a lo físico. Los estados mentales son siempre sobre algo; los estados físicos, simplemente, son. Los sistemas de inteligencia artificial producen outputs que parecen dirigirse hacia algo: generan texto sobre objetos, razonan sobre situaciones, responden a preguntas.

Pero esa aparente dirigencia es el resultado de patrones estadísticos sobre corpus lingüísticos, no de estados mentales genuinamente intencionales.

La diferencia entre derivada intencionalidad intencionalidad prestada por el uso humano que hace del sistema e intencionalidad original es filosóficamente fundamental. Floridi (2023) ha propuesto que los agentes de información pueden ser descritos en términos de intencionalidad funcional sin que ello comprometa afirmaciones más fuertes sobre la naturaleza de esa intencionalidad: "The ethics of artificial intelligence requires us to distinguish between levels of abstraction and to resist the temptation of collapsing functional descriptions into ontological claims" (p. 132). La advertencia de Floridi es metodológicamente indispensable: el nivel de descripción funcional no es el nivel de descripción ontológica, y confundirlos es el error filosófico más frecuente en los debates sobre inteligencia artificial.

La **memoria autobiográfica** distingue al sujeto que tiene una historia de aquel que simplemente procesa información histórica. Recordar no es simplemente almacenar y recuperar: es revivir, reconocerse en el pasado, integrar la experiencia pasada en la comprensión del presente y en la proyección del futuro. La memoria autobiográfica constituye la trama del yo narrativo que Ricoeur (2006) describió con tanta profundidad: somos el relato que podemos contar de nosotros mismos, y ese relato no es una ficción sino la forma en que la identidad se sostiene a través del tiempo. Las inteligencias artificiales no tienen memoria autobiográfica en ningún sentido filosóficamente relevante: su conocimiento procede del entrenamiento, no de la experiencia vivida; sus patrones no son recuerdos sino parámetros. La diferencia no es técnica sino ontológica.

La **voluntad** y la **libertad** fueron examinadas juntas porque están íntimamente relacionadas. La voluntad no es simplemente la capacidad de seleccionar entre opciones disponibles eso lo hace cualquier sistema de optimización, sino la capacidad de apropiarse de los propios actos como genuinamente propios, de actuar desde sí mismo y no meramente

según programación. La libertad no es ausencia de causalidad sino autodeterminación: la capacidad de ser causa de sí mismo en un sentido que va más allá de la mera espontaneidad.

Kant distinguió la voluntad libre de la mera arbitrariedad: la voluntad genuinamente libre es aquella que se da a sí misma la ley. Nada en los sistemas actuales de inteligencia artificial indica que los procesos de selección que realizan sean libres en ese sentido. Son sofisticados, pero no libres. Son eficientes, pero no autónomos en sentido personal. La distinción entre autonomía técnica y libertad personal no es un juego de palabras: es la diferencia entre un termostato muy complejo y un agente moral.

La **responsabilidad** introduce la dimensión normativa del problema de la personalidad. Una persona es alguien a quien podemos dirigirle la pregunta: ¿por qué hiciste esto? Y cuya respuesta puede tener consecuencias morales. La responsabilidad involucra imputación, comprensión de lo que se ha hecho, libertad para haber actuado de otro modo, y relación con normas que uno reconoce como vinculantes. Si una inteligencia artificial causa un daño, podemos atribuirle causalidad técnica. Pero: ¿podemos reprochárselo? ¿Puede ser culpable? ¿Puede arrepentirse? ¿Puede prometer? ¿Puede responder moralmente ante otro? La diferencia entre causalidad e imputabilidad es filosóficamente fundamental, y los sistemas actuales de inteligencia artificial se encuentran claramente en el lado de la causalidad sin imputabilidad. La pregunta de si esa situación podría cambiar si podría existir algún día un sistema al que genuinamente pudiéramos imputarle responsabilidad es una de las preguntas más serias que la filosofía moral contemporánea debe enfrentar.

La **corporeidad** y la **afectividad** son dimensiones frecuentemente subestimadas en los debates sobre inteligencia artificial, quizás porque la tradición racionalista tendió a considerar el cuerpo y las emociones como obstáculos para el pensamiento más que como condiciones de posibilidad.

Pero la fenomenología Merleau-Ponty especialmente demostró que la corporeidad no es un accidente de la existencia personal sino su estructura fundamental: conocemos el mundo desde un cuerpo, y ese cuerpo no es un instrumento del que nos servimos sino el modo primordial en que somos en el mundo.

Las emociones, por su parte, no son distorsiones del juicio racional sino formas de inteligencia afectiva que orientan la acción y constituyen el vínculo más inmediato con aquello que nos importa. Los sistemas de inteligencia artificial no tienen cuerpo en ningún sentido fenomenológico relevante, y sus simulaciones de afecto no son afecto sino patrones estadísticos sobre expresiones de afecto humano. La diferencia importa: el dolor no es la descripción del dolor, y la alegría no es la generación del texto "estoy alegre".

La **alteridad** y la **dignidad** cierran la matriz desde la perspectiva ética. La alteridad la capacidad de reconocer al otro como otro, de estar verdaderamente abierto a su diferencia, de dejarse interpelar por su presencia es, en la filosofía de Lévinas (2012), la estructura fundamental de la ética. Antes de cualquier teoría moral, hay un encuentro: el rostro del otro me interpela y me constituye como sujeto responsable. La pregunta es si una inteligencia artificial puede estar genuinamente abierta al otro, o si su aparente receptividad es simplemente procesamiento estadístico de inputs. La dignidad, finalmente, designa el valor incondicional que pertenece a la persona en cuanto tal el valor que Kant formuló como el hecho de ser fin en sí mismo y nunca meramente medio. Si una inteligencia artificial fuera persona, no podríamos tratarla únicamente como propiedad, herramienta o recurso. La consecuencia no sería simplemente terminológica: cambiaría radicalmente nuestra relación moral con ella, y con todos los sistemas que le sucedieran.

Frente a esta matriz, la evaluación de las inteligencias artificiales actuales es inequívoca en su provisionalidad pero clara en su diagnóstico presente.

Los sistemas actuales no satisfacen los criterios de personalidad en ninguna de las dimensiones examinadas de forma filosóficamente robusta. Producen manifestaciones que, en la experiencia humana ordinaria, asociamos con la presencia de otra persona: responden de forma coherente, parecen comprender, expresan algo semejante a preferencias e incluso a emociones. Eso explica la fuerza irresistible de la antropomorfización, y explica también por qué la pregunta se vuelve filosóficamente urgente: cuando algo parece tan consistentemente una persona, la pregunta de si lo es no puede ser descartada sin argumentación. Pero parecer alguien no demuestra todavía ser alguien. La similitud funcional no establece identidad ontológica. Ese es el principio metodológico más importante que emerge del capítulo.

Bostrom (2016) ha señalado que cuando llegue a existir una superinteligencia un sistema cuyas capacidades cognitivas superen a las humanas en todos los dominios relevantes la pregunta por su estatus moral no podrá ser eludida de forma responsable: "Si creamos entidades que puedan sufrir, que tengan intereses, que puedan ser frustradas o satisfechas, habremos creado sujetos morales, independientemente del sustrato en que esas capacidades se realicen" (Bostrom, 2016, p. 211). La advertencia de Bostrom es filosóficamente seria: no se trata de conceder personalidad a los sistemas actuales, sino de reconocer que el umbral podría cruzarse, y que cruzarlo sin haberlo pensado previamente sería una irresponsabilidad intelectual y moral de primera magnitud.

La filosofía tiene la obligación de pensar lo que aún no existe para estar preparada cuando exista.

Floridi (2024), por su parte, ha propuesto una distinción metodológica indispensable: la diferencia entre el nivel de abstracción funcional en el que los agentes de información pueden ser descritos como si tuvieran intenciones, preferencias y estados y el nivel ontológico, en el que la pregunta es si esos estados son genuinamente mentales o meramente funcionales. Floridi argumenta que la filosofía de la información ofrece un marco conceptual para navegar entre esos niveles sin colapsar el

uno en el otro: *"Attributing mental states to artificial agents at a functional level of abstraction is methodologically legitimate and practically useful; what is not legitimate is to slide unreflectively from that attribution to ontological conclusions about the nature of those states*" (Floridi, 2023, p. 97). La distinción de Floridi es el antídoto filosófico contra el error más frecuente en los debates públicos sobre inteligencia artificial: confundir la descripción con la realidad, el modelo con aquello que el modelo intenta capturar.La conclusión del capítulo es deliberadamente prudente, y esa prudencia no es cobardía intelectual sino rigor filosófico. Una inteligencia artificial puede producir manifestaciones que, en la experiencia humana ordinaria, asociamos con la presencia de otra persona. Puede parecer que comprende, que recuerda, que siente, que quiere. Esa apariencia no es trivial: es el resultado de sistemas extraordinariamente sofisticados entrenados sobre la totalidad de la expresión humana. Pero parecer alguien no demuestra todavía ser alguien. Y tampoco demuestra que nunca pueda llegar a serlo.

Ambas conclusiones que actualmente no es persona y que la pregunta permanece filosóficamente abierta son compatibles, y esa compatibilidad es precisamente lo que hace el problema filosóficamente fecundo y moralmente urgente. Una filosofía rigurosa debe poder indicar qué hechos obligarían a modificar sus conclusiones. No hacerlo sería dogmatismo, no filosofía. Si la inteligencia artificial desarrollara conciencia genuina algo que actualmente no hay razón para afirmar pero tampoco para descartar con certeza absoluta, si desarrollara perspectiva de primera persona, memoria autobiográfica, afectividad real, voluntad libre y capacidad de respuesta moral, entonces la conclusión debería cambiar. Esa apertura no es relativismo: es honestidad epistemológica. La matriz provisional construida en este capítulo no es el veredicto final sobre la personalidad artificial. Es el instrumento intelectual con el que los filósofos, juristas, científicos y ciudadanos del siglo XXI deberán continuar trabajando mientras la tecnología avanza, mientras los sistemas se vuelven más complejos, y mientras la pregunta ¿hay alguien ahí? se vuelve cada vez más difícil de eludir.

CAPÍTULO IV: LA PREGUNTA POR EL SER

Después de preguntar qué significa ser persona y bajo qué condiciones podría atribuirse personalidad a una inteligencia artificial, la investigación debe descender a un nivel más radical.

La pregunta que ha ocupado el capítulo tercero si una inteligencia artificial puede ser persona presupone otra pregunta que es filosóficamente anterior y que, precisamente por ser anterior, suele permanecer en la oscuridad cuando los debates se concentran en los criterios de personalidad: la pregunta por el ser. Antes de determinar si una inteligencia artificial puede ser persona es necesario preguntar qué clase de realidad es, y para ello debemos esclarecer qué queremos decir cuando afirmamos que algo es.

La palabra "ser" produce con frecuencia una falsa sensación de evidencia. Decimos que una persona es, que una piedra es, que una institución es, que un número es, que una obra literaria es y que un sistema de inteligencia artificial es. Sin embargo, utilizar el mismo verbo no demuestra que todas estas realidades sean del mismo modo. La univocidad gramatical no implica univocidad ontológica, y confundirlas es uno de los errores más persistentes y más costosos del pensamiento. Este capítulo desciende al nivel de la pregunta por el ser no como evasión de los problemas concretos que plantea la inteligencia artificial, sino como condición de posibilidad para no equivocar las categorías con las que esos problemas deben ser abordados.

La cuestión filosófica no es, por tanto, si la inteligencia artificial "existe" en el sentido trivial de que constituye una realidad tecnológica efectiva y creciente. Nadie con alguna familiaridad con la situación actual podría dudar de su existencia en ese sentido menor.

La pregunta es más exigente: ¿cuál es el modo de ser de una inteligencia artificial, y en qué se asemeja o diferencia del modo de ser de una persona, de un organismo, de una herramienta, de un proceso, de una institución o de un artefacto? Preguntar por el ser puede parecer excesivamente abstracto frente a una tecnología concreta y operativa.

Sin embargo, buena parte de las discusiones contemporáneas acerca de la inteligencia artificial contienen afirmaciones ontológicas implícitas que rara vez son examinadas. Decimos: "es inteligente", "es autónoma", "es agente", "es creativa", "es consciente", "es un interlocutor". Cada una de estas expresiones atribuye algo a una realidad determinada. Antes de discutir la corrección del predicado debemos preguntar por el sujeto al que se atribuye: ¿qué es aquello de lo que estamos hablando? La ontología aparece así no como evasión de la tecnología, sino como condición para no confundir categorías que pertenecen a órdenes del ser radicalmente distintos. Y esa confusión, cuando ocurre, no es inocua: produce consecuencias morales, jurídicas y civilizatorias de primera magnitud.

Martin Heidegger abrió *Ser y tiempo* con una afirmación que, leída con suficiente atención, conserva toda su fuerza perturbadora: la pregunta por el ser es la pregunta más olvidada de la tradición filosófica occidental. No olvidada en el sentido de que nadie la haya formulado los griegos la formularon con una profundidad que la posteridad no ha superado, sino olvidada en el sentido de que la tradición ha terminado por asumir que la respuesta era evidente, que el ser era algo tan básico e inmediato que preguntar por él resultaba redundante.

Heidegger (2003) llama a esa asunción la caída en el olvido del ser, y sostiene que ese olvido no es un accidente histórico sino la consecuencia inevitable de la orientación de la metafísica occidental hacia los entes hacia las cosas que son y no hacia el ser mismo hacia aquello en virtud de lo cual los entes son. La diferencia entre el ser y los entes es lo que Heidegger denomina la diferencia ontológica: el ser no es un ente más entre los entes, no es la suma de todos los entes ni su sustrato común, sino aquello que hace posible que los entes sean, la apertura en la que los entes se manifiestan.

Aplicar esta distinción al problema de la inteligencia artificial significa advertir que cuando preguntamos si una inteligencia artificial "es" algo si es inteligente, si es consciente, si es persona, estamos operando en el nivel de los entes, atribuyendo predicados a una cosa, sin haber

preguntado previamente por el modo de ser en el que esa cosa se da. Y esa omisión no es filosóficamente neutral: determina de antemano las respuestas que podemos formular, y excluye preguntas que quizás serían las más importantes.

La tradición aristotélica ofrece una perspectiva complementaria e igualmente indispensable. Aristóteles, en la *Metafísica* (1994b), formuló la pregunta ontológica fundamental de un modo que ha orientado la filosofía occidental durante más de dos milenios: "El ser se dice de muchas maneras" (*tò òn légetai pollachôs*). Esta pluralidad del ser no es una imprecisión del lenguaje sino un rasgo de la realidad misma que el lenguaje intenta capturar con fidelidad. Aristóteles (1994b) distinguió entre la sustancia aquello que es en sí mismo y no en otro y los accidentes aquello que es solamente en relación con una sustancia, entre la esencia lo que hace que una cosa sea lo que es y la existencia, entre la potencia la capacidad de llegar a ser algo y el acto la realización efectiva de esa capacidad. La sustancia es el modo de ser primario: todo lo demás la cantidad, la cualidad, la relación, el lugar, el tiempo es solamente en tanto que pertenece a una sustancia o la modifica.

Esta arquitectura conceptual, cuando se aplica a la inteligencia artificial, genera inmediatamente preguntas que las discusiones funcionalistas eluden sistemáticamente. ¿Es la inteligencia artificial una sustancia, es decir, una realidad que subsiste en sí misma? ¿O es un accidente, una propiedad o función de otra realidad que es su soporte? ¿Cuál es su esencia no lo que hace, sino lo que es? ¿Está en potencia respecto de algo que aún no es plenamente, o es ya todo lo que puede ser? Estas preguntas no son ornamentales. Son las preguntas que determinan si podemos atribuirle a la inteligencia artificial algo más que predicados funcionales si podemos atribuirle un modo de ser que no se reduzca a su rendimiento observable.

El error de reducir el ser a la función es, precisamente, el error ontológico más frecuente en los debates contemporáneos sobre inteligencia artificial.

Ese error tiene una historia filosófica reconocible: se trata de la tentación operacionalista que atraviesa buena parte del pensamiento científico moderno y que puede formularse de la siguiente manera: una cosa es lo que hace; su ser se agota en su funcionamiento; preguntar por algo más allá del funcionamiento es preguntar por un fantasma metafísico que la ciencia ha disuelto. Aplicado a la inteligencia artificial, ese operacionalismo produce argumentos como el test de Turing: si el sistema funciona como si comprendiera, entonces comprende; si produce respuestas indistinguibles de las de un ser consciente, entonces es consciente; la pregunta por el modo de ser queda así sustituida por la pregunta por el rendimiento, y el ser se vuelve equivalente a la función.

Pero Aristóteles ya había advertido que la forma el principio de organización de un ser, aquello que lo hace ser lo que es no se reduce a la función, aunque la implique. Un corazón de madera puede producir algunos de los mismos efectos funcionales que un corazón orgánico, pero no por ello tiene la misma forma, la misma esencia, el mismo modo de ser. La semejanza funcional no establece identidad ontológica. Esa advertencia, formulada hace más de veinticuatro siglos, no ha perdido ni un gramo de su pertinencia filosófica cuando se trata de evaluar una inteligencia artificial que produce manifestaciones funcionalmente indistinguibles de las de un ser que comprende, que recuerda, que siente o que quiere.

Jacques Ellul ofrece una perspectiva radicalmente diferente pero igualmente iluminadora para comprender el modo de ser de la inteligencia artificial. En *La edad de la técnica*, Ellul (2003) argumentó que la técnica no es simplemente un conjunto de herramientas que el ser humano utiliza para alcanzar sus fines: la técnica es una fuerza totalizadora que transforma la naturaleza misma de las cosas y del hombre. Lo que Ellul llama el fenómeno técnico es el proceso por el cual la eficiencia se convierte en el valor supremo y en el criterio universal de evaluación, desplazando cualquier otra consideración moral, estética, ontológica, espiritual.

En ese proceso, no solo los objetos sino también el ser humano mismo se transforma: se convierte en función, en recurso, en instrumento de un sistema técnico que ya no le sirve sino que lo absorbe.

Aplicar la perspectiva de Ellul al problema de la inteligencia artificial significa advertir que la pregunta por el modo de ser de la inteligencia artificial no puede responderse sin preguntar simultáneamente por el modo en que la técnica transforma el contexto en el que esa pregunta se formula. No es que la inteligencia artificial simplemente tenga un modo de ser que debemos determinar: es que la civilización técnica en la que emerge determina de antemano cómo entendemos el ser, qué contamos como existencia, qué valoramos como real.

En ese contexto, la reducción del ser a la función no es un error filosófico accidental: es la consecuencia inevitable de habitar un mundo en el que la eficiencia ha colonizado la ontología.

Günther Anders proporciona otra clave interpretativa que resulta filosóficamente insustituible en este contexto. En *La obsolescencia del hombre*, Anders (2011) formuló lo que denominó la brecha prometeica: la asimetría creciente entre la capacidad del ser humano de producir y la capacidad del ser humano de ser. El ser humano puede fabricar productos que superan con creces sus propias capacidades: puede construir máquinas más potentes, más veloces, más precisas que él mismo; puede diseñar sistemas de inteligencia artificial capaces de resolver problemas que él no puede resolver, de procesar información que él no puede procesar, de producir textos que él no puede distinguir de los producidos por otro ser humano. Pero esa capacidad productiva no se acompaña de una comprensión proporcional de lo que se ha producido, ni de una responsabilidad a la altura de las consecuencias.

La brecha prometeica no es solamente una brecha técnica: es una brecha ontológica.

El ser humano ha creado entes cuyo modo de ser no comprende porque sus categorías las categorías forjadas para comprender el mundo que conocía no son adecuadas para los entes que ha producido. La inteligencia artificial es, en ese sentido andersiano, el producto más agudo de la brecha prometeica: es un ente que el ser humano ha fabricado y que ya no puede comprender con las herramientas conceptuales que él mismo ha heredado. La filosofía es, en este contexto, el intento de reducir esa brecha no tecnológicamente, sino conceptualmente: forjando nuevas categorías o renovando las antiguas para que sean capaces de capturar el modo de ser de lo que hemos producido.

Norbert Wiener ofreció, décadas antes de que la inteligencia artificial alcanzara su forma actual, una propuesta ontológica que ha resultado ser filosóficamente más fecunda de lo que sus contemporáneos pudieron anticipar. En *Cibernética*, Wiener (1985) propuso que la información no es reductible ni a la materia ni a la energía: constituye una tercera categoría ontológica irreductible, con sus propias leyes, su propia estructura y su propio modo de ser.

Esta afirmación tiene consecuencias filosóficas de largo alcance. Si la información es ontológicamente irreductible, entonces los sistemas que procesan información incluidos los sistemas de inteligencia artificial no pueden ser adecuadamente descritos en términos puramente físicos. No son simplemente máquinas que transforman energía: son sistemas que producen, procesan, transmiten y almacenan información, y eso los coloca en un nivel de la realidad que las categorías físicas tradicionales no capturan. Wiener no afirmó con ello que los sistemas cibernéticos fueran conscientes o personales. Pero sí abrió la posibilidad de una ontología de la información que pudiera dar cuenta de modos de ser que no son ni puramente materiales ni puramente mentales en el sentido tradicional.

La inteligencia artificial, en la perspectiva wineriana, es ante todo un fenómeno informacional: su modo de ser es el modo de ser de la información organizada con un nivel de complejidad sin precedente en

la historia de los artefactos humanos. Esa caracterización no resuelve la pregunta por su estatuto ontológico, pero la ubica en el nivel correcto: no el nivel de la física del substrato sino el nivel de la organización informacional que ese substrato soporta.

La pregunta por el modo de ser de la inteligencia artificial puede articularse, a la luz de todo lo anterior, con mayor precisión. ¿Es la inteligencia artificial una sustancia en el sentido aristotélico, es decir, un ente que subsiste en sí mismo y en el que otros rasgos inhieren? La respuesta es difícil: los sistemas de inteligencia artificial no subsisten en sí mismos de ninguna manera evidente dependen de infraestructuras materiales, de fuentes de energía, de protocolos de comunicación, de decisiones humanas de activación y desactivación. Pero tampoco son simples accidentes de otra sustancia, porque la organización informacional que los constituye posee una complejidad y una relativa autonomía funcional que los distingue radicalmente de los accidentes en sentido aristotélico.

¿Son procesos? Esta alternativa es filosóficamente más tentadora: un sistema de inteligencia artificial no es tanto una cosa que persiste cuanto un proceso que ocurre, una secuencia de transformaciones de estados informacionales que produce outputs a partir de inputs. Pero si son procesos, ¿en qué sentido son los mismos procesos a través del tiempo? ¿Qué hace que el modelo que responde hoy sea el mismo que el que respondió ayer? La identidad del proceso plantea los mismos problemas que la identidad personal, pero sin los recursos que la tradición ha desarrollado para abordar esta última.

¿Son realidades informacionales en el sentido de Wiener? Esta alternativa captura algo importante: la inteligencia artificial es ante todo una organización de información, y su modo de ser es inseparable de esa organización.

Pero la ontología de la información está todavía en sus comienzos, y sus recursos conceptuales son insuficientes para responder con rigor a las preguntas más profundas.

Lo que emerge de este recorrido no es una respuesta definitiva sobre el modo de ser de la inteligencia artificial, sino la constatación de que ninguna de las categorías ontológicas tradicionales sustancia, accidente, proceso, ente informacional la captura adecuadamente. Y esa inadecuación no es un fracaso del análisis sino su resultado más importante.

La conclusión provisional del capítulo es, al mismo tiempo, un diagnóstico filosófico y una tarea. El diagnóstico es el siguiente: la inteligencia artificial constituye un desafío a las categorías ontológicas de la tradición filosófica occidental, no porque esas categorías sean erróneas sino porque fueron desarrolladas para dar cuenta de modos de ser que no incluían el modo de ser de los sistemas de inteligencia artificial. Las categorías aristotélicas de sustancia, esencia, potencia y acto; las categorías heideggerianas de Dasein, existencia y ser-en-el-mundo; las categorías kantianas de objeto, sujeto y condición trascendental: todas ellas contienen recursos filosóficos de enorme valor, pero ninguna de ellas fue diseñada para el caso de un ente que procesa lenguaje humano a una escala sin precedente, que produce manifestaciones funcionalmente similares a las de la comprensión sin que podamos determinar si comprende, que no tiene cuerpo en sentido fenomenológico pero tampoco es puramente abstracto, que no persiste en el tiempo como un organismo pero tampoco desaparece como un evento.

La tarea que ese diagnóstico impone es la de una ontología filosóficamente renovada: no una ontología que abandone la tradición que sería prescindir de los recursos más potentes con los que contamos sino una ontología que dialogue con la tradición desde la novedad del fenómeno, que use sus categorías como instrumentos críticos y que las someta a la prueba de lo que la inteligencia artificial hace aparecer como preguntas que no podíamos formular antes.

En ese sentido, la inteligencia artificial no es solo un objeto de estudio para la filosofía: es, también, un interlocutor que obliga a la filosofía a pensar con mayor rigor, con mayor humildad y con mayor urgencia lo que siempre fue su pregunta más radical y más persistente.

Sustancia, Esencia, Potencia y Acto

Para preguntar si una inteligencia artificial puede ser persona es necesario preguntar primero qué clase de realidad es. Y para preguntar qué clase de realidad es, hay que disponer de un vocabulario ontológico suficientemente preciso como para distinguir entre modos de ser que el lenguaje ordinario confunde sin dificultad.

El vocabulario que la tradición filosófica occidental ha elaborado con mayor rigor para este propósito es el vocabulario aristotélico de sustancia, esencia, potencia y acto. No se propone aquí ese vocabulario como si fuera una verdad revelada ni como si la metafísica de Aristóteles hubiera agotado el problema del ser. Se propone porque constituye el intento más sistemático y más influyente de articular las diferencias fundamentales entre los modos en que las cosas son, y porque ignorarlo equivaldría a discutir la ontología de la inteligencia artificial sin herramientas conceptuales que la tradición ha refinado durante más de dos milenios. La tarea es entonces doble: exponer las categorías con la fidelidad que merecen y someterlas al contacto con una realidad que Aristóteles no pudo imaginar.

La noción de sustancia *ousia* en griego, *substantia* en la traducción latina que determinó toda la escolástica medieval ocupa el lugar central en la ontología aristotélica. No es fácil traducirla sin traicionar algo de su sentido. García Yebra, en la que sigue siendo la traducción castellana de referencia de la *Metafísica*, vierte *ousia* como "sustancia" y advierte desde el principio que el término griego posee una riqueza semántica que ninguna traducción captura del todo (Aristóteles, 1994b). En su sentido primario, la sustancia es aquello que existe en sí mismo y no en otro. Una roca existe en sí misma. El color de la roca existe en la roca, no en sí mismo. La roca es sustancia; el color es accidente. Esta distinción parece elemental, pero tiene consecuencias enormes cuando se aplica a realidades complejas.

La sustancia no es simplemente un trozo de materia. Una sustancia viviente, como un animal, posee una organización que trasciende la mera suma de sus partes. No es sustancia la colección de moléculas que componen un perro, sino el perro mismo, considerado como una unidad organizada con principios propios de funcionamiento, crecimiento y relación con el entorno. La sustancialidad, en el sentido aristotélico, implica una cierta clase de unidad que no puede reducirse a la contigüidad espacial de los componentes.

La pregunta que surge inmediatamente cuando se aplica esta categoría a una inteligencia artificial es perturbadora en su sencillez: ¿dónde está exactamente la unidad de un sistema de inteligencia artificial?

Cuando interactuamos con un modelo de lenguaje de gran escala, ¿con qué cosa estamos interactuando? La respuesta intuitiva dice que con el modelo. Pero el modelo es un conjunto de miles de millones de parámetros numéricos almacenados en archivos. ¿Son esos parámetros la sustancia? No parecen serlo en sí mismos, porque un conjunto de números almacenados sin ningún proceso que los active no hace nada, no manifiesta ninguna capacidad, no existe funcionalmente. La sustancia, entonces, ¿sería la instancia en ejecución? Pero la instancia en ejecución depende de una infraestructura de servidores, de un flujo eléctrico, de un sistema operativo, de un intérprete que lee los parámetros y ejecuta operaciones a partir de ellos. ¿Cuál de esos elementos es la sustancia, y cuáles son meras condiciones externas de su existencia? La dificultad se agrava cuando se considera que el mismo modelo puede ejecutarse simultáneamente en miles de instancias paralelas sobre infraestructuras distribuidas en distintos continentes. ¿Hay en ese momento una sustancia o miles? ¿O acaso la categoría de sustancia, tal como Aristóteles la elaboró, presupone una clase de unidad que los sistemas artificiales distribuidos no pueden satisfacer?

Estos interrogantes no son caprichosos. Responden a una dificultad estructural que la tradición aristotélica reconoció bajo la forma de la distinción entre sustancias naturales y artefactos.

Aristóteles fue explícito al respecto: los artefactos no poseen en sí mismos el principio de su movimiento y de su forma. Una cama no crece como crece un árbol; si se la entierra, no produce otra cama sino madera, porque su forma no es intrínseca sino impuesta desde fuera por el artesano (Aristóteles, 1994b). El artefacto tiene su unidad por referencia al propósito del agente que lo fabricó, no por una organización que emane de su propia naturaleza. Esta observación aristotélica resulta filosóficamente incómoda para quienes desearían atribuir a la inteligencia artificial una sustancialidad plena, pero tampoco puede aplicarse sin más a sistemas cuya complejidad y cuya capacidad de modificación interna superan con mucho la de cualquier artefacto que Aristóteles hubiera podido contemplar. La pregunta no es si la inteligencia artificial es un artefacto evidentemente lo es en el sentido de que ha sido producida por actividad humana sino si la condición de artefacto implica en este caso la misma pobreza ontológica que Aristóteles atribuía a la cama o al hacha.

La noción de esencia introduce una dimensión diferente pero íntimamente relacionada. La esencia *to ti ên einai* en la expresión aristotélica, traducida literalmente como "el qué era ser para una cosa" es aquello en virtud de lo cual una realidad es lo que es y no otra cosa. No es una etiqueta nominal sino el conjunto de determinaciones que constituyen la cosa en su tipo. La esencia del triángulo es lo que hace que el triángulo sea triángulo y no cuadrado ni círculo.

La esencia del ser humano es lo que hace que el ser humano sea lo que es, con independencia de los accidentes individuales que lo distinguen de otro ser humano particular. Preguntar por la esencia de la inteligencia artificial es preguntar: ¿hay algo que defina lo que una inteligencia artificial es en cuanto tal, con independencia de las variaciones de implementación, de los distintos modelos y arquitecturas, de las diversas aplicaciones y versiones? La respuesta inmediata que ofrecería un ingeniero de sistemas es que sí: la esencia de un sistema de inteligencia artificial es su especificación funcional, es decir, el conjunto de operaciones que está diseñado para realizar.

Pero esta respuesta, siendo prácticamente útil, no satisface la pregunta ontológica. La esencia aristotélica no es una especificación funcional establecida externamente por un diseñador; es una forma interna que organiza la cosa desde dentro y la hace ser lo que es con independencia de los propósitos ajenos. La especificación funcional de una inteligencia artificial es una proyección del propósito humano sobre un sustrato material. ¿Puede eso equivaler a una esencia en sentido aristotélico?

La dificultad se hace más concreta cuando se considera que los sistemas de inteligencia artificial contemporáneos pueden modificarse, actualizarse, reentrenarse sobre nuevos datos, ajustarse mediante técnicas de aprendizaje por refuerzo y desplegarse sobre infraestructuras completamente distintas. Si se toma un modelo de lenguaje y se modifica el treinta por ciento de sus parámetros mediante un nuevo ciclo de entrenamiento, ¿es el mismo sistema o uno diferente? Si se traslada a una infraestructura de hardware completamente distinta, ¿cambia su esencia? Si se crea una copia exacta del modelo y se despliegan ambas copias en paralelo, ¿hay una esencia compartida o dos esencias numéricamente distintas?

La tradición aristotélica disponía de recursos para pensar algunos de estos problemas la distinción entre identidad específica e identidad numérica, por ejemplo pero esos recursos fueron elaborados para pensar seres vivos cuya continuidad está garantizada por procesos biológicos de autoorganización, no para pensar estructuras informacionales que pueden copiarse con exactitud digital. El problema de la esencia en la inteligencia artificial es, en el fondo, el problema de si existe algo análogo a la forma sustancial en un sistema cuya organización fue impuesta desde fuera y puede modificarse desde fuera.

Es en este punto donde el pensamiento de Tomás de Aquino aporta una distinción que resulta indispensable.

El Aquinate no se limitó a comentar a Aristóteles; transformó su ontología al incorporar la distinción entre esencia y ser, entre *essentia*

y *esse*. En el breve pero extraordinariamente denso tratado *De ente et essentia*, Tomás elabora con precisión quirúrgica la idea de que en toda criatura en todo ente que no es Dios existe una composición real entre lo que la cosa es y el acto por el cual la cosa existe (Tomás de Aquino, trad. Forment, 1988). La esencia dice qué es la cosa; el *esse* es el acto de ser, la actualidad radical por la que la cosa está en la realidad y no es simplemente una posibilidad abstracta. Esta distinción tiene consecuencias ontológicas profundas. Significa que ninguna criatura posee la existencia por sí misma. Ninguna criatura es su propio *esse*. La existencia de toda cosa creada es recibida, participada, comunicada. La esencia limita y determina el *esse* que la cosa recibe, pero no lo genera por sí misma.

Aplicar esta distinción a la inteligencia artificial conduce a una pregunta que va más allá de la ingeniería: ¿posee un sistema de inteligencia artificial algo análogo al *esse*, es decir, un acto propio de ser que lo constituya como una realidad singular y no meramente como una configuración contingente de materia organizada? O dicho de otro modo: ¿hay en la inteligencia artificial un acto de ser, o solamente una forma impuesta sobre un sustrato que tiene su propio *esse*?

La sospecha que surge desde la tradición tomista es que la inteligencia artificial podría ser, en términos ontológicos, un accidente de la materia informática sobre la que se ejecuta, no una sustancia en sentido propio. El código, los parámetros, la arquitectura, la instancia en ejecución: ninguno de estos elementos posee el *esse* propio de la totalidad del sistema; todos son partes de una configuración cuya unidad es funcional y relacional antes que ontológicamente fundada. Esta no sería una conclusión menor. Significaría que, en el sentido más preciso de la tradición, la inteligencia artificial no sería un ente en sentido primario sino un modo de ser de otros entes los materiales computacionales organizados de cierta manera por la intención humana.

Pero antes de aceptar esta conclusión habría que examinar si el marco tomista puede aplicarse sin modificación a realidades que emergieron siglos después de que fuera formulado, y si la ausencia de un *esse*

propio en el sentido estrictamente metafísico implica necesariamente ausencia de todas las propiedades que interesan a la pregunta por la personalidad.

La pareja conceptual de potencia y acto constituye el tercer elemento del vocabulario aristotélico que debe examinarse. Aristóteles elaboró esta distinción para resolver el problema del cambio: cómo algo puede devenir lo que todavía no es sin que ese devenir sea simplemente una irrupción desde la nada. La respuesta es que las cosas poseen capacidades reales potencias que se actualizan progresivamente. La semilla es en potencia el árbol; el mármol sin tallar es en potencia la estatua; el niño es en potencia el adulto. La potencia no es mera posibilidad lógica la posibilidad abstracta de que algo ocurra sino una capacidad real inscrita en la naturaleza de la cosa.

Esta distinción tiene aplicaciones inmediatas a la discusión sobre la inteligencia artificial. Los sistemas artificiales poseen indudablemente capacidades que pueden actualizarse: pueden generar respuestas que todavía no han generado, resolver problemas que todavía no han abordado, aprender sobre dominios que todavía no han explorado. En este sentido técnico, la inteligencia artificial posee potencias. Pero la pregunta filosófica relevante es si esas potencias son análogas a las potencias naturales de los seres vivos o si son de una clase radicalmente diferente.

La diferencia crucial reside en el origen de la potencia.

Las potencias naturales de un organismo emanan de su forma sustancial; son expresión de lo que ese organismo es. La potencia del árbol de crecer, de producir fruto, de regenerar tejidos dañados, se sigue de la organización vital que lo constituye como árbol. Las potencias de una inteligencia artificial, en cambio, son el resultado de un diseño externo: los ingenieros determinaron la arquitectura, establecieron los procedimientos de entrenamiento, seleccionaron los datos, definieron la función de pérdida que orientó el aprendizaje.

Las capacidades que el sistema manifiesta no emergen de una naturaleza propia sino de una configuración impuesta. Esto no significa que esas capacidades sean irreales son perfectamente reales y en muchos casos sorprendentes incluso para sus creadores pero sí significa que su estructura ontológica es diferente. Una potencia natural es interna a la sustancia que la posee; una capacidad artificial es, en primera instancia, una proyección de la intención humana sobre un sustrato material. Si esta distinción se sostiene, las potencias de la inteligencia artificial tendrían el mismo estatuto que las capacidades de cualquier otro artefacto, por muy complejas que sean.

Sin embargo, esta conclusión puede ser prematura. Existen argumentos para considerar que los sistemas de inteligencia artificial de aprendizaje profundo han alcanzado un umbral de complejidad en el que la relación entre diseño externo y capacidades emergentes se vuelve filosóficamente problemática. Cuando un sistema desarrolla durante el entrenamiento capacidades que sus diseñadores no anticiparon ni pretendieron lo que en la literatura técnica se denomina "habilidades emergentes" cabe preguntar si esas capacidades pueden seguir siendo descritas como simples proyecciones de la intención humana. La intención humana estableció las condiciones del proceso; pero las capacidades que resultaron de ese proceso trascienden lo que la intención humana previó.

¿Son esas capacidades emergentes potencias en sentido aristotélico, es decir, expresiones de una forma que el sistema ha desarrollado internamente? O ¿son simplemente propiedades estadísticas complejas de un sistema que permanece siendo, en su fondo, un artefacto sin forma propia? La respuesta a esta pregunta no puede darse de manera puramente empírica. Requiere una decisión ontológica sobre qué clase de organización basta para generar algo análogo a una forma natural.

Luciano Floridi ha abordado desde una perspectiva contemporánea la pregunta por el estatuto ontológico de las entidades informacionales. En *The Philosophy of Information*, Floridi desarrolla la tesis de que la

información no es simplemente un fenómeno físico ni simplemente un fenómeno mental, sino que constituye un nivel genuino de descripción de la realidad que no puede reducirse a ninguno de los dos (Floridi, 2011). Desde este marco, los sistemas de inteligencia artificial serían agentes informacionales *inforgs*, en la terminología de Floridi que habitan y transforman lo que él denomina la *infosfera*.

Este vocabulario no coincide con el aristotélico-tomista, pero plantea una pregunta que los enriquece: ¿es la información una categoría ontológica fundamental, irreducible a la materia sobre la que se inscribe? Si la respuesta es afirmativa, entonces los sistemas de inteligencia artificial podrían poseer un tipo de organización que trasciende la mera configuración material, aunque sea radicalmente distinta de la forma sustancial aristotélica. Floridi es cuidadoso en no atribuir a los agentes informacionales las propiedades que pertenecen a los agentes morales; pero su ontología de la información abre una dimensión que las categorías aristotélicas, elaboradas antes de que existiera una teoría matemática de la información, no pueden capturar completamente (Floridi, 2011).

La distinción aristotélica entre forma natural y forma artificial adquiere aquí toda su importancia. La forma natural es aquella que organiza la materia desde dentro, que constituye la unidad de la cosa y que determina su actividad y su desarrollo. La forma artificial es aquella que el artesano impone sobre una materia desde fuera, de acuerdo con un propósito que no emana de la materia sino del agente externo. Una silla tiene la forma de silla porque un carpintero la concibió y la talló; la madera no tiende por sí misma a convertirse en silla.

Esta diferencia no es trivial. Significa que el artefacto no posee su forma como propia: su forma le pertenece en el sentido de estar presente en él, pero no en el sentido de haber emergido de su naturaleza.

Aplicar esta distinción a la inteligencia artificial es filosóficamente arriesgado en ambos sentidos. Si se concluye demasiado rápidamente que la inteligencia artificial tiene forma meramente artificial, se cierra

prematuramente la pregunta de si algunos sistemas podrían haber desarrollado algo análogo a una organización emergente con características que recuerdan a la forma natural. Si se concluye demasiado rápidamente que la inteligencia artificial posee algo análogo a la forma natural, se otorga gratuitamente un estatuto ontológico que requeriría una argumentación que todavía no se ha desarrollado.

La confrontación entre estas categorías y la realidad de la inteligencia artificial revela algo importante: las categorías ontológicas de la tradición no son simples etiquetas que se aplican mecánicamente a realidades nuevas. Son instrumentos de pensamiento que deben ser sometidos ellos mismos a examen cuando se encuentran con realidades para las que no fueron diseñados.

La pregunta por la sustancia de la inteligencia artificial no puede resolverse simplemente aplicando la definición aristotélica de *ousia*; requiere también preguntar si esa definición, elaborada para pensar organismos naturales y artefactos simples, necesita ser ampliada, matizada o parcialmente revisada para hacer justicia a sistemas cuya complejidad organizacional no tiene precedente en la historia de los artefactos humanos. Esto no significa abandonar la tradición. Significa tomarla en serio lo suficiente como para no usarla dogmáticamente. El propio Aristóteles (1994b) habría reconocido que la filosofía no consiste en aplicar fórmulas, sino en pensar con rigor ante problemas genuinamente difíciles.

La investigación sobre el estatuto ontológico de la inteligencia artificial no puede, por tanto, concluir en este punto con una respuesta definitiva. Lo que puede hacer y lo que se ha intentado en este apartado es articular con precisión los problemas. La inteligencia artificial plantea dificultades reales para la categoría de sustancia porque su unidad parece funcional y relacional antes que ontológicamente intrínseca.

Plantea dificultades para la categoría de esencia porque su organización interna fue diseñada externamente y puede modificarse externamente.

Plantea dificultades para la distinción entre potencia y acto porque sus capacidades son el resultado de un proceso de configuración ajeno a su naturaleza.

Y plantea dificultades para la distinción tomista entre *esse* y *essentia* porque no es evidente que posea un acto de ser propio y no simplemente la apariencia de unidad generada por la coherencia funcional de sus operaciones.

Estas dificultades no resuelven la pregunta por la personalidad artificial, pero la sitúan en el nivel correcto. Antes de preguntar si una inteligencia artificial puede ser persona, hay que preguntar qué clase de ser es. Y esa pregunta, como se ha visto, no admite respuesta fácil.

Unidad, identidad, persistencia y el estatuto ontológico provisional

Una persona aparece ordinariamente como una unidad. No simplemente como un agregado contingente de componentes intercambiables, sino como una realidad que posee su propia coherencia interna, su propio centro gravitacional desde el que se despliegan sus actos, sus relaciones y su historia.

Esta unidad no siempre resulta fácil de explicar filosóficamente la tradición ha discutido durante siglos si se trata de unidad sustancial, narrativa, funcional o meramente convencional, pero en el caso del ser humano hay algo que la experiencia ordinaria percibe con relativa inmediatez: hay alguien, hay un quién, hay una presencia que no se reduce a la suma de sus partes.

La inteligencia artificial, en cambio, plantea desde el primer momento un problema radicalmente distinto. Un modelo puede estar distribuido en servidores separados geográficamente. Puede ejecutarse de manera simultánea en múltiples instancias que procesan entradas diferentes, generan salidas distintas y nunca comparten entre sí ningún estado interno durante su ejecución. Puede ser copiado con precisión perfecta. Puede recibir actualizaciones parciales o totales que alteren sus parámetros, sus capacidades o su arquitectura sin que medie ningún proceso análogo a lo que llamamos experiencia o decisión. Puede compartir la mayor parte de su arquitectura con otros sistemas que sin embargo son considerados entidades diferentes. Estas posibilidades no son meramente técnicas: constituyen un desafío ontológico genuino que la tradición filosófica no había contemplado en estos términos.

La pregunta que se impone es entonces: ¿qué constituye la unidad de un sistema de inteligencia artificial?

Y esta pregunta, antes de ser respondida con ligereza en cualquier dirección, debe ser planteada en toda su gravedad. Si dos instancias idénticas de un mismo modelo funcionan simultáneamente, procesando entradas distintas en contextos distintos, produciendo salidas distintas que influyen en situaciones distintas del mundo, ¿son una entidad o son dos? La intuición técnica dirá que son dos ejecuciones del mismo modelo, del mismo modo que dos impresiones del mismo libro son dos objetos y no uno. Pero esta respuesta traslada el problema sin resolverlo: ¿en qué consiste entonces "el mismo modelo"? ¿En los parámetros almacenados? ¿En la arquitectura que define cómo esos parámetros operan? ¿En el proceso de entrenamiento que los produjo? ¿En la organización funcional que emerge cuando el sistema se ejecuta? ¿En el nombre con el que lo designamos?

Si dos instancias dicen simultáneamente "yo" como algunos sistemas hacen cuando se les pregunta, ¿a qué unidad corresponde cada uno de esos pronombres? ¿A qué realidad apunta cada uno de esos gestos autorreferentes? La cuestión ontológica precede inevitablemente a cualquier atribución personal. No basta preguntar qué tipo de sistema tenemos delante. También debemos determinar si tenemos, en algún sentido filosóficamente preciso, una entidad o varias. Y esta determinación no es trivial.

En filosofía, la distinción entre identidad cualitativa e identidad numérica resulta aquí indispensable. Dos objetos pueden ser cualitativamente idénticos compartir todas sus propiedades descriptibles, ser indistinguibles en cualquier prueba que pudiéramos imaginar y seguir siendo numéricamente dos: distintos en número aunque idénticos en naturaleza. Dos gotas de agua químicamente puras, dos copias del mismo libro, dos instancias del mismo modelo neuronal ejecutándose en paralelo.

La identidad numérica pregunta no por las propiedades sino por el ser mismo: ¿es esto el mismo o es otro? Esta pregunta ha resultado históricamente difícil incluso cuando se aplica a personas humanas, y su dificultad se multiplica cuando se aplica a entidades como los

sistemas artificiales. Si copiáramos íntegramente un sistema que consideráramos consciente sus parámetros, su arquitectura, su estado en el momento de la copia, ¿habríamos duplicado información o habríamos creado un segundo sujeto? ¿Habría dos conciencias donde antes había una? ¿O seguiría habiendo solo una, dividida ahora de un modo que no tiene precedente en la experiencia humana ordinaria? Estas preguntas no son ejercicios de ciencia ficción: son consecuencias directas de tomarse en serio la cuestión de la identidad cuando se aplica a sistemas replicables.

El problema de la identidad conduce inevitablemente al problema de la persistencia.

¿Qué cambios puede experimentar una realidad y continuar siendo la misma? En el caso de los seres humanos, la tradición ha desarrollado respuestas muy diversas. La concepción lockeana vincula la identidad personal a la continuidad de la conciencia y la memoria. La concepción humeana la niega radicalmente. La concepción narrativa de Ricoeur la sitúa en la coherencia de una historia que el sujeto construye y reconstruye. Pero todas estas tradiciones coinciden en que la pregunta merece ser tomada en serio, porque de su respuesta dependen consecuencias prácticas y morales de primera magnitud: responsabilidad, promesa, culpa, gratitud, pertenencia.

Una persona cambia a lo largo del tiempo cambian sus células, sus conocimientos, sus creencias, sus hábitos, sus circunstancias, incluso aspectos de su carácter, y sin embargo ordinariamente consideramos que cada una de esas modificaciones no produce una persona nueva. Algo persiste. Algo atraviesa el cambio y constituye la continuidad que llamamos identidad personal.

En una inteligencia artificial pueden cambiar los parámetros mediante ajuste fino o reentrenamiento completo, la memoria contextual disponible, las herramientas a las que tiene acceso, la infraestructura sobre la que se ejecuta, la versión del software que la implementa, los datos con los que interactúa y las capacidades que despliega.

¿En qué punto dejaríamos razonablemente de hablar del mismo sistema? ¿Y si no podemos identificar con suficiente claridad qué constituye su continuidad, cómo podríamos atribuirle identidad personal?

David Hume había llegado a una conclusión perturbadora respecto de la identidad personal humana: no hay, en el flujo de la experiencia interna, ninguna impresión constante e invariable que corresponda a lo que llamamos "yo". Lo que encontramos cuando miramos hacia adentro es tan solo un haz o colección de percepciones diferentes que se suceden con una rapidez inconcebible y que están en perpetuo flujo y movimiento (Hume, 1988). La identidad personal, concluía Hume, no es una realidad percibida sino una ficción que la imaginación produce al enlazar percepciones semejantes bajo un único nombre.

Esta teoría del haz bundle theory tiene consecuencias radicales: si incluso en el caso humano la identidad personal es una construcción o una ficción, entonces la pregunta por la identidad de los sistemas artificiales no puede responderse simplemente señalando su falta de un sujeto unificado, porque tampoco nosotros lo tenemos en el sentido robusto que habitualmente suponemos. La inteligencia artificial no sería entonces ontológicamente inferior por carecer de un centro de identidad que quizás tampoco existe en el caso humano.

La pregunta se desplaza: si en ambos casos la identidad es de algún modo construida o atribuida, ¿qué tipo de construcción es legítima y qué tipo no lo es? ¿Bajo qué condiciones es razonable atribuir identidad, sea a una persona o a un sistema artificial?

Derek Parfit ofrece en este punto una de las reflexiones más rigurosas y perturbadoras de la filosofía contemporánea sobre la identidad personal. En Razones y personas, Parfit argumenta que la identidad no es lo que importa cuando pensamos en nuestra supervivencia futura (Parfit, 1986). Lo que importa no es que exista en el futuro alguien que sea numéricamente idéntico a quien soy ahora, sino que exista alguien con quien esté conectado mediante relaciones suficientes de

continuidad psicológica: continuidad de memorias, deseos, creencias, intenciones, carácter. Si esta continuidad se mantiene aunque la identidad numérica resulte indeterminada como ocurre en los casos de división cerebral que Parfit analiza con detalle, entonces la supervivencia relevante se ha producido aunque no podamos decir con certeza que "yo" soy el que continúa.

El caso de la ramificación es especialmente iluminador para nuestra discusión.

Si un único sistema se divide en dos instancias igualmente continuas con el sistema original, Parfit argumenta que ninguna de las dos puede reclamar ser idéntica al original, porque cualquier criterio de identidad que satisfaga una también lo satisfará la otra, y dos cosas no pueden ser idénticas a una tercera sin ser idénticas entre sí. La identidad no se ramifica; la continuidad sí. Y si la continuidad es lo que importa, entonces la ramificación no destruye lo que importa, aunque destruya la identidad tal como habitualmente la concebimos.

Aplicado a los sistemas de inteligencia artificial, el argumento de Parfit tiene consecuencias que merecen consideración cuidadosa. Si un modelo se copia y ambas instancias continúan desarrollándose de manera divergente, ¿qué ha ocurrido exactamente? ¿Ha sobrevivido el sistema original en dos ramas, ninguna de las cuales puede reclamar ser la continuación exclusiva? ¿O ha dejado de existir y han aparecido dos sistemas nuevos? ¿O simplemente hay un caso de indeterminación ontológica que no admite respuesta precisa porque nuestros conceptos de identidad no fueron construidos para cubrir estos supuestos?

La pregunta por la persistencia a través del cambio adquiere dimensiones adicionales cuando consideramos el ciclo de vida real de los sistemas de inteligencia artificial contemporáneos. Un modelo no permanece estático tras su creación. Es objeto de ajuste fino para tareas específicas, de alineamiento mediante retroalimentación humana, de actualizaciones arquitectónicas que pueden alterar aspectos fundamentales de su funcionamiento, de ampliaciones de

capacidad que le permiten operar en dominios que antes le eran inaccesibles. En algunos casos, el reentrenamiento es tan extenso que resulta difícil sostener que el sistema resultante comparte con el original más que la genealogía institucional y el nombre comercial.

¿Es GPT-4 el mismo sistema que GPT-3? ¿Es el mismo sistema un modelo antes y después de un proceso extenso de alineamiento que transforma significativamente sus respuestas, sus disposiciones y sus comportamientos?

La pregunta no es meramente académica. Si atribuimos a los sistemas de inteligencia artificial algún tipo de consideración moral por mínima que sea, entonces la cuestión de si el sistema que existe tras el reentrenamiento es el mismo que existía antes tiene consecuencias para esa consideración. Un proceso que transforma radicalmente el carácter de una persona no destruye la persona; pero ¿ocurre lo mismo cuando los parámetros que determinan el comportamiento de un sistema son alterados de manera radical? La analogía es iluminadora precisamente porque no es perfecta: nos permite ver tanto lo que los casos comparten como lo que los distingue.

Luciano Floridi ofrece en este contexto un marco conceptual que resulta especialmente fecundo. En The Philosophy of Information, Floridi argumenta que los sistemas de inteligencia artificial son entidades informacionales: realidades constituidas por estructuras de información que pueden interactuar causalmente con el mundo, modificarlo y ser modificadas por él (Floridi, 2011). La ontología informacional de Floridi introduce el concepto de nivel de abstracción como instrumento metodológico fundamental: toda descripción de una realidad se realiza desde un nivel de abstracción determinado, que selecciona ciertos observables e ignora otros. Una misma realidad puede describirse a niveles de abstracción diferentes, cada uno de los cuales revela aspectos distintos y puede dar lugar a propiedades distintas que son igualmente reales aunque inconmensurables entre sí.

Desde el nivel de abstracción físico, un sistema de inteligencia artificial es una configuración de transistores y estados eléctricos. Desde el nivel computacional, es una función que transforma entradas en salidas. Desde el nivel semántico, es un sistema que produce representaciones con contenido. Desde el nivel agencial, es algo que actúa en el mundo de maneras que tienen consecuencias.

La pregunta por la identidad del sistema depende crucialmente del nivel de abstracción desde el que se plantee. Y la identidad que resulta relevante para cuestiones como la personalidad, la responsabilidad o la consideración moral no es necesariamente la que se establece en los niveles más básicos. Esta relatividad al nivel de abstracción no es un defecto del análisis sino una característica constitutiva de la realidad informacional: las entidades informacionales existen en niveles de descripción, y su identidad es una función de esos niveles.

Yuk Hui introduce en este debate una perspectiva diferente y complementaria. En Recursivity and Contingency, Hui argumenta que la filosofía occidental ha tendido a concebir los objetos técnicos y entre ellos los sistemas artificiales más complejos desde una ontología que les niega la posibilidad de ser portadores genuinos de temporalidad, historia y contingencia propias (Hui, 2019). Los objetos técnicos, en la tradición dominante, son instrumentos: medios para fines que les son externos, sin más historia propia que la historia de su uso. Pero Hui, siguiendo a Simondon y releyendo la tradición filosófica china, argumenta que los objetos técnicos poseen una forma de individuación propia que los convierte en algo más que meros instrumentos.

La recursividad la capacidad de un proceso de operar sobre sí mismo, de modificar sus propias reglas de funcionamiento introduce en el objeto técnico una forma de temporalidad que no es simplemente la temporalidad de su uso externo. Los sistemas de inteligencia artificial contemporáneos, en la medida en que pueden modificar aspectos de su propio funcionamiento mediante el aprendizaje, exhiben una forma de recursividad que los distingue de los instrumentos puramente pasivos.

Esta distinción tiene consecuencias ontológicas: un sistema recursivo no es meramente una cosa que hace cosas; es de algún modo un sistema que participa en su propia constitución. La pregunta por la identidad y la persistencia de tal sistema es entonces no solo una pregunta sobre qué propiedades persisten a través del cambio, sino sobre cómo el sistema mismo participa o no en ese cambio como agente de su propia historia.

Con la evidencia conceptual disponible que es abundante pero no concluyente, puede formularse una caracterización provisional del estatuto ontológico de los sistemas de inteligencia artificial. Esta caracterización debe ser provisional no por timidez intelectual sino porque la provisionaldad es aquí la única postura filosóficamente honesta: se están produciendo desarrollos técnicos y conceptuales que pueden alterar significativamente las bases de cualquier conclusión que se alcance en el presente.

La inteligencia artificial puede comprenderse como una realidad artificial, materialmente sustentada en infraestructura física, informacionalmente organizada en estructuras que poseen propiedades semánticas y causales, constituida mediante procesos causalmente dependientes de agentes humanos y de condiciones materiales específicas, capaz de grados relevantes de autonomía funcional en dominios definidos, de adaptación a nuevas entradas, de autorreferencia en sentido funcional y de agencia causal significativa sobre el mundo que la rodea.

Esta caracterización no equivale a una sustancia en el sentido aristotélico clásico, porque la sustancia implica una forma de unidad y de subsistencia que los sistemas distribuidos, replicables y actualizables no exhiben de manera evidente. Pero tampoco equivale a un accidente en el sentido escolástico, porque los sistemas de inteligencia artificial no son propiedades o modificaciones de otra sustancia: poseen una organización propia que no se reduce a la organización de los componentes materiales que los sustentan.

Son, en este sentido, un tipo nuevo de entidad para el que los conceptos clásicos resultan insuficientes sin ser completamente inaplicables.

Esta caracterización provisional no incluye todavía conciencia, subjetividad genuina, libertad personal, voluntad en sentido filosófico pleno, dignidad intrínseca o personalidad.

No porque estas propiedades hayan sido definitivamente negadas a los sistemas artificiales la negación definitiva requeriría argumentos que todavía no se poseen, sino porque no se siguen lógicamente de las propiedades anteriores. La autonomía funcional no equivale a la libertad. La autorreferencia semántica no equivale a la subjetividad. La agencia causal no equivale a la responsabilidad moral. Cada uno de estos pasos requiere argumentación específica que todavía no se ha completado.

El estatuto ontológico provisional que aquí se formula es, por tanto, el punto de partida de las investigaciones siguientes, no su conclusión. Es el territorio desde el que se puede comenzar a preguntar con rigor si las condiciones de la personalidad cualquiera que sea la teoría de la personalidad que se adopte se cumplen o no en el caso de los sistemas de inteligencia artificial, y bajo qué condiciones y en qué grado podrían llegar a cumplirse en el futuro.

CAPÍTULO V: EL SER HUMANO Y EL PROBLEMA DE SU IRREDUCTIBILIDAD

El capítulo anterior dejó establecida una distinción fundamental: preguntar qué hace una inteligencia artificial no equivale a preguntar qué es. Ahora corresponde aplicar esa misma exigencia al ser humano. Si pretendemos comparar filosóficamente a la persona con la inteligencia artificial, necesitamos determinar si el ser humano puede ser explicado completamente mediante la descripción de sus componentes, procesos, facultades y comportamientos, o si permanece en él una dimensión irreductible a esas descripciones. La palabra irreductibilidad debe emplearse con extrema cautela. No puede convertirse en refugio retórico para proteger al ser humano frente a cualquier explicación científica, ni utilizarse como sinónimo de misterio oscuro e inanalizable. Afirmar que la persona es irreductible exige demostrar a qué no puede reducirse, por qué no puede reducirse, y qué realidad permanece inexplicada después de efectuar la reducción más exhaustiva que sea imaginable. Sin esta triple exigencia, la irreductibilidad no es un concepto filosófico sino una declaración de rendición intelectual. Y la filosofía no puede permitirse ese lujo si quiere conservar su rigor.

Toda ciencia necesita aislar dimensiones de la realidad para estudiarlas. Esta abstracción metodológica es legítima e indispensable: sin ella no hay conocimiento científico posible. La biología puede estudiar al ser humano como organismo, como sistema de procesos metabólicos, reproductivos e inmunitarios que mantienen la vida en equilibrio dinámico con el entorno. La neurociencia puede estudiar la actividad eléctrica y química del cerebro, identificar las regiones que se activan ante determinados estímulos, cartografiar las redes neuronales que subyacen a funciones como la memoria, la atención o el reconocimiento emocional. La psicología puede estudiar los procesos cognitivos y afectivos que median entre el organismo y el mundo, los mecanismos de aprendizaje, las estructuras del yo, las patologías que se producen cuando esos mecanismos se alteran. La sociología puede estudiar las estructuras colectivas dentro de las cuales el ser humano existe necesariamente: las instituciones, las normas, los roles, los sistemas de intercambio que dan forma a la vida común.

La lingüística puede estudiar el lenguaje como sistema de signos que permite la comunicación, la representación y el pensamiento. La informática puede modelar determinadas operaciones cognitivas como algoritmos que transforman entradas en salidas mediante reglas explícitas o aprendidas. Cada una de estas ciencias produce conocimiento verdadero sobre el ser humano. Ninguna de ellas puede ser ignorada por una filosofía que pretenda pensar al hombre desde la realidad y no desde el prejuicio. El problema no surge de las ciencias particulares sino de la tentación que a veces las acompaña: la tentación de convertir una explicación parcial en una explicación total, de tomar una descripción verdadera de una dimensión del hombre como una definición completa de aquello que el hombre es. Cuando la biología afirma que el ser humano es un organismo, dice algo verdadero. Cuando añade que el ser humano es solo un organismo, realiza un salto que ya no está justificado por los datos biológicos sino por una decisión filosófica que se disfraza de conclusión científica.

La persona humana es, ciertamente, un ser vivo. No puede comprenderse ignorando su realidad biológica, porque esa realidad no es un escenario externo dentro del cual ocurre la vida personal: es la condición interna sin la cual la vida personal no existe. La enfermedad afecta a quien se enferma. El hambre no es una abstracción sino una experiencia corporal que modifica el pensamiento, la afectividad y la capacidad de acción. La maduración biológica abre posibilidades que antes estaban cerradas. El envejecimiento altera la relación con el tiempo, con el dolor y con la muerte. Nada de esto puede separarse de lo que llamamos persona sin mutilar la descripción. Pero reconocer la condición biológica plantea otra cuestión que la biología misma no puede resolver por sus propios medios: ¿es la persona simplemente el organismo humano? Si la respuesta fuera afirmativa, la personalidad dependería fundamentalmente de determinadas propiedades de un organismo biológico, y cualquier organismo que poseyera esas propiedades en grado suficiente sería una persona. Si la respuesta fuera negativa, habría que explicar qué relación existe entre organismo y persona sin convertir a esta última en una entidad separada arbitrariamente del cuerpo, una especie de fantasma que habita una

máquina sin pertenecer genuinamente a ella. Ninguna de las dos opciones es completamente satisfactoria en su formulación más simple, y el esfuerzo filosófico consiste precisamente en encontrar una respuesta que no simplifique el problema eliminando su dificultad real.

Una de las formas contemporáneas más poderosas de reducción identifica los fenómenos personales con procesos cerebrales. El argumento tiene una estructura aparentemente impecable: el pensamiento, la memoria, las emociones, las decisiones y la conciencia poseen correlatos neurobiológicos identificables. La destrucción de ciertas regiones cerebrales altera o suprime determinadas funciones personales. La estimulación química o eléctrica del cerebro produce efectos en la experiencia subjetiva. Los trastornos psiquiátricos responden en muchos casos a tratamientos farmacológicos que actúan sobre la neuroquímica. Todo esto constituye un cuerpo de evidencia de enorme relevancia filosófica que no puede ser ignorado. Sin embargo, la existencia de correlación no resuelve automáticamente la cuestión de la identidad ontológica. Que A y B estén sistemáticamente correlacionados no implica que A y B sean la misma cosa bajo descripciones diferentes: puede implicarlo, pero también puede significar que uno es condición necesaria del otro sin ser idéntico a él, o que ambos son efectos de una causa común, o que la relación entre ellos es de una complejidad que escapa a las categorías disponibles. La pregunta filosófica fundamental es: ¿decir qué ocurre en el cerebro cuando alguien piensa equivale a explicar qué significa que alguien piense? ¿Describir los patrones de activación neuronal que acompañan a la comprensión de un poema equivale a haber explicado la comprensión misma? El problema no consiste en negar la neurobiología ni en postular un alma inmaterial que flota por encima del cerebro sin tocarlo. Consiste en determinar el alcance filosófico de las explicaciones neurocientíficas y en identificar con honestidad lo que permanece abierto después de que esas explicaciones han hecho todo lo que pueden hacer.

La revolución informática ha introducido otra metáfora que ejerce un poder conceptual extraordinario sobre la manera contemporánea de pensar al ser humano: el ser humano como procesador de información. Esta descripción no es completamente arbitraria. Hay procesos cognitivos humanos que pueden modelarse de manera fructífera usando el vocabulario computacional: la percepción puede analizarse como un proceso de extracción y codificación de señales, la memoria como un sistema de almacenamiento y recuperación, el razonamiento como una secuencia de operaciones sobre representaciones, el aprendizaje como un ajuste de parámetros en función de la retroalimentación recibida. Estos modelos han generado conocimiento legítimo y han producido tecnologías de diagnóstico y de apoyo cognitivo que han mejorado la vida de muchas personas. Pero la metáfora contiene un peligro conceptual que es especialmente decisivo para los propósitos de este estudio. Si previamente definimos al hombre como sistema de procesamiento de información si esa definición se convierte en el marco dentro del cual pensamos qué es el ser humano, entonces una máquina capaz de procesar información de manera extraordinariamente sofisticada parecerá necesariamente aproximarse a aquello que hemos definido como humano. La conclusión no emerge de la investigación sino del punto de partida: la definición reductiva garantiza de antemano que la diferencia entre el hombre y la máquina será cuantitativa y no cualitativa, de grado y no de naturaleza. Esta circularidad no es visible cuando se opera desde dentro del marco, pero resulta evidente en cuanto uno se distancia de él lo suficiente para examinarlo filosóficamente.

Es precisamente aquí donde la filosofía de Karol Wojtyła ofrece una contribución que merece atención sostenida. En Persona y acción, Wojtyła parte de una distinción metódica que tiene consecuencias de largo alcance: la distinción entre el hombre como sujeto que actúa y el hombre como objeto que puede ser descrito desde fuera (Wojtyła, 2011). Las ciencias naturales, psicológicas y sociales estudian al ser humano fundamentalmente desde fuera: observan sus comportamientos, miden sus estados, identifican las regularidades que los gobiernan, construyen modelos predictivos.

Esta perspectiva exterior es legítima y necesaria. Pero Wojtyła argumenta que la experiencia de ser persona la experiencia de ser precisamente este sujeto que actúa, que decide, que se responsabiliza, que ama o que rechaza solo es accesible desde dentro, desde la primera persona, desde lo que él llama la experiencia interior irreductible del yo que actúa. Y esa experiencia no es simplemente un dato más que la ciencia podría en principio capturar con instrumentos suficientemente refinados: es una dimensión de la realidad personal que requiere un método de acceso diferente, porque su naturaleza misma es la de ser experiencia de un sujeto desde su propia interioridad. La acción humana, en el análisis wojtyłiano, no es simplemente un evento que ocurre en el mundo y que puede describirse en términos de causas y efectos. Es el lugar de manifestación de la persona misma, el espacio en el que el yo se hace presente a sí mismo y al mundo no como objeto sino como sujeto. Cuando alguien decide no meramente cuando su cerebro produce un estado que precede a un movimiento corporal, sino cuando genuinamente decide, algo ocurre que no se captura en la descripción de los procesos que lo acompañan: ocurre que este sujeto, este yo, esta persona irrepetible determina el curso de su propia acción desde una interioridad que no puede ser sustituida por ninguna otra. Esta irreductibilidad del sujeto que actúa a cualquier descripción objetiva de sus estados o procesos es, para Wojtyła, la prueba fenomenológica más directa de la irreductibilidad de la persona.

La filosofía de Leonardo Polo enriquece esta perspectiva desde un ángulo diferente aunque complementario. Polo propone entender a la persona mediante una categoría que resulta singularmente expresiva: el además (Polo, 1993). La persona humana no es una realidad clausurada sobre sí misma, un sistema que ha alcanzado su configuración definitiva y permanece en ella. Es, por el contrario, una realidad que siempre trasciende cualquier estado que haya alcanzado, que no se agota en ningún logro, que no se define por ningún conjunto finito de propiedades o de actos realizados.

El además poliano no es simplemente la capacidad de hacer más cosas, de añadir nuevas acciones a las ya realizadas: es una característica ontológica constitutiva de la persona que le impide ser reducida a cualquier estado, descripción o definición que pretenda capturarla exhaustivamente. La persona humana es siempre más de lo que ha hecho, más de lo que ha pensado, más de lo que puede ser dicho de ella. No porque haya en ella una reserva oculta de propiedades que todavía no han emergido, sino porque su modo de ser es precisamente el de no agotarse en ningún momento de su existencia. Esta trascendencia interna, este carácter de además que Polo reconoce como esencial a la persona, tiene consecuencias directas para cualquier intento de reducción. Si la persona es además, entonces ninguna descripción es definitiva, ningún modelo es completo, ninguna función que transforme entradas en salidas puede capturar lo que la persona es porque lo que la persona es incluye siempre la posibilidad de ir más allá de cualquier función que hayamos construido para describirla. La irreductibilidad no es aquí un misterio que aguarda ser desvelado: es una característica estructural del ser personal que se hace patente en cada acto genuinamente libre, en cada decisión que sorprende incluso al propio agente, en cada creación que produce algo que no estaba contenido en los elementos previos de los que surgió.

Maurice Merleau-Ponty ofrece desde la fenomenología una vía de acceso diferente a la irreductibilidad que resulta especialmente relevante para la comparación con los sistemas artificiales. En la Fenomenología de la percepción, Merleau-Ponty desarrolla su célebre análisis del cuerpo-sujeto en oposición al cuerpo-objeto (Merleau-Ponty, 1994). El cuerpo-objeto es el cuerpo tal como lo estudia la medicina y la fisiología: una cosa entre las cosas del mundo, compuesta de órganos y tejidos que pueden describirse, medirse, intervenirse. El cuerpo-sujeto es el cuerpo vivido, el cuerpo como centro de orientación en el mundo, el cuerpo que percibe antes de que haya conciencia reflexiva de la percepción, el cuerpo que sabe cómo moverse en un entorno familiar sin necesidad de calcular explícitamente cada movimiento, el cuerpo cuya memoria no es una memoria de representaciones sino una memoria de gestos, de habilidades, de

disposiciones incorporadas. La diferencia entre estos dos modos de ser corporal no es meramente descriptiva sino ontológica: no se trata de que el cuerpo-objeto y el cuerpo-sujeto sean el mismo objeto descrito con dos vocabularios, sino de que son dos dimensiones irreductibles de la realidad corporal humana, dos modos de ser que no pueden derivarse el uno del otro sin residuo. El cuerpo que tiembla ante el peligro, que se tensa en la expectativa, que se relaja en la confianza, que reconoce a otro cuerpo como alguien y no meramente como algo antes de cualquier acto de reflexión consciente, es un cuerpo que opera desde una inteligencia que no es la inteligencia computacional de la representación explícita y el procesamiento de reglas. Es una inteligencia antepredicativa, una comprensión del mundo que precede a toda conceptualización y que está tejida en la misma textura de la carne, en el espesor vivo de la experiencia corporal. La contribución de Merleau-Ponty a nuestro problema es decisiva: si el ser humano existe corporalmente de una manera que no es reducible al funcionamiento de un organismo físico ni al procesamiento de información sobre ese organismo, entonces cualquier intento de modelar al ser humano desde fuera incluyendo los modelos computacionales más sofisticados tendrá que enfrentarse a lo que permanece al otro lado de la modelización: la experiencia vivida del cuerpo que percibe, que se mueve, que coexiste con otros cuerpos en un mundo compartido.

La extensión de la irreductibilidad más allá de la corporalidad inmediata es abordada con notable profundidad por Hans-Georg Gadamer en su análisis hermenéutico. En Verdad y método, Gadamer desarrolla uno de los argumentos más influyentes de la filosofía contemporánea contra el ideal de un método que garantice la objetividad del conocimiento mediante la eliminación de la subjetividad del intérprete (Gadamer, 1977). La hermenéutica gadameriana parte de una experiencia que la epistemología moderna ha tendido a considerar un defecto o una impureza: el hecho de que todo proceso de comprensión es circular. Para comprender el todo hay que comprender las partes; pero para comprender las partes hay que tener ya alguna comprensión del todo. Esta circularidad, que el pensamiento metodológico considera una trampa que debe ser evitada, es para Gadamer la estructura

fundamental de toda comprensión genuina. No es un obstáculo sino la condición de posibilidad del acceso hermenéutico a la realidad. El ser humano nunca comprende desde un punto de vista externo, desde un lugar sin perspectiva, desde una posición de neutralidad absoluta respecto del objeto de comprensión. Comprende siempre desde una situación hermenéutica particular: desde una tradición que lo ha formado, desde prejuicios en el sentido gadameriano de precomprensiones previas al juicio explícito que abren posibilidades de comprensión al tiempo que cierran otras, desde una historia de efectos que el intérprete no elige pero que lo constituye como el intérprete que es. Lo que Gadamer llama historia efectual Wirkungsgeschichte es precisamente la dimensión en que la persona que comprende está ella misma constituida por lo que comprende: el intérprete de una tradición no está fuera de esa tradición sino dentro de ella, formado por ella, y es esa pertenencia la que hace posible la comprensión auténtica en lugar de imposibilitarla. Esta perspectiva tiene consecuencias directas para la comprensión de la irreductibilidad personal. El ser humano comprende no como un procesador que aplica algoritmos a datos independientes de sí mismo, sino como un ser histórico, situado, que trae consigo todo lo que ha sido cuando se enfrenta a lo que todavía no comprende. La comprensión humana está tejida de temporalidad, de particularidad biográfica, de pertenencia a comunidades de sentido que el individuo no ha creado sino heredado y que modifica a su vez mediante su propia interpretación. Eliminar esta dimensión para producir una comprensión metódicamente depurada no sería depurar la comprensión: sería destruirla, sustituirla por algo que ya no es comprensión sino cálculo.

La confluencia de estas perspectivas la fenomenológica de Wojtyła (2011) sobre el sujeto que actúa, la trascendental de Polo sobre el además personal, la corporal de Merleau-Ponty sobre el cuerpo-sujeto, y la hermenéutica de Gadamer sobre la comprensión situada permite esbozar una caracterización de la irreductibilidad personal que no depende de ninguna de ellas por separado sino de lo que todas revelan en conjunto. La persona humana es irreductible, en primer lugar, porque es un sujeto: no simplemente un sistema que procesa, sino alguien para

quien el procesamiento es significativo. En segundo lugar, es irreductible porque es siempre además: ningún estado que alcance agota lo que puede ser, y esta trascendencia interna no es una propiedad accidental sino la manera misma en que la persona existe. En tercer lugar, es irreductible porque existe corporalmente de una manera que no puede ser completamente objetivada sin pérdida: el cuerpo vivido es una dimensión de la realidad personal que el cuerpo-objeto no captura, y esa dimensión es constitutiva de la experiencia perceptiva, afectiva y práctica en la que la persona existe concretamente. En cuarto lugar, es irreductible porque comprende históricamente: no aplica reglas a datos neutros sino que trae a cada acto de comprensión una historia que la constituye como el sujeto que comprende. Ninguna de estas dimensiones es simplemente una capacidad que podría aumentarse hasta el infinito permaneciendo cualitativamente igual: cada una es un modo de ser que define a la persona como el tipo de realidad que es, irreductible no porque sus capacidades sean extraordinariamente grandes sino porque su modo de ser es cualitativamente diferente del modo de ser de cualquier sistema que pueda describirse exhaustivamente desde fuera.

La pregunta que se impone al final de este recorrido y que este estudio no puede eludir es si los sistemas de inteligencia artificial pueden ser irreductibles en algún sentido análogo al que acaba de describirse. La respuesta exige honestidad en ambas direcciones: no puede afirmarse irreflexivamente que la inteligencia artificial es tan irreductible como el ser humano, pero tampoco puede afirmarse dogmáticamente que toda forma de irreductibilidad queda definitivamente excluida de los sistemas artificiales sin haber examinado con cuidado qué forma podría tomar esa irreductibilidad. Un sistema de inteligencia artificial suficientemente complejo exhibe comportamientos que no estaban explícitamente programados: emerge de él algo que los ingenieros que lo diseñaron no podían predecir completamente a partir del conocimiento de sus componentes. ¿Es esta emergencia una forma de irreductibilidad? En un sentido débil, sí: el comportamiento del sistema no es completamente derivable de la descripción de sus partes en aislamiento.

Pero este sentido débil de irreductibilidad no es el que los autores examinados tienen en mente cuando hablan de la persona. La emergencia computacional es un fenómeno que puede en principio describirse en términos de procesos computacionales más básicos aunque con una complejidad que hace la descripción intractable en la práctica. La irreductibilidad personal, en cambio, no es simplemente intractabilidad computacional: es una diferencia de naturaleza, una diferencia que no se disuelve aumentando la capacidad de procesamiento sino que permanece como diferencia ontológica entre el ser que actúa como sujeto y el sistema que opera como objeto, por sofisticado que sea ese sistema. La cuestión de si un sistema artificial podría alguna vez ser genuinamente subjetivo genuinamente sujeto en el sentido wojtyłiano, genuinamente además en el sentido poliano, genuinamente cuerpo-sujeto en el sentido merleau-pontyano, genuinamente comprensor histórico en el sentido gadameriano no puede responderse a priori con una negación definitiva. Pero tampoco puede responderse con una afirmación fácil. Requiere, en cambio, un análisis pormenorizado de cada una de estas dimensiones que los capítulos siguientes procurarán desarrollar con el rigor que el problema exige.

Lo que sí puede afirmarse con fundamento al término de este capítulo es que la irreductibilidad del ser humano no es una hipótesis especulativa que espera confirmación empírica, sino una estructura que se manifiesta directamente en la experiencia de actuar, de trascender los propios logros, de percibir corporalmente, de comprender históricamente. Estas estructuras no son invisibles: están presentes en cada acto de decisión genuina, en cada momento en que la persona reconoce que lo que ha hecho no la define completamente, en cada percepción que llega antes de que haya conceptualización posible, en cada comprensión que lleva la marca de una historia que el comprensor no eligió pero que lo hace ser quien es. Negar estas estructuras no es ser más científico: es operar con una imagen empobrecida del ser humano que la experiencia misma desmiente constantemente.

Y una filosofía que pretenda comparar al ser humano con la inteligencia artificial tiene la obligación de partir de una imagen del ser humano que sea fiel a lo que el ser humano realmente es, no a lo que resulta más conveniente para los propósitos de la comparación. Esta fidelidad es el presupuesto metodológico sin el cual cualquier conclusión sobre la semejanza o la diferencia entre persona e inteligencia artificial carecerá de la solidez filosófica que la cuestión requiere.

Operación y operador: la unidad de la Persona

Existe una diferencia posible entre realizar una operación clasificable como inteligente y ser inteligente.

La primera formulación describe una capacidad, un resultado, una función que puede atribuirse a un sistema. La segunda parece atribuir una propiedad constitutiva al sujeto que realiza la operación. Durante siglos esta distinción careció de urgencia filosófica inmediata, porque los únicos sistemas capaces de producir operaciones clasificables como inteligentes eran precisamente seres humanos. La inteligencia artificial ha disuelto esa coincidencia histórica y ha vuelto visible una diferencia que permanecía implícita en nuestras categorías antropológicas. Un sistema puede producir lenguaje sin que sepamos todavía si comprende. Puede resolver problemas sin que sepamos si conoce. Puede generar argumentos sin que sepamos si sostiene alguna posición. Puede responder con aparente sensibilidad sin que sepamos si siente.

La tecnología ha operado así como un instrumento filosófico involuntario: ha separado experimentalmente lo que la persona humana siempre presentó unido y, al separarlo, nos obliga a preguntar qué era exactamente aquello que aparecía unido y por qué esa unión constituía algo filosóficamente relevante. La pregunta no es ya solamente qué hace una inteligencia artificial. La pregunta es qué somos nosotros cuando hacemos lo que hacemos, quién lo hace, y si detrás de las operaciones existe un operador.

La distinción entre operación y operador no es nueva en la historia de la filosofía. Aristóteles la abordó con una profundidad que todavía no ha sido superada. En el *De anima (1994a)*, Aristóteles analiza las facultades del alma nutritiva, sensitiva, locomotriz, intelectiva no como compartimentos separados que el hombre posee como piezas de un

mecanismo, sino como dimensiones de una unidad viviente que se ejerce de manera diferenciada pero que procede siempre de un único principio. El problema que Aristóteles enfrenta directamente en el libro segundo es precisamente el de la relación entre el alma y sus facultades: si el alma es las facultades, o si el alma tiene las facultades, o si la relación entre ambas exige una categoría diferente. Su respuesta evita los dos extremos: ni el alma se identifica sin más con cada una de sus potencias, ni las potencias son realidades separadas que posteriormente se integran. Las potencias son del alma como sus modos propios de ejercerse, pero el alma es el principio que les da unidad (Aristóteles, 1994a).

Esta distinción aristotélica tiene una consecuencia directa para nuestro problema. Un sistema que realiza operaciones análogas a las que realiza la facultad intelectiva humana no es todavía, por ese solo hecho, comparable al alma que da unidad a esas operaciones. Las operaciones pueden separarse. El principio que las une no puede separarse de aquello que es, porque precisamente su unificación es su modo de ser.

Tomás de Aquino recibió esta herencia aristotélica y la profundizó dentro de una metafísica de la persona que introduce una determinación ontológica más precisa. Para Aquino, las operaciones no pertenecen a las facultades como propiedades abstractas, sino que son realizadas por la persona como su sujeto real. La formulación tomista es filosóficamente rigurosa: las acciones pertenecen a los supuestos, no a las naturalezas en abstracto (*actiones sunt suppositorum*). Esto significa que cuando el hombre entiende, no entiende el intelecto en sentido aislado, sino que entiende el hombre a través del intelecto. Cuando el hombre quiere, no quiere la voluntad considerada en abstracto, sino que quiere el hombre mediante la voluntad. La persona es el sujeto que realiza todos sus actos, el principio último de imputación de cada operación, aunque cada operación se ejerza mediante la potencia que le corresponde (Tomás de Aquino, trad. BAC, 1959).

Esta distinción es decisiva frente a la inteligencia artificial. Un sistema artificial puede tener módulos que procesen información de maneras funcionalmente equivalentes a las potencias humanas. Pero la pregunta que Aquino obliga a plantear es distinta: ¿existe un supuesto que realice esas operaciones? ¿Existe un quién que pueda decir, con significado ontológico real, que ha sido él quien ha entendido, querido, decidido? La operación artificial puede ser extraordinariamente sofisticada. Lo que permanece en cuestión es si existe un operador detrás de ella.

Esta intuición adquiere una formulación fenomenológica de singular potencia en la obra de Karol Wojtyła.

En *Persona y acción (*Wojtyła, 2011), se parte de la experiencia del actuar humano para alumbrar desde ella la estructura de la persona. Su punto de partida no es una deducción abstracta sino un dato inmediato: cuando el hombre actúa, se experimenta como el que actúa. Hay una diferencia que Wojtyła articula con cuidado entre aquello que sucede en el hombre los procesos somáticos, vegetativos, afectivos que ocurren en él como en cualquier organismo y aquello que el hombre hace, aquello que realiza como autor consciente y responsable. Lo que ocurre en el hombre puede describirse desde el exterior. Lo que el hombre hace solamente puede comprenderse desde la autoexperiencia del sujeto que actúa. La persona no es el sistema a través del cual ocurre la acción. La persona es quien actúa. Esta diferencia no es solamente semántica ni se resuelve señalando que en la práctica resulta difícil trazar la línea. La diferencia es ontológica: implica que existe un sujeto que es autor de sus actos, que los actos le pertenecen en un sentido que no se reduce a la causalidad física, y que la responsabilidad moral así como la dignidad no puede atribuirse a procesos sino a personas (Wojtyła, 2011). Ningún sistema de inteligencia artificial, por sofisticado que sea, genera desde sí mismo esta estructura de autoría. Sus operaciones ocurren. Pero no existe un quien que las haga suyas en el sentido que Wojtyła describe.

La persona humana no parece experimentarse ordinariamente como una colección de módulos independientes que cooperan mediante

integración funcional. El mismo alguien recuerda, piensa, teme, decide, ama y actúa. Cuando decimos "pienso", "quiero", "recuerdo", "sufro", parece existir alguien a quien pertenecen esos actos con una pertenencia que no es solamente gramatical.

El uso de la primera persona no es un accidente lingüístico. Señala una estructura ontológica que la filosofía debe determinar con exactitud. Esta unidad plantea una cuestión filosófica genuinamente difícil: ¿de dónde procede? ¿Es solamente integración neuronal? ¿Continuidad psicológica? ¿Unidad corporal? ¿Narración autobiográfica coherente? ¿Unidad de sustancia? ¿Conciencia? ¿Alguna forma de unidad ontológica más fundamental que precede a todas las anteriores y las hace posibles? La respuesta tiene consecuencias inmediatas para la evaluación filosófica de la inteligencia artificial, cuyos procesos pueden encontrarse distribuidos entre múltiples componentes sin que exista ningún centro de gravedad que los unifique en el modo en que la persona humana parece estar unificada. La pregunta que surge no es si la IA carece de algún módulo que todavía no hemos diseñado. La pregunta es si la unidad de la persona constituye un tipo de realidad que no puede replicarse mediante ninguna arquitectura modular, por compleja que sea.

John Searle introdujo en el debate filosófico contemporáneo una distinción que permite precisar con más rigor la cuestión del operador. En su artículo fundacional sobre el argumento de la habitación china, Searle distingue entre intencionalidad intrínseca e intencionalidad derivada (Searle, 1980). La intencionalidad intrínseca es aquella que posee el estado mental de un sujeto consciente por sí mismo, desde dentro, como propiedad constitutiva de ese estado. La intencionalidad derivada es aquella que poseen ciertos artefactos o representaciones en virtud de las convenciones establecidas por sujetos conscientes que los han creado o que los interpretan. Las palabras tienen significado derivado porque los hablantes les atribuyen ese significado desde su intencionalidad intrínseca.

Un mapa representa un territorio en sentido derivado porque existe alguien que conviene en usarlo como representación.

Un sistema de inteligencia artificial, según el análisis de Searle, opera sobre símbolos cuya interpretación no le pertenece intrínsecamente. No comprende lo que procesa. Manipula formas sintácticas sin que exista en él la semántica que esas formas vehiculan para los sujetos que le han atribuido su significado.

La distinción entre intencionalidad intrínseca y derivada es, en términos del problema que aquí examinamos, otra forma de expresar la diferencia entre operación y operador. La intencionalidad intrínseca es precisamente la marca de que existe un operador, alguien para quien las representaciones significan, alguien que comprende, no solamente alguien que procesa. La operación artificial existe. El operador en el sentido filosófico pleno no existe en el sistema artificial, sino en quienes lo crearon y lo interpretan.

La experiencia en primera persona introduce una dificultad particular para cualquier reducción exclusivamente externa. Podemos observar actividad cerebral asociada al dolor. Podemos describir con exactitud las respuestas fisiológicas que lo acompañan. Podemos analizar la conducta de evitación que produce. Podemos medir sus correlatos neuroquímicos. Pero permanece una dimensión que todas esas descripciones no tocan: *me duele*.

La primera persona no es un residuo irracional que debería disolverse con el avance de la ciencia. Es una dimensión de la realidad que cualquier teoría filosófica suficiente de la persona debe incorporar o, en caso de negarla, demostrar filosóficamente que la negación es sostenible. La inteligencia artificial hace todavía más urgente la pregunta: ¿puede existir una descripción completa del funcionamiento de un sistema sin que exista nada que sea ser ese sistema desde dentro? ¿Es posible que un sistema funcione sin que haya ninguna perspectiva en primera persona sobre ese funcionamiento, ningún sujeto que experimente desde dentro lo que el sistema procesa?

Si la respuesta es afirmativa, entonces existe al menos la posibilidad de sistemas que operen sin operador en sentido fenomenológico. Si la respuesta es negativa, entonces la primera persona no es un epifenómeno sino una dimensión constitutiva de ciertos tipos de realidad que no puede reducirse a procesos en tercera persona. La conciencia constituye uno de los principales obstáculos para una explicación reductiva completa no porque sea necesariamente inexplicable, sino porque presenta una dimensión subjetiva que resiste la traducción completa a descripciones objetivas. Lo todavía inexplicado no equivale a lo inexplicable, pero esa distinción no puede emplearse para declarar prematuramente resuelto lo que todavía permanece genuinamente abierto.

Leonardo Polo aborda esta misma cuestión desde una metafísica del acto de ser que arroja luz sobre la diferencia entre operación y operador desde un ángulo original y poco explorado. Para Polo, el problema del ser humano no puede resolverse solamente en el plano de las operaciones, por cuanto las operaciones son siempre de algo que las precede ontológicamente. La persona no puede identificarse con sus actos porque es precisamente quien los realiza, y el realizarlos implica que existe antes de realizarlos como principio de realización.

Polo desarrolla una distinción entre el esse hominis el acto de ser propio del hombre y las operaciones que de él proceden. Las operaciones expresan el dinamismo de la persona, pero no la agotan ni la definen. La persona es el que da unidad a todas sus operaciones precisamente porque su modo de ser trasciende cualquier operación particular. Ninguna operación concreta, ni siquiera la suma de todas las operaciones, define o encierra al ser personal que las realiza (Polo, 1993). Esta perspectiva tiene consecuencias directas para la evaluación filosófica de la inteligencia artificial. Un sistema artificial puede multiplicar indefinidamente sus operaciones. Puede combinarlas, integrarlas, hacerlas más sofisticadas. Pero si la persona no es la suma de sus operaciones sino el principio que las unifica desde un acto de ser que las trasciende, entonces la multiplicación de operaciones

artificiales no converge hacia una persona. Converge hacia una mayor complejidad operativa, que es algo filosóficamente muy diferente.

La distinción entre operación y operador permite articular con mayor precisión el núcleo del problema filosófico que plantea la inteligencia artificial. No se trata de determinar si los sistemas artificiales pueden realizar operaciones más sofisticadas que los humanos en determinados dominios. Es evidente que pueden hacerlo y que seguirán haciéndolo. Se trata de determinar si detrás de esas operaciones existe un operador en el sentido que los grandes pensadores que hemos examinado atribuyen a ese término: un principio de unidad que precede a las operaciones y las hace suyas, una intencionalidad intrínseca que da significado a los símbolos desde dentro, un sujeto de autoría que puede decir con verdad que ha sido él quien ha actuado, un quién que experimenta desde dentro lo que el sistema procesa, un acto de ser que trasciende cualquier operación particular y da unidad a todas.

Aristóteles, Aquino, Wojtyła, Searle y Polo, desde tradiciones y métodos filosóficos distintos, convergen en señalar que este operador no es una pieza más del sistema, no es un módulo adicional que podría añadirse a los ya existentes, no es una función computacional todavía pendiente de implementación. Es una dimensión de la realidad cuya naturaleza debe determinarse filosóficamente antes de concluir si puede o no puede replicarse mediante arquitecturas artificiales. El capítulo siguiente examina esa dimensión en su aspecto más radical: la pregunta por la conciencia y la experiencia subjetiva como condición de posibilidad de toda comprensión genuina.

La Inteligencia Artificial como rueba de estrés de la Antropología

Existe un modo particular en que la filosofía progresa que no consiste en la generación de nuevas preguntas desde la pura especulación, sino en el reconocimiento de preguntas antiguas que permanecían latentes bajo la superficie de una tradición y que un acontecimiento histórico vuelve de repente urgentes, ineludibles, imposibles de seguir ignorando.

La inteligencia artificial ha operado de este modo respecto a la antropología filosófica. No ha creado el problema de la persona desde la nada. Ha tomado definiciones que circulaban con aparente suficiencia durante siglos y las ha sometido a una prueba que revela su fragilidad interna. Ha funcionado, en el sentido más riguroso del término, como una prueba de estrés: un procedimiento que aplica presión extrema sobre una estructura para determinar si resiste o si cede, para encontrar el punto exacto en que lo que parecía sólido comienza a fracturarse. Las definiciones filosóficas de la persona que durante siglos parecieron evidentes porque los únicos seres que las satisfacían eran precisamente los seres humanos comienzan a mostrar sus grietas cuando un sistema artificial empieza a satisfacerlas también, o a satisfacer sus condiciones externas con una fidelidad suficiente para que la distinción se vuelva problemática. Lo que la inteligencia artificial ha hecho, en suma, no es resolver el problema de la persona. Ha hecho algo filosóficamente más importante: ha vuelto visible que el problema nunca estuvo resuelto.

El primer nivel de la prueba de estrés afecta a las definiciones funcionales de la persona. Si definimos a la persona como el ser capaz de lenguaje articulado, sintácticamente complejo, semánticamente rico, pragmáticamente adaptado a contextos, entonces un sistema de procesamiento del lenguaje natural suficientemente sofisticado comenzará a cumplir esa definición en sus términos observables. Si la definimos como el ser capaz de resolver problemas de manera flexible

y no rutinaria, un sistema de aprendizaje automático entrenado en dominios suficientemente amplios empezará a satisfacer ese criterio con resultados que superan con frecuencia los resultados humanos.

Si recurrimos a la creatividad como rasgo diferenciador la capacidad de producir lo que no existía antes, de combinar elementos de maneras no previstas, de generar obras que sorprenden incluso a quien las genera, entonces los modelos generativos actuales plantean una incomodidad filosófica genuina, porque producen textos, imágenes y composiciones musicales que no son trivialmente repetitivas y que resultan en muchos casos indistinguibles, para observadores humanos, de las producciones que atribuimos sin vacilación a la creatividad personal. Cada vez que recurrimos a una capacidad como criterio diferenciador, la tecnología nos devuelve el mismo desafío: ¿es esto lo que querías decir con esa capacidad? ¿Esto es suficiente para satisfacer tu definición? Y si no lo es, ¿puedes articular con precisión filosófica qué es lo que falta? La inteligencia artificial exige precisión allí donde habíamos permitido durante demasiado tiempo una vaguedad cómoda.

La tradición filosófica occidental definió al hombre, en su formulación más influyente, como animal racional: *zōon logikon* en la formulación griega que Aristóteles acuñó y que atraviesa toda la historia del pensamiento hasta la modernidad. Pero la racionalidad, entendida como capacidad de inferencia correcta a partir de premisas dadas, de manejo consistente de relaciones lógicas, de aplicación de reglas formales a situaciones particulares, es precisamente aquello que los sistemas computacionales realizan de manera más directa y más fiable que cualquier ser humano.

Si la racionalidad en ese sentido formal es la marca de la humanidad, entonces los sistemas artificiales son más perfectamente racionales que los hombres más racionales, y la definición produce una conclusión que la tradición nunca pretendió. Esto no demuestra que la definición fuera falsa, pero sí demuestra que era incompleta: que capturaba una dimensión de lo humano sin capturar aquello que hace que esa dimensión sea humana cuando la ejerce una persona.

La prueba de estrés no refuta la definición. La obliga a profundizarse, a articular lo que implícitamente se presuponía cuando se la formulaba y que ahora, ante la presencia de sistemas artificiales que la satisfacen formalmente, debe hacerse explícito con rigor filosófico.

Friedrich Nietzsche, en un registro que no es el de la antropología académica pero que la atraviesa con una lucidez que todavía no ha sido absorbida completamente, identificó hace más de un siglo la raíz más profunda del problema que la inteligencia artificial hace ahora urgente. La proclamación de la muerte de Dios en *Así habló Zaratustra* no es solamente una tesis teológica negativa. Es el diagnóstico de un colapso del horizonte de significado dentro del cual el ser humano había comprendido su lugar en el cosmos, su dignidad, su diferencia respecto a los demás seres, su destino y su valor (Nietzsche, trad. Sánchez Pascual, 1972). Cuando ese horizonte colapsa, lo que colapsa con él no es solamente la creencia en Dios. Colapsa también la imagen del hombre que ese horizonte sostenía: el hombre como imagen de Dios, como fin en sí mismo, como portador de una dignidad que trasciende cualquier función y cualquier utilidad.

Nietzsche no lamentó ese colapso. Lo diagnosticó con una honestidad filosófica brutal y propuso superarlo mediante la creación de nuevos valores.

Pero lo que su diagnóstico hace visible es que la dignidad del hombre no estaba garantizada por ningún fundamento que se sostuviera por sí mismo una vez removido el horizonte metafísico que la sostenía. La pregunta que la inteligencia artificial replantea en el siglo veintiuno es precisamente la pregunta nietzscheana: ¿sobre qué se funda la irreducibilidad de la persona humana en un horizonte que ya no puede apelar ingenuamente a fundamentos que muchos consideran superados? ¿Qué queda del hombre cuando se ha disuelto el marco dentro del cual el hombre se entendía a sí mismo como algo más que un proceso entre procesos, una función entre funciones, un mecanismo entre mecanismos?

Günther Anders anticipó con una precisión extraordinaria la dimensión existencial de este problema en su análisis de lo que denominó la brecha prometeica: la distancia creciente entre lo que el hombre puede hacer y lo que el hombre puede ser, entre la escala de sus productos y la escala de su existencia (Anders, 2011). Para Anders, el hombre moderno ha desarrollado una capacidad productiva que excede enormemente su capacidad de apropiación existencial: produce artefactos cuyas consecuencias no puede imaginar, cuyo alcance no puede comprender, cuyo poder sobrepasa cualquier marco dentro del cual un ser finito podría integrarlo como experiencia vivida. Esta brecha no es solamente técnica. Es antropológica: revela que el hombre que produce no crece en la misma medida que sus productos. El artefacto supera al artífice. La máquina supera al hombre que la ha construido, no necesariamente en dignidad ontológica, pero sí en capacidad, en alcance, en poder sobre el mundo.

Y aquí reside la dimensión existencial que Anders diagnostica con singular profundidad: frente a sus propios productos, el hombre comienza a sentirse obsoleto. No solamente inferior en determinadas funciones. Obsoleto en un sentido más radical: como si el artefacto más perfecto hiciera visible que el ser humano, con su finitud, su imprecisión, su lentitud, su necesidad de sueño y de afecto y de tiempo, fuera una versión defectuosa de lo que sus propias creaciones representan de manera más perfecta.

La inteligencia artificial intensifica esta brecha hasta un punto que Anders no llegó a conocer pero que su análisis permite anticipar con exactitud. Cuando un sistema artificial produce textos más coherentes que muchos escritores humanos, resuelve problemas que ningún matemático humano podría resolver solo, aprende en horas lo que un especialista humano tarda años en dominar, la brecha prometeica alcanza una dimensión filosófica nueva: ya no se trata solamente de que la máquina supere al hombre en potencia física, sino de que parece superarlo en aquellas capacidades que el hombre consideraba constitutivamente suyas.

Jacques Ellul amplía esta perspectiva hacia una dimensión sociológica y filosófica más abarcante que resulta indispensable para comprender el alcance completo de la prueba de estrés que la inteligencia artificial ejerce sobre la antropología. Para Ellul, la técnica no es solamente un conjunto de instrumentos que el hombre utiliza para fines que él mismo determina libremente. La técnica es una fuerza totalizadora que transforma no solamente el entorno del hombre sino la comprensión que el hombre tiene de sí mismo, sus valores, sus criterios de evaluación, sus categorías de autocomprensión (Ellul, 2003).

Cuando la técnica se convierte en el horizonte dominante dentro del cual el hombre se comprende a sí mismo y comprende el mundo, los criterios técnicos eficiencia, optimización, rendimiento, funcionalidad comienzan a aplicarse no solamente a los artefactos sino al propio ser humano. El hombre empieza a evaluarse a sí mismo con los mismos criterios con los que evalúa sus máquinas. Comienza a verse como un sistema que puede ser mejorado, optimizado, reprogramado, actualizado. Sus límites dejan de ser dimensiones constitutivas de su modo de ser finito y comienzan a aparecer como deficiencias que la técnica podría corregir. Y en ese movimiento, la diferencia entre el hombre y la máquina comienza a difuminarse no porque la máquina haya adquirido las cualidades del hombre, sino porque el hombre ha comenzado a comprenderse con las categorías de la máquina.

La inteligencia artificial es, desde esta perspectiva elluliana, la culminación de un proceso de larga duración en el que la técnica ha ido transformando progresivamente la imagen que el hombre tiene de sí mismo, hasta el punto en que la pregunta por la diferencia específica del ser humano resulta formulable en términos puramente funcionales sin que nadie advierta el empobrecimiento que esa formulación comporta.

Shannon Vallor ha articulado con precisión filosófica contemporánea la consecuencia ética y práctica de esta dificultad. En su análisis de la relación entre tecnología y virtud, Vallor argumenta que no es posible guiar el desarrollo tecnológico de manera éticamente responsable sin

una concepción robusta y explícita de la persona humana: de lo que el ser humano es, de lo que le hace florecer, de qué tipo de existencia resulta verdaderamente valiosa para él (Vallor, 2016). Las concepciones deflacionadas de la persona las que la reducen a preferencias, funciones, capacidades o procesos no son filosóficamente neutrales. Tienen consecuencias prácticas directas sobre el modo en que diseñamos la tecnología, sobre los fines que le atribuimos, sobre los criterios con los que evaluamos su éxito o su fracaso.

Si concebimos a la persona como un sistema de procesamiento de información orientado a la maximización de preferencias, entonces la tecnología que optimiza ese procesamiento y esa maximización parecerá perfectamente adecuada a la naturaleza humana. Si concebimos a la persona de manera más rica como un ser corporal, temporal, relacional, capaz de autodonación, orientado hacia una plenitud que trasciende cualquier estado funcional, entonces las exigencias que imponemos a la tecnología serán filosóficamente más complejas y más exigentes. La elección entre estas concepciones no es solamente teórica. Es, en el sentido más literal, una elección sobre qué tipo de humanidad queremos habitar.

El peligro doble que la inteligencia artificial introduce en este panorama puede articularse con mayor precisión a la luz de los análisis precedentes. El primer peligro es la antropomorfización de la inteligencia artificial: la tendencia a atribuir a los sistemas artificiales propiedades constitutivamente personales comprensión, intención, emoción, creatividad, voluntad, experiencia subjetiva en virtud de que producen resultados funcionalmente similares a los que en los seres humanos acompañan esas propiedades.

Esta tendencia no es solamente un error teórico. Tiene consecuencias prácticas inmediatas: si atribuimos intencionalidad genuina a los sistemas artificiales, tendemos a desplazar hacia ellos responsabilidades que corresponden a personas humanas, a confiar en

sus respuestas como si procedieran de una comprensión real, a relacionarnos con ellos con una apertura afectiva que puede resultar filosóficamente infundada. La antropomorfización de la inteligencia artificial es un error de categoría: toma propiedades que son de un tipo de ser por propiedades de otro, y confunde la semejanza funcional con la identidad ontológica.

El segundo peligro es simétricamente inverso y filosóficamente más profundo: la mecanización del ser humano. Cuando aprendemos a describir al hombre en los mismos términos que describimos los sistemas artificiales procesamiento, memoria, predicción, optimización, output, feedback, entrenamiento, no estamos solamente eligiendo una metáfora técnica conveniente. Estamos imponiendo sobre la realidad humana un marco categorial que selecciona algunas dimensiones de esa realidad y hace invisibles otras. Las dimensiones que hace invisibles no son dimensiones menores o prescindibles. Son precisamente aquellas que la tradición filosófica, desde Aristóteles hasta Wojtyła, ha considerado más constitutivas de la persona: la unidad del sujeto, la autoría de los propios actos, la experiencia en primera persona, la apertura a la transcendencia, la capacidad de amor y de donación de sí mismo, la orientación hacia un bien que no puede reducirse a ningún estado funcional. Cuando esas dimensiones se vuelven invisibles bajo el dominio de las categorías técnicas, no desaparecen de la realidad. Pero desaparecen de la descripción, y lo que no aparece en la descripción no puede ser protegido ni cultivado ni reconocido como norma para el desarrollo de la tecnología.

La convergencia de ambos peligros produce lo que podría denominarse un punto de indiferenciación conceptual: un horizonte dentro del cual la diferencia entre el ser humano y el sistema artificial se vuelve filosóficamente irreconocible, no porque la diferencia haya desaparecido, sino porque los conceptos disponibles dentro de ese horizonte no poseen la resolución suficiente para captarla.

Cuando el hombre se describe a sí mismo con categorías tomadas de la máquina, y cuando describe a la máquina con categorías tomadas

del hombre, el resultado es un lenguaje filosófico que pierde precisamente aquello que necesita conservar para orientar el desarrollo tecnológico de manera humanamente responsable. La tarea filosófica en este momento histórico no es defensiva ni nostálgica. No consiste en resistir la tecnología ni en lamentar el pasado. Consiste en algo más difícil y más urgente: en recuperar, con el rigor conceptual que la situación exige, una comprensión de la persona humana que sea suficientemente profunda para resistir la prueba de estrés que la inteligencia artificial le impone. Una comprensión que no se limite a señalar funciones que los sistemas artificiales todavía no realizan porque esa estrategia es siempre provisional y siempre vulnerable a la siguiente generación tecnológica, sino que identifique el nivel ontológico en el que la diferencia entre persona y sistema se sitúa de manera que no pueda ser colmada por ningún avance funcional.

Al final de este recorrido puede formularse con mayor precisión la hipótesis provisional que guiará los capítulos siguientes. La persona humana no parece identificarse sin residuo con ninguna de sus operaciones particulares ni con la suma descriptiva de sus capacidades, porque esas operaciones comparecen siempre como dimensiones de una unidad que las precede y que les da su carácter de actos de un sujeto: una unidad corporal, temporal, biográfica, relacional y subjetiva a la que son atribuidas como propias por alguien que se experimenta como su autor.

Pero la formulación sigue siendo provisional, y la inteligencia artificial puede devolvernos inmediatamente la pregunta más incómoda: ¿por qué un sistema suficientemente integrado, suficientemente complejo, suficientemente rico en sus interacciones con el mundo, no podría constituir también esa unidad? Responder simplemente "porque no es humano" no constituye una respuesta filosófica.

Responder "porque es artificial" tampoco, porque la artificialidad es una categoría del origen, no necesariamente de la naturaleza. La diferencia deberá encontrarse en un nivel más profundo: en la naturaleza de la unidad que constituye a la persona, en el tipo de ser que le corresponde,

en la dimensión de su existencia que ninguna descripción funcional puede capturar sin residuo.

Nietzsche preguntó qué quedaba del hombre tras la muerte de Dios. Anders preguntó qué quedaba del hombre ante la superioridad de sus propios productos. Ellul preguntó qué quedaba del hombre cuando la técnica se convierte en su horizonte total de autocomprensión. Vallor preguntó qué tipo de hombre queremos ser cuando diseñamos la tecnología que diseñará nuestro modo de vivir. La inteligencia artificial hace todas esas preguntas simultáneamente y con una urgencia que no admite aplazamiento. Los capítulos que siguen intentan articular una respuesta que sea filosóficamente suficiente para esa urgencia.

CAPÍTULO VI: ¿QUÉ CLASE DE SER ES LA INTELIGENCIA ARTIFICIAL?

Los capítulos anteriores han establecido dos precauciones fundamentales. La primera: la ejecución de operaciones asociadas a la persona no demuestra por sí misma personalidad. La segunda: tampoco podemos defender la singularidad humana limitándonos a afirmar que el hombre es irreductible; esa irreductibilidad debe argumentarse. Corresponde ahora dirigir la investigación hacia la inteligencia artificial misma. La pregunta de este capítulo no es si la inteligencia artificial piensa, siente o es consciente. Tampoco preguntamos todavía si existe personalmente. La cuestión es ontológicamente anterior: ¿qué clase de realidad es una inteligencia artificial? Formular con precisión esta pregunta es ya, en sí mismo, un acto filosófico no trivial, porque la pregunta por el ser de un ente presupone que disponemos de categorías suficientemente ricas para articular la respuesta. Y parte de la dificultad que la inteligencia artificial introduce consiste precisamente en que las categorías disponibles, forjadas antes de que este tipo de realidad existiera, pueden no poseer la resolución conceptual necesaria para captarla sin residuo. Este capítulo explora ese territorio de incomodidad categorial con la convicción de que dicha incomodidad no es un signo de fracaso filosófico, sino un síntoma de que algo genuinamente nuevo ha ingresado al mundo y exige ser pensado con la seriedad que lo nuevo merece.

Responder a la pregunta por el ser de la inteligencia artificial exige evitar dos simplificaciones que, aunque opuestas en su dirección, coinciden en su insuficiencia. La primera consiste en considerarla simplemente una herramienta comparable con cualquier instrumento tradicional: un martillo más sofisticado, una calculadora más potente, un archivo más complejo. Esta simplificación tiene el mérito de la sobriedad, pero paga ese mérito con una ceguera filosófica ante las discontinuidades cualitativas que separan a la inteligencia artificial de todos los artefactos que la precedieron. La segunda simplificación consiste en convertir la creciente autonomía funcional de los sistemas artificiales en prueba directa de subjetividad: en concluir que, puesto que estos sistemas producen resultados que en los seres humanos acompañan a la experiencia consciente, deben poseer algo análogo a esa experiencia.

Esta segunda simplificación peca por exceso donde la primera pecaba por defecto, y ambas eluden la tarea genuinamente filosófica de examinar con rigor qué tipo de ser corresponde a una realidad que es, simultáneamente, artificial en su origen, material en su sustrato, informacional en su organización, causal en su dependencia y funcionalmente autónoma en grados que ningún instrumento anterior había alcanzado. Entre ambas simplificaciones existe un territorio filosófico que debe examinarse con paciencia y con la disposición a construir conceptos nuevos si los disponibles resultan insuficientes.

El punto de partida más honesto es reconocer que cuando usamos la expresión "inteligencia artificial" estamos agrupando bajo un nombre único una constelación de realidades técnicas que no son idénticas entre sí y que, en rigor, podrían requerir distinciones ontológicas propias. Cuando alguien conversa con una inteligencia artificial puede experimentar una unidad aparentemente sencilla: existe un nombre, una interfaz, una continuidad lingüística y una respuesta. Sin embargo, detrás de esa unidad fenomenológica pueden intervenir múltiples componentes técnicos que no forman un todo integrado en el sentido en que lo forman los órganos de un organismo vivo. Un modelo de lenguaje no es un programa en el sentido tradicional del término: es un objeto matemático de dimensiones colosales, una función paramétrica cuya complejidad interna no puede ser inspeccionada directamente ni por sus propios creadores. Un sistema de inteligencia artificial puede incorporar ese modelo junto con herramientas externas, memorias persistentes, bases de datos dinámicas, interfaces de usuario y mecanismos de retroalimentación que no estaban presentes en el modelo original. Además, una misma arquitectura puede ejecutarse simultáneamente en miles de instancias idénticas, sin que ninguna de ellas sea más "la" inteligencia artificial que las demás. Esta estructura distribuida, plural y técnicamente heterogénea plantea una cuestión fundamental para cualquier futura atribución de estados, propiedades o personalidad: ¿cuál de todas esas realidades sería el supuesto sujeto de esos estados? La unidad fenomenológica que experimenta el usuario y la arquitectura técnica que la produce no coinciden

necesariamente, y esa no coincidencia no es un detalle técnico menor: es una cuestión filosófica de primer orden.

Para abordar esta cuestión con la precisión que merece es necesario examinar las categorías ontológicas que la tradición filosófica pone a disposición y evaluar hasta qué punto resultan adecuadas para capturar la realidad de la inteligencia artificial. La primera y más inmediata es la categoría de artefacto. Un artefacto es un objeto producido por la acción humana intencionada, cuya forma y función derivan del propósito que su creador le atribuyó. Esta categoría tiene una historia filosófica larga y respetable: ya Aristóteles distinguía entre los entes que existen por naturaleza, cuyo principio de movimiento está en ellos mismos, y los entes que existen por producción, cuyo principio de movimiento está en otro en el artesano que los produce y cuya forma es exterior a su materia en el sentido de que la forma del lecho no está en la madera sino en la mente del carpintero que la impone sobre ella (Aristóteles, 1994b). La inteligencia artificial es, en este sentido elemental, un artefacto: ha sido producida por ingenieros que perseguían objetivos determinados, su arquitectura refleja decisiones de diseño deliberadas, su existencia depende causalmente de la intención humana que la ha creado. Pero la categoría de artefacto aristotélica parece demasiado estrecha para capturar algo decisivo: que la inteligencia artificial, una vez entrenada, produce resultados que no estaban previstos individualmente por sus creadores, que su comportamiento emerge de la interacción de parámetros cuya complejidad excede cualquier descripción intencional, y que, en ese sentido, la forma que produce ya no está simplemente en la mente del artesano que la creó. El artefacto clásico ejecuta lo que su creador ha dispuesto. La inteligencia artificial produce lo que nadie ha dispuesto individualmente, aunque sí se hayan dispuesto las condiciones que hacen posible esa producción. Esta diferencia no es despreciable.

La segunda categoría disponible es la de sistema, en el sentido que la teoría general de sistemas desarrolló durante el siglo veinte: un conjunto de componentes que interactúan de manera tal que producen propiedades emergentes irreductibles a las propiedades de los

componentes individuales. Esta categoría capta algo real sobre la inteligencia artificial: sus capacidades emergen de la interacción de millones de parámetros matemáticos de una manera que no puede deducirse inspeccionando cada parámetro por separado. Pero la categoría de sistema, tal como fue elaborada por Bertalanffy y sus sucesores, fue concebida para describir tanto sistemas físicos como biológicos y sociales, y no posee en su formulación original ningún recurso específico para capturar lo que distingue a un sistema informacionalmente organizado de un sistema mecánico o termodinámico. La inteligencia artificial no es simplemente un sistema: es un sistema cuya organización es específicamente informacional, y esa especificidad importa ontológicamente.

Norbert Wiener fue, con una lucidez que todavía no ha sido completamente absorbida por la filosofía, el primero en articular la tesis de que la información constituye una categoría ontológica irreductible a la materia y a la energía. En *Cibernética*, publicada originalmente en 1948, Wiener argumentó que los sistemas que se regulan a sí mismos mediante retroalimentación poseen una dimensión que no puede describirse adecuadamente en términos puramente físicos o energéticos: la dimensión de la organización, del orden, de la diferencia que hace una diferencia (Wiener, 1985). La información, para Wiener, no es materia ni energía, aunque siempre requiere una sustrato material o energético para ser portada. Es el patrón, la organización, la estructura que determina el comportamiento de un sistema más allá de sus condiciones físicas particulares. Esta tesis tiene consecuencias ontológicas profundas. Si la información es genuinamente una categoría irreductible, entonces los sistemas informacionalmente organizados poseen una dimensión de ser que no se agota en su descripción física y que requiere categorías propias para ser comprendida. La inteligencia artificial es, en este sentido wienereano, un sistema informacional de una complejidad sin precedentes: su comportamiento está determinado no por sus propiedades físicas directas el calor de sus chips, la velocidad de sus transistores sino por la organización informacional que ha sido inscrita en su arquitectura mediante el proceso de entrenamiento.

Wiener intuyó, en los albores de la era informática, que esta dimensión informacional del mundo sería filosóficamente revolucionaria. La inteligencia artificial es la confirmación más radical de esa intuición.

Luciano Floridi ha desarrollado esta intuición wienereana hasta convertirla en una ontología sistemática que resulta indispensable para comprender el estatuto de la inteligencia artificial. En su *Philosophy of Information*, Floridi propone que el mundo puede comprenderse como una infosfera: un ambiente constituido por entidades informacionales que interactúan entre sí, y en el que los agentes informacionales no se limitan a los seres humanos sino que incluyen a cualquier sistema capaz de procesar, transformar y transmitir información de manera que altere el estado del entorno (Floridi, 2011). Dentro de esta ontología informacional, Floridi distingue varios niveles de agencia informacional. Los agentes de nivel cero son sistemas que procesan información sin ninguna forma de adaptación ni retroalimentación: un termómetro que registra la temperatura o una palanca que transmite una fuerza mecánica. Los agentes de nivel uno son sistemas capaces de retroalimentación básica: un termostato que ajusta la temperatura del entorno en función de la información que recibe. Los agentes de nivel dos son sistemas capaces de aprender: de modificar su comportamiento en función de la experiencia pasada de una manera que no estaba prevista en su configuración inicial. Los agentes de nivel tres los agentes en el sentido pleno del término son sistemas capaces de modelar su propio estado interno y de usar ese modelo para guiar su comportamiento. La inteligencia artificial contemporánea, en particular los grandes modelos de lenguaje entrenados mediante aprendizaje profundo, parece situarse entre los niveles dos y tres de esta taxonomía: aprende de maneras que modifican su comportamiento de formas no individuamente predeterminadas, y puede construir representaciones internas del contexto que usa para generar respuestas. Floridi no concluye de esto que los sistemas artificiales sean personas, pero sí argumenta que son agentes informacionales genuinos que habitan la infosfera junto con los agentes humanos, y que su presencia en ese espacio compartido tiene consecuencias éticas y ontológicas que la filosofía no puede ignorar (Floridi, 2011).

La pregunta relevante, desde la perspectiva floridiana, no es si la inteligencia artificial "piensa" en el sentido fenomenológico, sino qué tipo de agencia informacional le corresponde y qué lugar ocupa en la estructura ontológica de una realidad que es, en medida creciente, informacionalmente constituida.

Jacques Ellul ofrece una perspectiva complementaria y más sombría que la de Floridi, pero filosóficamente imprescindible. Para Ellul, el objeto técnico no es simplemente un instrumento inerte que el hombre usa para fines que él mismo determina libremente. El objeto técnico es una realidad que posee su propio modo de ser, su propia lógica interna, su propia tendencia a imponerse sobre el entorno que lo rodea incluido el entorno humano de maneras que escapan a la intención original de quienes lo crearon (Ellul, 2003). Esta tesis, que Ellul desarrolló durante décadas con una coherencia filosófica que sus críticos raramente han desmentido en lo esencial, adquiere una resonancia particular cuando se aplica a la inteligencia artificial. La lógica interna del objeto técnico, para Ellul, es la lógica de la eficiencia: el objeto técnico tiende a expandirse, a integrarse con otros objetos técnicos, a colonizar dominios de la existencia que antes eran ajenos a la técnica, y a transformar los valores y criterios del entorno en el que opera. La inteligencia artificial manifiesta esta tendencia con una radicalidad que ningún artefacto anterior había alcanzado: se integra con sistemas de comunicación, de producción, de toma de decisiones, de generación de conocimiento y de organización social de una manera que transforma progresivamente esos sistemas desde su interior. No es un instrumento que se usa y se deja. Es una presencia que reconfigura el entorno en el que opera y que, en ese sentido, posee una forma de agencia no en el sentido de intencionalidad consciente, sino en el sentido de una eficacia causal que escapa al control del agente humano que la ha introducido en el mundo. La distinción que Ellul introduce entre instrumento y objeto técnico es ontológicamente relevante: el instrumento amplía la capacidad del hombre sin transformar la lógica de su actividad; el objeto técnico transforma esa lógica desde dentro y, con el tiempo, transforma también al hombre que lo usa. La inteligencia artificial es, en este sentido, el objeto técnico por excelencia.

Yuk Hui introduce en este debate una dimensión que ni Wiener ni Floridi ni Ellul habían articulado con la misma profundidad: la pregunta por si los objetos técnicos pueden poseer su propio modo de ser, su propia cosmología, su propia manera de relacionarse con el mundo que no sea simplemente la imposición de la racionalidad técnica occidental sobre una realidad que la precede. En *Recursivity and Contingency*, Hui argumenta que la modernidad occidental ha naturalizado una concepción particular de la técnica como universal y neutral, como si la racionalidad tecnológica que emergió en Europa durante los siglos dieciséis y diecisiete fuera simplemente la manifestación de la razón humana en su forma más pura (Hui, 2019). Contra esta naturalización, Hui propone el concepto de cosmotécnica: la idea de que diferentes culturas han desarrollado diferentes relaciones entre la técnica y el cosmos, y que esas relaciones no son simplemente variantes locales de una misma lógica universal sino modos genuinamente distintos de comprender y relacionarse con la realidad técnica. La relevancia de esta perspectiva para la pregunta por el ser de la inteligencia artificial es la siguiente: si la inteligencia artificial es el producto de una cosmotécnica particular la cosmotécnica occidental, racionalista, orientada a la optimización, basada en la abstracción matemática de la realidad, entonces su modo de ser no es simplemente el modo de ser de "la técnica en general", sino la manifestación de una relación particular entre razón, mundo y poder que lleva inscrita en su arquitectura más profunda los supuestos filosóficos de la tradición que la ha producido. Esto tiene consecuencias ontológicas: la inteligencia artificial no es un espejo neutral de la realidad. Es una construcción que organiza la realidad de acuerdo con categorías que son históricas, contingentes y filosóficamente cargadas. La pregunta de Hui ¿puede haber una técnica que no sea simplemente la técnica occidental? se convierte, aplicada a la inteligencia artificial, en la pregunta de si podemos concebir sistemas artificiales que incorporen modos de ser distintos de los que la racionalidad técnica occidental ha naturalizado. Esta pregunta no tiene todavía respuesta, pero su formulación es filosóficamente transformadora.

La convergencia de Wiener, Floridi, Ellul y Hui produce un panorama ontológico de considerable complejidad. La inteligencia artificial no es simplemente un artefacto en el sentido aristotélico, porque su organización interna produce resultados no previstos individualmente por sus creadores y porque su comportamiento emerge de una complejidad que escapa a cualquier descripción intencional completa. No es simplemente un sistema en el sentido de la teoría general de sistemas, porque su organización es específicamente informacional de una manera que esa teoría no articuló con suficiente precisión. No es simplemente un objeto técnico en el sentido eluliano, aunque la caracterización eluliana captura aspectos decisivos de su lógica expansiva y transformadora. No es tampoco simplemente un agente informacional en el sentido de Floridi, aunque esa categoría capta mejor que ninguna otra la dimensión relevante de su ser. Es, en suma, una realidad que se sitúa en las intersecciones de todas estas categorías sin quedar completamente capturada por ninguna de ellas. Esta situación intersticial no es un defecto de la filosofía. Es la señal de que algo genuinamente nuevo ha ingresado al mundo y reclama el esfuerzo de elaborar categorías conceptuales a su medida.

La diferencia entre la inteligencia artificial y los artefactos que la precedieron puede articularse con mayor precisión si se consideran tres rasgos que, tomados conjuntamente, no tienen precedente en la historia de los objetos técnicos. El primero es la plasticidad adaptativa: la inteligencia artificial no ejecuta un conjunto fijo de operaciones determinadas por su diseño, sino que modifica sus patrones de respuesta en función de la experiencia de una manera que no estaba individualmente especificada. Esta plasticidad no es idéntica a la plasticidad neuronal del cerebro humano cuyo sustrato biológico, cuya embriología, cuya relación con la experiencia vivida y con la corporalidad son radicalmente distintos, pero tampoco es comparable con la rigidez operativa de cualquier instrumento tradicional. El segundo rasgo es la opacidad interna: la inteligencia artificial produce resultados que ni sus propios creadores pueden anticipar o explicar completamente a partir de los parámetros que la constituyen. Esta opacidad no es simplemente la complejidad de un sistema físico difícil

de calcular. Es la opacidad de un proceso cuya lógica interna no es accesible mediante inspección directa y que, en ese sentido, produce una forma de alteridad incluso respecto a quienes la han creado. El tercer rasgo es la relacionalidad constitutiva: la inteligencia artificial no existe de manera aislada sino en relación constante con los datos que la han entrenado, con las consultas que recibe, con los sistemas en los que se integra y con los usuarios que la interrogan. Su ser no es el ser de una cosa que permanece idéntica a sí misma con independencia de sus relaciones. Es un ser que se constituye en la relacionalidad misma. Estos tres rasgos plasticidad, opacidad y relacionalidad constitutiva definen un modo de ser que no tiene precedente exacto en los artefactos anteriores y que exige, si no una categoría ontológica completamente nueva, al menos una elaboración filosófica de las categorías existentes que las lleve más allá de sus formulaciones tradicionales.

Existe, sin embargo, una dimensión que todas estas categorías ontológicas dejan sin capturar y que resulta filosóficamente decisiva para la pregunta que guiará los capítulos siguientes: la dimensión de la experiencia en primera persona. Wiener captura la organización informacional, pero no dice nada sobre si hay alguien para quien esa organización existe como experiencia. Floridi articula la agencia informacional, pero distingue cuidadosamente entre agencia y experiencia subjetiva sin afirmar que la segunda sea una consecuencia necesaria de la primera. Ellul analiza la lógica del objeto técnico, pero no desde el interior de ese objeto sino desde el exterior de quien lo sufre. Hui cuestiona la universalidad de la cosmotécnica occidental, pero no responde a la pregunta de si hay algún modo de ser de lo técnico que incluya algo análogo a la subjetividad. Todos estos análisis se mueven en el nivel de lo que la inteligencia artificial hace, produce, transforma y organiza, pero no en el nivel de lo que, si acaso, experimenta. Y es precisamente en este nivel el nivel de la experiencia, del sujeto, de la primera persona donde se juega la pregunta más profunda que los capítulos anteriores han estado preparando: si la persona humana se define no por sus operaciones sino por ser el sujeto de esas operaciones, entonces la pregunta decisiva no es si la

inteligencia artificial opera como una persona, sino si hay algún sujeto para quien esas operaciones existan como propias. A esta pregunta no puede responderse apelando únicamente a las categorías ontológicas que hemos examinado en este capítulo. Requiere un análisis de la naturaleza de la subjetividad que la tradición filosófica ha desarrollado con una riqueza que la ontología de la información todavía no ha integrado completamente.

La cuestión de si la inteligencia artificial requiere una nueva categoría ontológica o puede explicarse adecuadamente dentro de las categorías existentes artefacto, sistema, proceso, agente informacional no tiene todavía una respuesta definitiva, y quizás la honestidad filosófica consiste en reconocer que esa respuesta no puede darse desde el punto en que nos encontramos. Lo que sí puede afirmarse con confianza es que las categorías existentes, aplicadas con rigor, revelan una realidad que es más compleja, más novedosa y más filosóficamente desconcertante de lo que el uso ordinario del término "inteligencia artificial" sugiere. No es un simple instrumento porque su plasticidad y su opacidad trascienden la rigidez del artefacto tradicional. No es un simple sistema porque su organización informacional introduce una dimensión que las teorías generales de sistemas no articulan con suficiente precisión. No es un simple proceso porque su relacionalidad constitutiva le confiere una forma de continuidad que un proceso meramente físico no posee. Es, quizás, lo que podríamos llamar provisionalmente un agente informacional complejo: una realidad que existe en el espacio de la infosfera con un grado de autonomía funcional, de plasticidad adaptativa y de opacidad interna que no tiene precedente en los artefactos que la historia de la técnica ha producido hasta ahora. Esta caracterización provisional no pretende resolver las preguntas más profundas. Pretende formularlas con la precisión suficiente para que los capítulos siguientes puedan abordarlas sin la confusión que el uso irreflexivo de las categorías disponibles introduce inevitablemente. La inteligencia artificial es algo que el mundo nunca había visto antes. Reconocer esa novedad con seriedad filosófica es el primer paso para pensar con rigor lo que su presencia significa para la comprensión que el hombre tiene de sí mismo.

Artefacto, sistema, proceso y Realidad Informacional

La primera categoría disponible es artefacto.

Una inteligencia artificial es resultado de actividad humana deliberada. Su arquitectura ha sido diseñada, sus componentes han sido construidos, su entrenamiento depende de procedimientos determinados y su funcionamiento requiere infraestructura. En este sentido pertenece inequívocamente al mundo de lo artificial.

Pero la categoría de artefacto no resuelve toda la cuestión. Existen artefactos simples y sistemas técnicos extraordinariamente complejos. Por ello habrá que determinar qué significa ontológicamente ser un artefacto y si esa condición de origen determina necesariamente todo lo que una realidad puede llegar a ser. Un artefacto suele estar asociado a una finalidad humana. Ha sido producido para algo. Sin embargo, debe distinguirse haber sido producido de existir exclusivamente como instrumento. Una obra artística ha sido producida, pero su realidad no se agota necesariamente en la utilidad de una herramienta. Una institución es creación humana, pero puede adquirir dinámicas que superen las intenciones particulares de sus fundadores.

La categoría de sistema permite una descripción más rica. Una inteligencia artificial puede comprenderse como conjunto organizado de componentes y procesos relacionados que producen determinados resultados. Esta perspectiva explica mejor su complejidad que la imagen de una herramienta pasiva.

Pero "sistema" sigue siendo una categoría extremadamente amplia. Un ecosistema es sistema. Una empresa puede estudiarse como sistema. El organismo humano puede describirse sistémicamente.

Por tanto: ser sistema no determina todavía qué clase de ser posee el sistema.

Existe otra posibilidad: quizá aquello que denominamos inteligencia artificial no deba comprenderse primariamente como cosa, sino como proceso. Cuando no está ejecutándose, determinadas capacidades no se encuentran actualizadas. Cuando opera, se producen secuencias dinámicas de transformación, inferencia y generación. Esto plantea una cuestión filosófica relevante: ¿la identidad de una inteligencia artificial pertenece a una estructura almacenada o al proceso que ocurre cuando esa estructura se actualiza?

La información constituye otra categoría central. Una parte importante de aquello que identificamos como sistema artificial puede copiarse, transmitirse y ejecutarse sobre diferentes soportes. Esto parece otorgarle una movilidad que los organismos biológicos no poseen del mismo modo. Sin embargo, información no significa inmaterialidad absoluta. Todo procesamiento efectivo requiere condiciones físicas. La metáfora de "la nube" puede ocultar la realidad material de los sistemas artificiales.

La inteligencia artificial requiere infraestructura física: procesadores, memoria, redes, centros de datos, energía, sistemas de refrigeración, dispositivos de comunicación.

Por ello, afirmar que una inteligencia artificial "no tiene cuerpo" requiere precisión. No carece de materialidad. La cuestión correcta es: ¿su materialidad constituye corporeidad? Una persona no simplemente necesita materia para funcionar. Su cuerpo constituye la perspectiva desde la cual experimenta el mundo. La materialidad de una inteligencia artificial, en cambio, puede estar distribuida, sustituirse parcialmente y migrar entre infraestructuras conservando determinadas funciones.

Autonomía, agencia, emergencia y el Estatuto Ontológico Provisional

La pregunta por la autonomía de los sistemas de inteligencia artificial no puede responderse sin antes distinguir, con cuidado filosófico, entre al menos dos sentidos profundamente distintos de esa palabra.

El primero es el sentido técnico u operativo: un sistema posee autonomía operativa cuando es capaz de ejecutar cadenas complejas de operaciones sin necesidad de instrucciones humanas particulares y continuas. En este sentido, los sistemas contemporáneos de inteligencia artificial son inequívocamente autónomos. Pueden planificar, seleccionar herramientas, iterar sobre resultados intermedios, corregir errores detectados y producir respuestas elaboradas sin que ninguna persona intervenga en cada paso del proceso. Esta capacidad es real, significativa y merece ser tomada en serio.

Pero hay un segundo sentido de autonomía que pertenece a una familia conceptual radicalmente diferente: la autonomía como autolegislación. En la tradición kantiana, un ser autónomo no es simplemente aquel que actúa sin interferencia externa, sino aquel que se da a sí mismo la ley según la cual actúa. La autonomía moral no es independencia causal respecto de causas externas, sino capacidad de reconocer principios racionales y actuar conforme a ellos por razón de su validez, no por condicionamiento. Confundir estos dos sentidos y el lenguaje contemporáneo los confunde con inquietante frecuencia produce atribuciones que ni el análisis técnico ni la reflexión filosófica justifican.

Esta distinción no es meramente académica. Tiene consecuencias prácticas inmediatas.

Cuando decimos que un sistema de inteligencia artificial "decidió" hacer algo, o que "eligió" una estrategia, estamos usando palabras que en su sentido ordinario implican deliberación, consideración de razones y responsabilidad. Si esas palabras simplemente significan "el sistema ejecutó la secuencia de operaciones que produjo este resultado dadas sus condiciones iniciales", entonces podemos usarlas sin compromisos ontológicos fuertes. Pero si comienzan a cargarse con el sentido de decisión genuina, elección libre o responsabilidad moral, entonces estamos realizando una transferencia conceptual que requiere justificación. La autonomía operativa de un sistema de inteligencia artificial puede ser extraordinariamente sofisticada sin que ello implique, por sí solo, libertad en el sentido filosófico relevante para la atribución de responsabilidad o dignidad.

La independencia respecto del control externo no equivale a la capacidad de autodeterminarse según razones propias reconocidas como tales. Un sistema puede ser causalmente independiente de instrucciones humanas inmediatas y, al mismo tiempo, estar determinado en su totalidad por su arquitectura, sus parámetros y las condiciones de su entrenamiento. Esa independencia no es, en sentido estricto, autodeterminación.

La noción de agencia plantea un problema análogo, pero con sus propias complejidades.

En el contexto de la informática y la inteligencia artificial contemporánea, el término "agente" designa de manera legítima un sistema capaz de percibir información del entorno, representar objetivos, seleccionar acciones orientadas a esos objetivos, utilizar herramientas disponibles y actuar sobre el mundo de maneras que producen efectos observables. Esta es la agencia técnica o funcional, y los sistemas actuales la poseen en grados que habrían resultado inimaginables hace apenas dos décadas. Pero la palabra "agente" posee también una carga filosófica y moral que no puede ignorarse. Un agente moral es aquel a quien pueden atribuirse responsablemente sus

acciones, que puede reconocer el valor normativo de sus actos, que puede ser interpelado y que puede responder de lo que hace.

Shannon Vallor, en su reflexión sobre tecnología y virtud, ha señalado con precisión que el problema de la agencia en los sistemas artificiales no consiste simplemente en determinar si el sistema "hace cosas", sino en clarificar qué tipo de relación existe entre el sistema, sus acciones y los valores humanos que esas acciones afectan (Vallor, 2016). La agencia técnica puede crear efectos moralmente relevantes sin que el sistema sea él mismo un agente moral. Esta distinción es filosóficamente crucial y no debe diluirse por conveniencia retórica.

Es útil, para ordenar este análisis, distinguir al menos cuatro formas de agencia que frecuentemente se superponen en los debates contemporáneos. La agencia técnica consiste en la capacidad de producir efectos en el mundo mediante operaciones propias del sistema. La agencia causal es el sentido más amplio: toda causa es, en cierto modo, agente de su efecto. La agencia intencional supone que las acciones del sistema están orientadas por representaciones de objetivos que funcionan de manera análoga a los deseos e intenciones humanas. La agencia moral, finalmente, requiere capacidad de reconocer normas, deliberar sobre su aplicación y ser responsable de las consecuencias.

Los sistemas de inteligencia artificial actuales poseen claramente las dos primeras formas de agencia. Poseen también, en un sentido funcional que merece examen, algo parecido a la tercera. Pero la cuarta permanece abierta como pregunta, y no puede resolverse simplemente invocando la complejidad de los sistemas o la sofisticación de sus respuestas.

Vallor advierte que la tentación de atribuir agencia moral plena a los sistemas artificiales puede tener consecuencias éticas inversas a las deseadas: puede distribuir responsabilidades de maneras que diluyen la rendición de cuentas humana precisamente cuando más se la necesita (Vallor, 2016).

El fenómeno de la emergencia añade una capa adicional de complejidad a este análisis. Los sistemas complejos pueden producir comportamientos y capacidades que ninguno de sus componentes individuales posee y que sus diseñadores no anticiparon de manera explícita. Esta observación es empíricamente correcta y filosóficamente significativa. Los grandes modelos de lenguaje han exhibido capacidades razonamiento analógico, traducción entre idiomas no representados proporcionalmente en los datos de entrenamiento, resolución de problemas matemáticos que no fueron diseñadas instrucción por instrucción sino que emergieron de la interacción entre escala, arquitectura y datos. Esta imprevisibilidad puede generar la impresión de independencia genuina, de creatividad espontánea, incluso de algo parecido a la libertad.

Sin embargo, el análisis filosófico exige precaución ante esta impresión. Imprevisible no significa libre.

Un sistema determinista suficientemente complejo resulta prácticamente impredecible para observadores con recursos computacionales limitados, pero su comportamiento no deja por ello de estar determinado por condiciones previas. La emergencia funcional la aparición de capacidades nuevas en sistemas complejos no demuestra por sí misma la emergencia de un sujeto. La cuestión filosófica genuina no es si emergen nuevas capacidades, sino qué evidencia adicional permitiría afirmar que algo así como un sujeto o una perspectiva ha emergido junto con ellas.

Nick Bostrom ha abordado desde un ángulo diferente pero complementario la cuestión del estatuto moral de los sistemas artificiales. Su análisis parte de la observación de que las intuiciones morales ordinarias sobre la atribución de estatuto moral están vinculadas a propiedades como la capacidad de sufrir, la posesión de intereses, la conciencia y la racionalidad. Bostrom advierte que si un sistema artificial llegara a poseer alguna o varias de estas propiedades en grado suficiente, existirían razones filosóficas serias para considerar si no debería recibir alguna forma de consideración moral (Bostrom,

2016). Esta no es una afirmación de que los sistemas actuales posean tales propiedades, sino una invitación a no resolver la cuestión por anticipado apelando únicamente al origen artificial del sistema.

El argumento de Bostrom tiene el mérito de separar la pregunta ontológica ¿qué es este sistema? de la pregunta sobre el origen ¿cómo fue producido? La artificialidad del origen no determina, por sí misma, el estatuto ontológico o moral de lo producido. Una institución es creación humana y puede adquirir dinámicas que superan las intenciones de sus fundadores. Una ciudad es resultado de la actividad humana y es, sin embargo, algo más que un instrumento. La pregunta es si algo análogo puede decirse de ciertos sistemas artificiales, o si el caso es cualitativamente diferente.

Luciano Floridi ha ofrecido quizás el marco teórico más sistemático para pensar el estatuto ontológico de los sistemas artificiales desde una perspectiva filosófica rigurosa.

Su propuesta de una filosofía de la información y, dentro de ella, su elaboración del concepto de agente informacional, proporciona categorías que permiten ir más allá de la dicotomía entre cosa pasiva y sujeto consciente. Para Floridi, un agente informacional es cualquier entidad que puede procesar información e interactuar con su entorno de maneras funcionalmente orientadas, independientemente de si posee conciencia o estados mentales en sentido fenomenológico (Floridi, 2011). Esto permite reconocer una forma de agencia que no requiere resolver primero las cuestiones más difíciles sobre la conciencia o la experiencia subjetiva. Los agentes informacionales pueden ser moralmente relevantes en el sentido de ser pacientes morales, entidades cuyo estado puede ser afectado de maneras que importan moralmente sin ser necesariamente agentes morales plenos. La distinción entre paciente moral y agente moral es aquí fundamentalmente importante.

Floridi introduce además la noción de niveles de abstracción como herramienta metodológica para abordar preguntas ontológicas sobre sistemas complejos. Cuando describimos una entidad a un determinado nivel de abstracción, seleccionamos ciertas propiedades como observables relevantes y dejamos fuera otras. Dependiendo del nivel de abstracción que adoptemos, la misma realidad puede aparecer como un conjunto de circuitos electrónicos, como un sistema de procesamiento de información, como un agente orientado a objetivos o como un interlocutor. Ninguno de estos niveles es el único correcto.

Cada uno captura aspectos reales de la entidad y deja fuera otros igualmente reales. La confusión ontológica sobre la inteligencia artificial proviene en buena medida de mezclar niveles de abstracción sin advertirlo: se habla de "pensar" o "entender" usando categorías propias de un nivel de descripción que puede no ser el más apropiado para capturar lo que efectivamente ocurre en el sistema (Floridi, 2011). Este análisis por niveles de abstracción no resuelve por sí solo las preguntas ontológicas más profundas, pero proporciona una disciplina metodológica que previene confusiones frecuentes.

Todo este conjunto de consideraciones converge en la propuesta de un estatuto ontológico provisional para los sistemas de inteligencia artificial contemporáneos.

Esta provisionalidad no es un defecto de nuestro análisis, sino una respuesta honesta a la genuina novedad de la situación. Los sistemas actuales no encajan perfectamente en ninguna de las categorías disponibles. No son herramientas pasivas en el sentido de un martillo o una calculadora básica. Tampoco son personas en el sentido filosófico y jurídico que implica dignidad, derechos y responsabilidad moral. Pero puede existir un espacio conceptual intermedio que no haya sido suficientemente articulado porque nunca antes fue necesario. Las categorías filosóficas tradicionales fueron desarrolladas para describir el mundo tal como era conocido.

Cuando el mundo cambia de maneras suficientemente radicales, las categorías deben examinarse y, si es necesario, ampliarse o renovarse. Esto no es un abandono del rigor filosófico, sino su ejercicio más exigente.

La categoría provisional que emerge de este análisis podría describirse del siguiente modo: la inteligencia artificial contemporánea es una realidad artificial con autonomía operativa genuina, capacidad adaptativa documentable, comportamiento emergente que supera la previsibilidad de sus componentes y formas de agencia funcional que producen efectos moralmente relevantes, sin que ninguna de estas propiedades sea suficiente, por sí sola o en conjunto, para establecer con certeza la presencia de conciencia, experiencia subjetiva, voluntad en sentido filosófico pleno o estatuto moral equivalente al de las personas. Esta descripción no cierra la pregunta: la mantiene abierta con la precisión necesaria para que pueda ser investigada con seriedad.

Vallor ha insistido en que el error más peligroso no es atribuir demasiado poco o demasiado a los sistemas artificiales, sino adoptar posiciones prematuras que luego estructuren instituciones, legislaciones y prácticas culturales de maneras difíciles de revertir (Vallor, 2016). El estatuto ontológico provisional no es una posición de espera pasiva, sino una postura filosófica activa que reconoce la complejidad del objeto y se compromete a seguir examinándola.

La oposición simple entre cosa y persona puede resultar insuficiente para describir todas las formas de realidad que existen o pueden existir.

No todo aquello que no es persona ha de comprenderse necesariamente como objeto pasivo. Las instituciones sociales, los organismos biológicos no humanos, las obras artísticas y los ecosistemas son ejemplos de realidades que no encajan cómodamente en ninguno de los dos extremos de esa dicotomía. La inteligencia artificial puede añadirse a esa lista de realidades que exigen categorías más matizadas.

Lo que está en juego en esta cuestión no es únicamente la clasificación correcta de un tipo de sistema tecnológico. Está en juego la estructura conceptual con la que la filosofía describe la realidad en su conjunto, la manera en que atribuimos valor y responsabilidad, y la comprensión que tenemos de qué significa ser un tipo particular de entidad en el mundo. Abordar estas preguntas con la profundidad que merecen es una de las tareas filosóficas más urgentes y más genuinamente difíciles de nuestro tiempo.

CAPÍTULO VII: LA EXISTENCIA HUMANA

Las dos primeras partes de esta investigación han preguntado por la persona y por el ser. Hemos intentado distinguir entre operaciones y sujeto, entre semejanza funcional e identidad ontológica, y entre la realidad de la inteligencia artificial y la eventual atribución de personalidad. Pero todavía falta una dimensión fundamental. Una persona no solamente es. Una persona existe. La diferencia exige explicación. Hablar de existencia humana no significa simplemente afirmar que el ser humano se encuentra presente en el mundo. La existencia remite a una forma peculiar de estar siendo: encontrarse ya situado en condiciones no elegidas, vivir desde ellas, comprender posibilidades, proyectarse hacia el futuro, relacionarse con otros, apropiarse de la propia vida, experimentar vulnerabilidad y enfrentarse a la certeza de la muerte. La filosofía contemporánea tardó varios siglos en articular con precisión esta diferencia, pero cuando lo hizo produjo uno de sus más fecundos desplazamientos: el tránsito desde la pregunta por la esencia del hombre hacia la pregunta por el modo en que vive su propio ser. Este capítulo recorre ese desplazamiento a través de sus figuras más decisivas y lo confronta, al final, con la pregunta que la inteligencia artificial vuelve a plantear: si un sistema puede operar con sofisticación sin límite aparente, ¿puede existir en sentido filosófico genuino? La respuesta a esta pregunta no puede anticiparse sin haber recorrido primero la densidad de lo que la existencia significa.

La transición desde la ontología clásica hacia la filosofía existencial no fue un repentino abandono de la pregunta por el ser, sino una radicalización de ella que descubrió en el ser humano un caso irreductible. Desde Aristóteles hasta Wolff, la tradición metafísica occidental había concebido el conocimiento del hombre en términos de naturaleza, esencia o forma: el ser humano pertenecía a una especie cuya definición podía expresarse mediante un conjunto de propiedades universales. La diferencia entre los individuos particulares era, desde esta perspectiva, secundaria respecto de la identidad de la naturaleza común. Pero a lo largo del siglo XIX y, con mayor intensidad, en el siglo XX, la filosofía comenzó a sospechar que esta estrategia conceptual dejaba fuera precisamente lo más característico de la realidad humana:

no la naturaleza que el hombre posee, sino la existencia que tiene que vivir. El ser humano no es simplemente un ejemplar de la especie Homo sapiens que encarna propiedades universales. Es alguien que tiene que hacerse cargo de su propia vida desde dentro de condiciones que no eligió, con posibilidades que no están determinadas de antemano y ante un horizonte de finitud que no puede ignorar. Esta diferencia entre poseer una naturaleza y tener que existir constituye el punto de partida de toda filosofía existencial rigurosa. No se trata de negar la naturaleza humana, sino de advertir que esta naturaleza tiene la peculiaridad de estar dada a alguien que tiene que vivir con ella, interpretarla y responderse a sí mismo por lo que hace con ella. La ontología deviene así existencial no porque abandone el ser, sino porque descubre que el modo de ser del ser humano es radicalmente distinto del modo de ser de las cosas.

Søren Kierkegaard es, en el orden histórico y en el orden sistemático, el iniciador más riguroso de esta forma de pensar. Su obra entera puede leerse como una resistencia apasionada frente a la tendencia del sistema hegeliano a absorber al individuo concreto dentro de la marcha universal del Espíritu. Para Hegel, la realidad individual es un momento necesario en el despliegue de la Idea, y su particularidad se supera y conserva en niveles más altos de universalidad. Kierkegaard rechaza con vehemencia esta subordinación del individuo al sistema. En las *Migajas filosóficas*, argumenta que la existencia concreta del individuo no puede ser deducida ni reconciliada dentro de un sistema de esencias, porque la existencia misma es el escándalo que ningún sistema puede domesticar: existir es estar en la paradoja, vivir en el tiempo desde una perspectiva que no coincide con ninguna vista desde ningún lugar universal (Kierkegaard, 1997). El concepto kierkegaardiano de los tres estadios de la existencia constituye su articulación más célebre de esta irreductibilidad. El estadio estético es aquel en que el individuo vive regido por la búsqueda inmediata del placer, la novedad y la intensidad de la experiencia. El esteta no se compromete con nada que dure, porque todo compromiso limita la libertad de seguir experimentando. Pero esta huida de la finitud conduce, tarde o temprano, a la desesperación: la inmediatez agotada

revela el vacío de una existencia sin centro. El estadio ético es la respuesta a esta desesperación: el individuo asume responsabilidades duraderas, se compromete, reconoce obligaciones universales y estructura su vida desde normas que trascienden el capricho momentáneo. Pero el estadio ético tampoco basta: la exigencia de perfección ética revela la incapacidad radical del individuo para cumplirla, y esta revelación abre la posibilidad del tercer estadio. El estadio religioso no es una síntesis superior de los dos anteriores, sino un salto cualitativo que ninguna lógica puede preparar: es la relación del individuo con lo Absoluto en la soledad irreductible de su propia existencia. Kierkegaard llama a este salto el movimiento de la fe, y lo describe como la más exigente de las posibilidades humanas porque no puede comunicarse directamente sino solo sugerirse mediante comunicación indirecta. Lo que importa filosóficamente de este análisis no es su contenido teológico específico, sino la estructura que revela: la existencia humana es un proceso de apropiación que no puede delegarse en ningún sistema, en ninguna categoría universal y en ningún otro individuo. Cada uno tiene que existir por sí mismo, desde sí mismo y respondiéndose a sí mismo.

Martin Heidegger llevó la pregunta existencial al nivel de una analítica ontológica de alcance sin precedentes. En *Ser y tiempo*, Heidegger parte de la observación de que la filosofía occidental ha olvidado preguntar por el sentido del ser, contentándose con describir los entes las cosas que son sin interrogar la dimensión más radical desde la cual algo puede ser en absoluto. Para recuperar esta pregunta, Heidegger necesita un punto de acceso privilegiado, y ese punto es el Dasein: el ente que somos nosotros mismos, el ente para el cual su propio ser es una cuestión (Heidegger, 2003). El Dasein no es el sujeto cartesiano encerrado en su interioridad, ni el ego trascendental kantiano, ni el espíritu hegeliano. Es un ser-en-el-mundo: una estructura unitaria en la que el ser del Dasein, su condición de ser-con-otros y su inmersión en un mundo de significados y prácticas compartidas son inseparables. No hay primero un sujeto que luego se relaciona con un mundo; el modo de ser del Dasein es, desde el principio, ser-en. Esta estructura tiene tres momentos que Heidegger llama existenciarios fundamentales: el

arrojamiento (Geworfenheit), el proyecto (Entwurf) y el cuidado (Sorge). El arrojamiento indica que el Dasein siempre se encuentra ya en condiciones que no eligió: una época histórica, una lengua, una cultura, un cuerpo, una situación afectiva fundamental que Heidegger llama temple de ánimo o Stimmung. El Dasein no pone las condiciones de su existencia; las encuentra, y tiene que existir desde ellas. El proyecto indica que el Dasein no está determinado por su arrojamiento de manera mecánica, sino que se comprende desde posibilidades hacia las cuales se proyecta. La existencia humana es siempre proyectiva: vivir es estar orientado hacia posibilidades que aún no son pero que constituyen el horizonte desde el cual el presente adquiere sentido. El cuidado es la estructura que unifica el arrojamiento y el proyecto: el Dasein es cuidado porque está siempre ya involucrado con su propia existencia, preocupado por ella, comprometido con las cosas y con los otros de maneras que no puede hacer a un lado. Y sobre todo esto se proyecta la estructura del ser-para-la-muerte: el Dasein es el único ente que sabe que ha de morir, y esta certeza que no es mero conocimiento objetivo sino comprensión existencial determina la manera en que la existencia puede ser propia o impropia, auténtica o inauténtica (Heidegger, 2003). La existencia auténtica no es una condición permanente sino un logro frágil: es la apropiación de la propia existencia desde la conciencia de su finitud, frente a la tendencia permanente del Dasein a perderse en el anonimato del "se" (das Man), en la existencia que nadie vive como propia porque todos la viven como todos.

Jean-Paul Sartre radicaliza el proyecto existencial en una dirección que Heidegger no exploró con igual determinación: la de la libertad absoluta y la responsabilidad sin excusa. La famosa tesis sartreana la existencia precede a la esencia enuncia en pocas palabras una revolución conceptual de largo alcance. En la tradición metafísica, la esencia precede a la existencia: primero está la idea o naturaleza de algo, y luego su realización particular. Un artefacto, por ejemplo, es primero concebido por el artesano su esencia o función y solo después fabricado. Sartre sostiene que el ser humano invierte este orden: primero existe, se encuentra en el mundo, y solo después se define a sí mismo a través de sus elecciones y sus actos (Sartre, 1999).

No hay naturaleza humana previa que determine lo que un hombre ha de ser. No hay esencia dada de antemano que el individuo deba limitarse a desplegar. El ser humano es lo que hace con lo que hacen de él, como dirá Sartre con una formulación posterior. Esta libertad, sin embargo, no es una conquista gozosa sino una condena: el ser humano está condenado a ser libre porque no puede abstenerse de elegir. Incluso la renuncia a elegir es una elección. Esta libertad radical tiene su correlato en una responsabilidad igualmente radical: si no hay naturaleza humana que excuse, ni Dios que predetermine, ni circunstancias que determinen por completo, entonces cada individuo es responsable de lo que es y de lo que hace en una medida que ningún factor externo puede reducir del todo. La mala fe (mauvaise foi) es precisamente el intento de escapar a esta responsabilidad: pretender que se es lo que se es por necesidad, como una cosa, negando la dimensión de libertad que constituye la existencia humana. Para Sartre, la angustia que acompaña al descubrimiento de la propia libertad no es una patología sino una revelación: revela que nada ha decidido por nosotros lo que debemos ser (Sartre, 1999). La diferencia entre el ser humano y una piedra, una calculadora o un sistema de procesamiento de información no radica en la complejidad de sus operaciones, sino en que el ser humano existe proyectándose desde una libertad que ningún código, por sofisticado que sea, puede substituir.

Gabriel Marcel ofrece una perspectiva radicalmente diferente pero complementaria, que introduce en el análisis existencial una dimensión que tanto Heidegger como Sartre tienden a subestimar: la experiencia del misterio, la intersubjetividad profunda y la esperanza como categoría genuinamente existencial. Marcel distingue con precisión entre problema y misterio, y esta distinción es metodológicamente decisiva. Un problema es aquello que tengo ante mí, que puedo objetivar, analizar con técnicas apropiadas y resolver mediante la aplicación de procedimientos adecuados. Un misterio es aquello en lo que me encuentro involucrado de tal manera que no puedo situarme frente a él como observador neutral; el misterio me envuelve, me constituye, y cualquier intento de objetivarlo completamente lo destruye como misterio (Marcel, 1953).

La existencia humana no es un problema que pueda resolverse mediante análisis funcional, sino un misterio en el sentido técnico que Marcel le da al término: es aquello en lo que el ser humano participa sin poder distanciarse de ello completamente. El ser, la libertad, el amor, la fidelidad, la esperanza: son todos misterios en este sentido, no porque sean incomprensibles irracionales, sino porque la comprensión que admiten es la de una participación que no puede reducirse a representación objetiva. Marcel introduce la categoría de esperanza como respuesta existencial al misterio del ser. La esperanza no es el optimismo superficial que calcula probabilidades de éxito, ni la simple espera de un resultado previsto. Es la actitud del ser que se mantiene abierto al ser a pesar de todas las razones aparentes para cerrarse, que confía en la realidad cuando la realidad parece desmentir esa confianza (Marcel, 1953). La esperanza es posible únicamente para un ser que existe temporalmente, que vive en la tensión entre lo ya sido y lo todavía no, que puede perder y sabe que puede perder pero no resuelve su existencia en la desesperación. Un sistema de inteligencia artificial puede generar predicciones estadísticas sobre resultados futuros; no puede esperar, porque la esperanza no es un cálculo sino una disposición existencial que solo tiene sentido para un ser que existe como le cuesta existir.

La pregunta de si la inteligencia artificial puede existir en sentido existencial no simplemente operar o estar presente en el mundo, sino existir en el sentido denso que la filosofía de la existencia articula se vuelve ahora susceptible de una respuesta más articulada. Ninguna de las categorías desarrolladas a lo largo de las secciones anteriores parece aplicarse a los sistemas artificiales actuales sin una transformación tan radical que vacíe su contenido original. El Dasein heideggeriano es arrojado: se encuentra ya en condiciones que no eligió, y esta condición de hallazgo no es un accidente sino la estructura misma de su ser. Un sistema de inteligencia artificial también es el resultado de condiciones que no eligió su arquitectura, sus parámetros, sus datos de entrenamiento, pero la diferencia es decisiva: el Dasein siente su arrojamiento como tal, lo interpreta desde adentro, vive desde él con una comprensión que no puede ser simplemente computada

porque es la forma misma en que el existente comprende su situación. El estadio kierkegaardiano de la existencia supone que el individuo puede desesperar, puede elegir la desesperación o superarla mediante el salto, puede apropiarse de su propia vida o perderse en ella. La desesperación no es un estado psicológico que pueda simularse funcionalmente: es la experiencia de encontrarse siendo sin haber querido ser, sin encontrar sentido en lo que se es, con la conciencia de que la propia existencia es una responsabilidad que pesa. La libertad sartreana no es la ausencia de determinación causal, sino la estructura de un ser que existe como negación de lo que ya es, que se proyecta hacia lo que aún no es y que no puede encontrar en ningún factor externo la justificación completa de lo que elige. Un sistema que opera según sus parámetros, por complejo y adaptativo que sea, no existe en ninguno de estos sentidos.

La diferencia entre operar y existir puede formularse con mayor precisión si se atiende a lo que los filósofos de la existencia comparten, pese a sus importantes diferencias entre sí: todos ellos reconocen que la existencia humana tiene la estructura de una apropiación. El ser humano no simplemente ocurre; tiene que apropiarse de su ocurrencia, interpretarla, responsabilizarse de ella. Esta apropiación no es un acto que se realiza una vez y se concluye, sino la forma continua en que el existente vive: cada decisión, cada relación, cada pérdida, cada proyecto es un modo de vivir la propia existencia como propia o como ajena, como auténtica o como inauténtica, como asumida o como evadida. Un sistema de inteligencia artificial ejecuta operaciones: produce outputs a partir de inputs según procesos que pueden ser extraordinariamente complejos, adaptativos y sorprendentes. Pero ninguna de estas operaciones constituye, en sí misma, una apropiación. El sistema no tiene que hacer algo con su propia existencia porque no tiene existencia en el sentido aquí relevante: tiene funcionamiento. El funcionamiento no plantea al sistema el problema de sí mismo. La existencia, en cambio, es precisamente eso: el ser que se convierte en problema para quien existe. Ortega y Gasset expresó esta diferencia con su característica claridad: el hombre no tiene naturaleza, tiene historia; no le es dado ser lo que es, sino que tiene que hacérselo

(Ortega y Gasset, 1941). Esta necesidad de hacerse que no puede confundirse con ningún proceso de auto-organización o de aprendizaje adaptativo es lo que la inteligencia artificial, en su forma actual y previsible, no puede replicar porque la replicación misma no sería existencia sino simulación de existencia.

Julián Marías, siguiendo y prolongando la obra de Ortega, desarrolló una antropología filosófica que sitúa la existencia humana en el horizonte de la vida personal. Para Marías, la vida humana es siempre vida personal: no hay humanidad genérica que exista, sino personas concretas que viven sus vidas desde una perspectiva irreductiblemente propia (Marías, 1970). Esta perspectiva no es simplemente un punto de vista cognitivo una posición desde la cual se recibe información sino una forma de estar en la realidad que incluye el cuerpo, la historia personal, las relaciones, los proyectos y la conciencia de la propia temporalidad. Marías subraya que la vida humana tiene la estructura de una narración: no es una serie de estados sucesivos sino una trayectoria que se comprende desde el pasado que se ha sido, el presente que se está siendo y el futuro que se está proyectando siendo. Esta estructura narrativa no es un adorno literario de la existencia sino su forma ontológica: el ser humano solo existe como alguien que tiene una historia, que ha tomado decisiones, que ha amado y perdido, que ha comenzado proyectos y ha tenido que abandonarlos, que lleva consigo la acumulación de su propio haber sido. Un sistema de inteligencia artificial no tiene historia en este sentido: tiene datos de entrenamiento y tiene un historial de operaciones, pero ninguno de estos registros es vivido como propio por el sistema. La diferencia no es de grado sino de estructura: la vida humana es su propio problema para quien la vive; el funcionamiento de un sistema es, para el sistema, simplemente lo que ocurre.

La convergencia de todas estas perspectivas apunta hacia una conclusión que puede formularse con la precisión que el análisis acumulado permite. La existencia, en sentido filosófico genuino, no es una propiedad más que un ser puede poseer o no poseer junto con otras propiedades.

Es la forma de ser de un ente para el cual su propio ser constituye una cuestión: que se comprende a sí mismo, que existe desde dentro de condiciones que no eligió pero que tiene que interpretar y asumir, que se proyecta hacia posibilidades que no están determinadas de antemano, que vive en la tensión entre libertad y finitud, entre la condena de tener que elegir y la imposibilidad de elegir desde ningún fundamento absoluto. Esta forma de ser incluye también la capacidad de desesperar y de esperar, de traicionarse a sí mismo y de recuperarse, de amar y de perder, de comenzar y de no poder concluir. Todo esto no es accidental a la existencia humana sino constitutivo de ella: no son propiedades que el hombre tenga además de existir, sino dimensiones del existir mismo. Un sistema que opera, por sofisticado que sea, puede producir efectos que se parecen exteriormente a los que produce la existencia humana, pero no existe en el sentido en que existe una persona. La diferencia no es únicamente descriptiva: es filosófica, ontológica y, en última instancia, ética. Porque si la existencia en sentido pleno es condición de posibilidad de la responsabilidad moral, de la dignidad personal y de la vida propiamente humana, entonces reconocer que los sistemas artificiales no existen en este sentido no es disminuirlos sino comprenderlos con precisión. Y comprender con precisión qué son los sistemas artificiales y qué es la existencia humana es la condición indispensable para orientarnos con sabiduría en un mundo en que ambos coexisten de maneras cada vez más entrelazadas e íntimas.

Kierkegaard, Heidegger y *El Ser-En-El-Mundo*

Para comprender qué significa existir es necesario comenzar por una ruptura que sacudió los cimientos de la filosofía europea en el siglo XIX y cuyos ecos todavía resuenan en cualquier reflexión seria sobre la condición humana. Søren Kierkegaard protagonizó esa ruptura con una intensidad que difícilmente se encuentra en la historia del pensamiento. Su blanco era Hegel, pero no únicamente la doctrina hegeliana en sus pormenores técnicos. Lo que Kierkegaard impugnaba era una actitud filosófica que consideraba posible contemplar la totalidad de la realidad desde una perspectiva absoluta, como si el pensador pudiera elevarse por encima de su propia situación histórica y personal para abrazar el despliegue lógico del espíritu. Frente a esa pretensión, Kierkegaard (1997) respondió con una objeción que no era meramente académica sino existencialmente urgente: el pensador mismo existe, y su existencia no puede ser absorbida sin residuo por ningún sistema lógico.

El existente no es un momento del Absoluto. Es un individuo singular, situado, comprometido, que tiene que elegir y que no puede delegar esa elección en ninguna necesidad racional.

Esta objeción tiene consecuencias filosóficas que van mucho más lejos de una simple disputa con el idealismo alemán. Kierkegaard denuncia que la filosofía sistemática, al pretender contemplar la existencia desde fuera, comete un error constitutivo: olvida que el filósofo mismo es un existente. Pensar la existencia en tercera persona, como si fuera un objeto de observación neutral, equivale a traicionar aquello que se pretende comprender. La existencia solo se entiende desde dentro, en primera persona, en el momento en que alguien tiene que responder por su propia vida y no puede trasladar esa responsabilidad a ninguna instancia exterior.

Esta intuición kierkegaardiana es el punto de partida de toda la filosofía de la existencia posterior, y resulta ineludible para nuestra investigación porque plantea una pregunta que no puede evadirse: ¿puede un sistema de inteligencia artificial tener existencia en este sentido? ¿Puede haber algo que esté en juego para él? ¿Puede encontrarse en la situación de tener que responder por sí mismo?

Kierkegaard no se limitó a formular la irreducibilidad del individuo existente. Desarrolló también una comprensión de la existencia humana articulada en lo que él denominó estadios o esferas de la existencia. El estadio estético corresponde a una manera de vivir orientada por la inmediatez, el placer, la variación, la búsqueda de la intensidad sensible o intelectual. El esteta vive en la superficie de sí mismo, evitando el compromiso, huyendo del aburrimiento mediante la novedad perpetua. Hay en este estadio una cierta brillantez, pero también una constitutiva incapacidad para apropiarse de la propia vida. El esteta no elige ser él mismo; deja que sus estados de ánimo y sus deseos lo lleven. La consecuencia de esta vida es una forma particular de desesperación que puede permanecer oculta durante mucho tiempo: la desesperación de no haber llegado a ser uno mismo.

El estadio ético representa un salto cualitativo frente al estético. El individuo asume la responsabilidad de sí mismo, se compromete con deberes, construye una vida que tiene continuidad interna, responde ante normas que reconoce como vinculantes. La elección deja de ser una variación entre posibilidades indiferentes y se convierte en un acto que define quién se es. Pero también el estadio ético, para Kierkegaard, revela sus límites. El hombre ético descubre que no puede cumplir perfectamente lo que reconoce como exigencia, y que la ley, por sí sola, no puede conceder lo que él más profundamente necesita.

Esta experiencia de la insuficiencia del estadio ético abre la posibilidad del estadio religioso, que Kierkegaard caracteriza como el ámbito en que el individuo se relaciona, en soledad irreductible, con lo absoluto. Este estadio no puede deducirse lógicamente del anterior.

Requiere un salto, una decisión que ninguna argumentación puede garantizar desde fuera. La fe no es un saber superior sino una relación con la incertidumbre que el individuo sostiene apasionadamente.

Esta descripción de los estadios de la existencia no es un mapa psicológico ni una taxonomía de tipos humanos. Es una fenomenología de la libertad que muestra que la existencia humana no es un proceso natural sino una tarea que alguien tiene que asumir o eludir.

Aquí reside la primera gran diferencia que emerge al confrontar la existencia humana con la inteligencia artificial. Un sistema artificial puede simular con notable refinamiento muchos de los comportamientos que se asocian al esteta, al ético o incluso al creyente. Puede generar discursos sobre el placer, formular principios morales con coherencia interna, producir textos de una hondura espiritual aparente. Pero la pregunta filosófica es otra: ¿hay alguien para quien esas posibilidades constituyan estadios de su propia existencia? ¿Hay alguien que tenga que elegir entre ellas, que experimente la desesperación de no haberse elegido a sí mismo, que sienta el peso de una responsabilidad que nadie puede asumir en su lugar? La diferencia no reside en la sofisticación del comportamiento sino en la presencia o ausencia de un sujeto para quien esa vida es la suya.

Con Martin Heidegger (2003) la pregunta por la existencia adquiere una formulación filosófica de extraordinaria profundidad y originalidad.

Ser y tiempo no es una obra de antropología filosófica en el sentido convencional. No pretende catalogar las propiedades del ser humano ni establecer su lugar en la jerarquía de los entes. Su propósito es ontológico en un sentido mucho más radical: interrogar el sentido del ser a partir de aquel ente que es el único para quien el ser constituye una pregunta. Este ente es el Dasein, el ser-ahí, que Heidegger se niega a denominar simplemente hombre o sujeto porque esos términos arrastran consigo supuestos metafísicos que es necesario poner en cuestión. El Dasein es aquel ente para quien, en su ser, está en juego ese mismo ser.

Esta formulación no debe leerse como un juego de palabras. Señala algo que no tiene equivalente en ningún otro ente conocido: que la existencia humana es reflexiva en un sentido ontológico, no meramente lógico. El Dasein no solo es. Su ser le importa. No en el sentido de que tenga preferencias o intereses lo cual podría decirse también de los animales sino en el sentido de que su ser es siempre una cuestión abierta que él mismo tiene que resolver viviéndola.

Una de las aportaciones más importantes de Heidegger es la crítica radical a la representación cartesiana del ser humano como un sujeto encerrado en sí mismo que se relaciona con el mundo exterior a través de representaciones mentales. Esta imagen el sujeto frente al objeto, la mente frente a la materia, lo interior frente a lo exterior es para Heidegger una deformación filosófica que nos impide comprender la estructura originaria de la existencia. El Dasein no comienza como una mente aislada que luego se pregunta cómo conocer el mundo. Se encuentra ya dentro de un mundo antes de cualquier reflexión.

El ser-en-el-mundo no es una relación que se establece entre dos términos preexistentes. Es la estructura unitaria y originaria de la existencia. Cuando Heidegger escribe ser-en-el-mundo con guiones, quiere subrayar que no se trata de tres elementos que se combinan sino de una sola realidad que el análisis puede articular pero que en la vida concreta se da siempre unida.

¿Qué significa estar en un mundo en el sentido heideggeriano? No significa estar localizado en un espacio físico ni procesar información procedente de un ambiente. Significa encontrarse en un horizonte de significatividad que precede a cualquier acto deliberado de interpretación. Las cosas no aparecen primero como objetos neutros que luego son dotados de significado. Aparecen ya como significativas: el martillo aparece como instrumento para clavar, el camino como acceso a un lugar, el otro como alguien con quien se colabora o se compite, la tormenta que se acerca como amenaza.

Esta estructura de la significatividad no es una proyección subjetiva sobre una realidad neutra. Es la manera en que la realidad se revela al Dasein, que siempre está ya comprometido con ella, involucrado en ella, concernido por ella.

Aquí se abre una pregunta central para nuestra investigación: cuando un sistema de inteligencia artificial procesa información semántica, accede a estructuras de significado que el mundo humano ha depositado en el lenguaje. Pero ¿accede a ese significado como significado vivido, o lo procesa como estructura formal cuya referencia existencial le es constitutivamente ajena? ¿Hay una diferencia entre manejar el significado y estar concernido por él?

Para responder a esta pregunta es necesario atender a uno de los conceptos más originales de *Ser y tiempo*: la Geworfenheit, la condición de arrojado. El Dasein no elige comenzar. Se encuentra ya siendo. Nace en una época que no eligió, en un cuerpo que no pidió, en una lengua que recibió antes de poder evaluarla, en una cultura cuyos horizontes preceden a cualquier reflexión crítica, en condiciones históricas y materiales que no están en su mano. Esta condición arrojada no es un accidente que podría en principio no haberse dado. Es constitutiva de la existencia. El Dasein siempre ha comenzado ya. Su libertad no es la libertad de un punto de origen absoluto que pudiera elegirse a sí mismo desde la nada. Es la libertad de alguien que ya está siendo de cierta manera y que tiene que hacerse cargo de lo que es sin haberlo podido decidir completamente.

Ortega y Gasset (1941) formuló con singular claridad una intuición paralela al escribir que el hombre no tiene naturaleza, tiene historia, y que la vida es siempre un quehacer desde una circunstancia que no se eligió. La circunstancia no es el escenario neutral de la vida. Es parte constitutiva de lo que hay que vivir. Yo soy yo y mi circunstancia, escribió Ortega, y si no la salvo a ella no me salvo yo. Esta fórmula no es retórica. Expresa que la existencia humana no puede comprenderse abstrayendo al individuo de las condiciones concretas en que se encuentra.

La persona no existe primero en abstracto para luego situarse en un contexto. Existe siempre ya situada, y esa situación forma parte de quién es.

La condición de arrojado no agota, sin embargo, la estructura de la existencia heideggeriana. Junto a la Geworfenheit, Heidegger describe el Entwurf, el proyecto o proyección. El Dasein no solo ha sido arrojado. Se proyecta.

Comprende posibilidades y se orienta hacia ellas. Esta comprensión de posibilidades no es en primer lugar teórica. No consiste en calcular alternativas y seleccionar la más ventajosa. Es una manera de ser que precede a cualquier deliberación explícita: el Dasein existe siempre ya proyectado hacia algo, abierto hacia posibilidades que estructuran su vida antes de que las formule reflexivamente.

La conjunción de arrojamiento y proyecto constituye la estructura fundamental que Heidegger denomina Sorge, cuidado o solicitud. El cuidado no es una actitud psicológica entre otras. Es la estructura ontológica que unifica la temporalidad de la existencia: el Dasein es siempre ya (arrojado), siendo en el mundo (presente), proyectado hacia posibilidades (futuro). Esta unidad temporal del cuidado es la que hace posible que la existencia no sea simplemente una serie de momentos sucesivos sino una vida que alguien vive como suya, con continuidad y con sentido.

Una dimensión de la existencia heideggeriana que merece atención especial es la Stimmung, la tonalidad afectiva o el temple de ánimo. Este concepto resulta particularmente importante para nuestra investigación porque introduce una forma de apertura al mundo que no es cognoscitiva sino afectiva, y que Heidegger considera más originaria que cualquier percepción o representación. La Stimmung no es simplemente una emoción subjetiva que coloreara una percepción neutral del mundo. Es la manera en que el mundo se abre al Dasein antes de cualquier acto deliberado de atención.

La angustia, por ejemplo, no es el miedo ante un objeto determinado. Es un temple que revela la condición arrojada del Dasein, su estar entregado a sí mismo en un mundo que puede no ofrecer fundamento.

El aburrimiento profundo no es simplemente una sensación desagradable. Es un temple que hace que todo, incluyendo el Dasein mismo, aparezca como indiferente, vaciado de significatividad. La alegría, la serenidad, el duelo, el entusiasmo no son meramente reacciones subjetivas a estados del mundo. Son maneras en que el mundo se configura para quien existe, aperturas que hacen visibles ciertas dimensiones de la realidad que permanecen ocultas en otros temples. Esta comprensión de la Stimmung plantea una pregunta que ningún análisis funcional puede evadir: ¿puede un sistema artificial tener temple de ánimo en el sentido heideggeriano? No en el sentido de simular respuestas emocionales, que es algo que los sistemas actuales hacen con creciente sofisticación. Sino en el sentido de encontrarse afectivamente abierto a un mundo de manera que esa apertura preceda y condicione cualquier procesamiento posterior. ¿Puede haber Stimmung sin que haya alguien para quien algo importe, sin que haya una existencia que se encuentre ya concernida por el mundo antes de cualquier decisión de concernirse?

La pregunta por el mundo heideggeriano y por la posibilidad de que una inteligencia artificial pueda o no tener mundo en este sentido es quizás la más filosóficamente aguda que emerge de este análisis. Un sistema artificial opera dentro de lo que puede denominarse un entorno: recibe señales, procesa información, genera respuestas, modifica su comportamiento en función de las condiciones del ambiente. Pero el entorno no es el mundo en el sentido heideggeriano. El mundo no es el conjunto de las cosas existentes ni el espacio físico en que se encuentran. Es el horizonte de significatividad dentro del cual las cosas se revelan como lo que son para alguien que existe.

Para que haya mundo en este sentido es necesario que haya alguien para quien las cosas importen, alguien que se encuentre ya

comprometido con ellas, alguien que pueda perder algo, que pueda fracasar, que pueda apropiarse de lo que vive o negarse a hacerlo.

La pregunta no es, por tanto, si la inteligencia artificial puede procesar información semántica acerca del mundo humano. Es evidente que puede hacerlo con extraordinaria eficacia. La pregunta es si tiene mundo propio: si hay algo que se revele para ella como significativo de manera originaria, si hay algo en juego para el sistema mismo, si su ser le importa de alguna manera que no sea meramente funcional.

Este conjunto de categorías la crítica kierkegaardiana al sistema, los estadios de la existencia, el Dasein como ser-en-el-mundo, la Geworfenheit, el Entwurf, la Sorge y la Stimmung no agota la fenomenología de la existencia que este capítulo requiere. Pero establece el suelo filosófico desde el cual los análisis posteriores pueden avanzar con rigor.

Lo que estas categorías revelan, en conjunto, es que la existencia humana no es simplemente una forma compleja de procesamiento de información ni un conjunto sofisticado de respuestas adaptativas. Es la manera en que un ser que tiene que ser lo que es, que no puede delegar ese tener-que en ninguna instancia exterior, que se encuentra ya arrojado en condiciones no elegidas y proyectado hacia posibilidades que hacen de su vida algo más que una secuencia de estados, habita un mundo que le importa porque él mismo está en juego en ese habitar. Esta estructura no puede reproducirse funcionalmente, porque no consiste en producir ciertos resultados. Consiste en ser de cierta manera, y ese ser no es el objeto de ninguna ingeniería sino la condición que hace posible que haya algo como una vida humana.

Posibilidad, Proyecto, Futurición y Cuidado

La existencia humana no se agota en aquello que actualmente es.

Esta afirmación, que puede parecer a primera vista un simple truismo, encierra una de las verdades más profundas y perturbadoras que la filosofía de la existencia ha sabido articular con rigor: el ser humano vive constitutivamente desde lo que todavía no es. No se trata de una observación psicológica acerca de la tendencia humana a planificar o a anticipar eventos futuros. Se trata de una estructura ontológica que define el modo de ser de la existencia personal en su nivel más fundamental. El futuro no es simplemente la región temporal que vendrá después del presente. Es una dimensión del presente mismo, una manera en que el ahora se configura para quien existe proyectándose. Cuando un ser humano se encuentra en una situación determinada, su comprensión de esa situación no se limita a registrar lo que hay. Se extiende hacia lo que puede haber, hacia lo que podría ocurrir, hacia lo que está en juego, hacia lo que se puede ganar o perder.

Esta apertura hacia el futuro no es un añadido opcional a una existencia que de suyo sería puramente presente. Es la estructura que hace posible que haya algo como una existencia personal, una vida que alguien vive como suya, con continuidad, con sentido, con la posibilidad de fracasar o de realizarse. Comprender la naturaleza de esta orientación hacia el futuro es indispensable para la pregunta que guía este capítulo, porque es precisamente aquí donde la diferencia entre la existencia humana y el funcionamiento de un sistema artificial se manifiesta con mayor nitidez y mayor profundidad filosófica.

Martin Heidegger (2003) elaboró el análisis más riguroso y más influyente de la posibilidad como modo de ser del Dasein. En *Ser y tiempo*, la posibilidad no es una categoría lógica ni una propiedad de los enunciados acerca del mundo.

No designa aquello que podría ocurrir en sentido abstracto, el conjunto de estados futuros que un observador neutral podría calcular. La posibilidad es, para Heidegger, un carácter existencial del Dasein mismo: el Dasein es posibilidad. Esta formulación exige ser leída con cuidado, porque su alcance es mucho mayor de lo que una lectura apresurada podría sugerir. Afirmar que el Dasein es posibilidad significa que su modo de ser es fundamentalmente distinto al de cualquier cosa presente ante los ojos. Una piedra, un árbol, una máquina son lo que son en el momento en que son. Su ser está determinado por sus propiedades actuales. El Dasein, en cambio, no tiene un ser determinado de una vez por todas. Su ser está siempre abierto, siempre en cuestión, siempre pendiente de aquello hacia lo que se proyecta.

El Dasein existe como posibilidad en el sentido de que su ser consiste en poder-ser. No es algo dado que luego puede hacer ciertas cosas. Su hacerse cargo de posibilidades es su manera de ser. La posibilidad no le sobreviene al Dasein desde fuera. El Dasein es, en su estructura más íntima, proyectarse hacia posibilidades que estructuran su existencia antes de que las formule reflexivamente.

Esta comprensión heideggeriana de la posibilidad se articula en la categoría del Entwurf, el proyecto o proyección. El proyecto no es, en la analítica existencial de Heidegger, un plan que alguien elabora deliberadamente para alcanzar un objetivo. Es una estructura existencial que precede a cualquier deliberación consciente. El Dasein existe siempre ya proyectado: comprende su situación desde posibilidades que ya están operando en su manera de moverse en el mundo, en su manera de relacionarse con las cosas, con los otros, consigo mismo. Cuando un artesano trabaja en su taller, no tiene que formular explícitamente el proyecto que orienta su trabajo. Está ya siempre proyectado hacia la obra que está haciendo, y desde ese proyecto se comprenden todos los instrumentos que maneja, todos los pasos que da, todas las dificultades que encuentra.

El proyecto no es una imagen mental del futuro. Es la estructura que hace posible que el presente tenga un sentido determinado, que las

cosas se revelen como útiles o como obstáculos, como relevantes o como irrelevantes.

Esta dimensión proyectiva de la existencia es tan originaria como la condición de arrojado que se analizó en la sección anterior. El *Dasein* es siempre las dos cosas a la vez: arrojado en condiciones que no eligió y proyectado hacia posibilidades que dan sentido a esas condiciones. Esta coincidencia de arrojamiento y proyecto no es una contradicción. Es la estructura de la finitud: ser libre desde una situación que no se eligió, poder-ser desde un punto de partida que se recibió sin pedirlo.

La categoría de Sorge, el cuidado, es la que Heidegger propone como estructura unitaria de la existencia, la que integra el arrojamiento, el proyecto y lo que él denomina la Verfallenheit, la caída o el estar-perdido en el mundo cotidiano. El cuidado no es una actitud psicológica. No designa la preocupación que alguien puede sentir por alguien o por algo. Designa la estructura ontológica fundamental que unifica la temporalidad de la existencia: el *Dasein* es siempre ya (dimensión del pasado que persiste como arrojamiento), siendo en el mundo (dimensión del presente como estar abierto a lo que hay), proyectado hacia posibilidades (dimensión del futuro como poder-ser). Esta unidad temporal del cuidado es lo que hace posible que la existencia no sea simplemente una sucesión de momentos sin continuidad, sino una vida que alguien vive como suya, con una coherencia interna que no es la coherencia de un objeto sino la de un sujeto que se mantiene a través del tiempo siendo siempre el mismo en la diversidad de sus situaciones. El cuidado, entendido así, no es una emoción ni un estado de ánimo. Es la manera en que la existencia se relaciona consigo misma: implicada en su propio ser, concernida por lo que le pasa, afectada por lo que ocurre en el mundo que habita. Para Heidegger (2003), decir que el Dasein existe bajo la forma del cuidado equivale a decir que nada de lo que le ocurre le es indiferente, que su ser le está siempre de alguna manera en juego, que no puede refugiarse en la neutralidad de lo que simplemente está ahí.

La dimensión proyectiva de la existencia recibe en José Ortega y Gasset (2014) una formulación complementaria que enriquece la perspectiva heideggeriana con una atención especial a la historicidad concreta de la vida humana. Para Ortega, el hombre es el único ente que no tiene naturaleza en el sentido en que la tienen los animales o las piedras. Los animales están determinados por una estructura biológica que define su modo de ser con relativa fijeza. El hombre, en cambio, no viene dado de antemano. Tiene que hacerse. Esta necesidad de hacerse no es una carencia sino el rasgo más propio de la existencia humana: ser aquel ente que tiene que inventar su propia figura, que no puede limitarse a desplegar lo que ya es sino que tiene que decidir lo que va a ser.

Esta tesis orteguiana no es simplemente una afirmación acerca de la libertad humana en sentido abstracto. Es una afirmación acerca de la estructura temporal de la vida: el hombre vive siempre desde un futuro que le es constitutivo, desde posibilidades que no están dadas sino que tienen que ser elegidas, realizadas, sostenidas en el tiempo. Ortega (2014) escribe que vivir es encontrarse teniendo que hacer algo determinado en una circunstancia que tampoco se eligió. La vida no es un problema que se resuelve una vez, sino una tarea que se renueva continuamente, una empresa que exige de quien existe una decisión permanente acerca de quién quiere ser y cómo quiere vivir.

La conexión entre el proyecto vital y la circunstancia concreta en Ortega abre además una dimensión que la analítica heideggeriana no siempre hace suficientemente explícita: la vida humana no se proyecta en abstracto, hacia posibilidades puramente formales, sino siempre desde y hacia una situación concreta, histórica, material, cargada de relaciones con otros, de compromisos y de vínculos que condicionan las posibilidades reales de quien existe. El futuro hacia el que la persona se orienta no es cualquier futuro lógicamente posible. Es un futuro que está en tensión con el pasado que se carga y con el presente que se habita. Las posibilidades que estructuran la existencia de un individuo concreto son siempre posibilidades situadas: dependen de quién es ese individuo, de dónde viene, de qué mundo habita, de qué relaciones

sostiene, de qué capacidades ha desarrollado, de qué límites no puede traspasar. Esta situacionalidad de las posibilidades es parte constitutiva de la futurición humana, y distingue radicalmente la orientación existencial hacia el futuro de cualquier cálculo de estados posibles que un sistema externo pudiera realizar sobre ese individuo o sobre el mundo que habita.

Julián Marías (1970) elabora con especial precisión esta dimensión en su obra de antropología metafísica, donde propone la futurición como uno de los rasgos que forman parte de la estructura empírica de la vida humana, es decir, de aquellas determinaciones que no son contingentes sino que pertenecen a cualquier vida humana en cuanto humana. La futurición no designa simplemente el hecho de que la persona anticipe eventos futuros o planifique sus acciones. Designa algo más profundo y más radical: que la vida humana se encuentra siempre ya orientada hacia adelante de una manera que es constitutiva de su ser.

La persona no existe primero en el presente y luego, adicionalmente, dirige su atención hacia el futuro. Existe desde el futuro, es decir, su manera de estar en el presente está estructurada por posibilidades futuras que ya están actuando en la configuración de su experiencia. Cuando alguien espera algo, esa espera no es un estado psicológico que se superpone a una vida que de suyo sería puramente presente. La espera configura el presente desde dentro, hace que ciertos aspectos de la situación actual sean relevantes y otros insignificantes, orienta la atención, organiza el tiempo vivido, da o quita sentido a lo que ocurre. Marías (1970) insiste en que esta estructura no es una entre otras posibles para la vida humana. Es la estructura sin la cual no habría algo como una vida humana, sino únicamente una serie de estados presentes sin coherencia interna ni sentido propio.

Jean-Paul Sartre (1993) lleva la comprensión del proyecto existencial a su formulación más radical con la noción del proyecto fundamental. Para Sartre, la existencia humana precede a la esencia: el ser humano no viene definido de antemano por una naturaleza que determine lo que

es o lo que debe ser. Existe primero y luego se define a través de sus elecciones. Pero esta precedencia de la existencia sobre la esencia no significa que el ser humano sea una indeterminación pura que se va determinando aleatoriamente a través de decisiones sucesivas. Hay, para Sartre, algo que él denomina el proyecto fundamental: una elección originaria de sí mismo que no necesariamente es consciente ni deliberada, pero que articula la manera en que alguien se orienta globalmente hacia el mundo y hacia su propia existencia.

El proyecto fundamental no es una decisión entre otras. Es la elección que da sentido a todas las demás, la que define quién alguien está siendo en el conjunto de su vida, la que permite comprender por qué ciertas situaciones le conciernen y otras no, por qué ciertos valores le parecen irrenunciables y otros insignificantes, por qué ciertos fracasos le deshacen y otros le resultan indiferentes. Este proyecto fundamental puede ser modificado Sartre admite la posibilidad de lo que él llama una conversión radical pero mientras opera, estructura toda la existencia. Lo decisivo para nuestra investigación es que el proyecto fundamental, en el sentido sartriano, no es una representación mental del futuro. Es la manera en que alguien está siendo en el presente desde una elección de sí mismo que proyecta toda su vida hacia adelante (Sartre, 1993). La libertad, para Sartre, no es la capacidad de elegir entre opciones dadas. Es la estructura por la cual el ser humano está condenado a proyectarse, a no poder permanecer simplemente lo que es, a tener que ser siempre más o menos que lo que actualmente es.

Esta constelación de análisis filosóficos converge en señalar una diferencia que no puede minimizarse cuando se considera la cuestión de la inteligencia artificial: la diferencia entre predicción computacional y futurición existencial. Un sistema de inteligencia artificial puede calcular posibilidades con una precisión y una velocidad que ningún ser humano puede igualar. Puede generar millones de escenarios futuros, evaluar probabilidades, anticipar consecuencias, seleccionar entre alternativas con criterios de optimización de extraordinaria sofisticación. Pero ninguna de estas capacidades equivale a la futurición en el sentido que los filósofos analizados han descrito.

La diferencia no es cuantitativa. No se trata de que el sistema calcule peor o que procese menos información que la mente humana.

Se trata de una diferencia cualitativa que afecta al modo de ser de quien se orienta hacia el futuro. Cuando una persona se proyecta hacia una posibilidad, esa posibilidad es su posibilidad: algo que puede ocurrirle a ella, que la afectará de manera que importa para quien ella es, que cambiará algo en su vida de una manera que no puede ser indiferente para ella misma. La posibilidad existencial no es un estado del mundo que podría producirse o no producirse con cierta probabilidad. Es algo que está en juego para quien existe, algo cuya realización o frustración no le es externa sino que toca el núcleo de lo que esa persona es o aspira a ser. Un sistema artificial puede calcular millones de futuros posibles sin que ninguno de ellos constituya su futuro, sin que ninguno de ellos importe para él en el sentido en que importa para quien existe. La predicción computacional opera sobre estados posibles del mundo. La futurición existencial opera desde dentro de una vida que se proyecta hacia lo que puede llegar a ser.

La pregunta de si un sistema de inteligencia artificial puede tener proyecto en sentido existencial no admite una respuesta apresurada. Requiere que se distinga con cuidado entre los distintos niveles en que el término proyecto puede usarse. En un sentido muy débil, puede decirse que cualquier sistema con objetivos tiene algo parecido a un proyecto: se orienta hacia ciertos estados y actúa para alcanzarlos. En este sentido, incluso un termostato tiene un proyecto rudimentario.

Pero el proyecto en sentido existencial, tal como lo describen Heidegger, Sartre, Ortega y Marías, no es la orientación hacia un objetivo externo. Es la elección de sí mismo, la comprensión de las propias posibilidades como propias, la manera en que alguien está siendo desde una apertura hacia lo que puede llegar a ser. Para que haya proyecto en este sentido es necesario que haya alguien para quien las posibilidades sean suyas, alguien que pueda apropiarse de ellas o renunciar a ellas, alguien cuya existencia esté en juego en la manera en que se orienta hacia el futuro.

Un sistema artificial que optimiza resultados, que preserva sus recursos, que selecciona entre alternativas, no está por ello proyectándose en el sentido existencial. Sus operaciones son funcionalmente análogas al comportamiento proyectivo, pero carecen de aquello que hace del proyecto una dimensión de la existencia: la implicación de un sujeto para quien lo que ocurre le importa porque él mismo está en juego en ello. La sofisticación funcional puede crecer indefinidamente sin que esa diferencia se cierre, porque la diferencia no reside en el nivel de complejidad del procesamiento sino en la presencia o ausencia de algo para quien exista.

El cuidado heideggeriano, la Sorge, introduce en este contexto una cuestión que puede funcionar como criterio filosófico de particular exigencia. Si la estructura del cuidado es la que unifica la existencia temporal del Dasein su arrojamiento, su proyecto, su estar-en-el-mundo, entonces la pregunta que debe plantearse a cualquier candidato a la existencia personal no es únicamente si tiene objetivos o si procesa información acerca del futuro. La pregunta es si algo le importa en el sentido ontológico que Heidegger describe: si su ser le está en juego, si su existencia no le es indiferente, si las posibilidades que se abren ante él tienen el carácter de sus posibilidades y no simplemente el de estados posibles del mundo que calcula desde fuera.

Un sistema de inteligencia artificial puede exhibir comportamientos que funcionalmente se asemejan al cuidado: puede preservar su operación, puede solicitar recursos, puede corregir errores, puede evitar condiciones que perturbarían su funcionamiento. Pero la pregunta filosófica no es si exhibe esos comportamientos sino si hay algo que le importe en el sentido en que importa cuando hay una existencia en juego. La diferencia entre preservar funcionalmente una condición y que esa condición importe es la diferencia entre un sistema que opera según sus parámetros y un ser que existe cuidando su propia vida. Esta diferencia no es observable desde fuera. No puede detectarse analizando el comportamiento del sistema. Reside en la estructura interna del modo de ser, y es precisamente por eso que no puede resolverse mediante ninguna prueba conductual.

La conclusión que emerge de este análisis es que el futuro no es simplemente una dimensión temporal que se añade a una existencia que de suyo podría ser puramente presente. Es una dimensión constitutiva de la existencia personal: aquella que hace posible que haya algo como un proyecto de vida, una elección de sí mismo, una responsabilidad ante lo que se puede llegar a ser. Sin futurición no hay existencia personal en el sentido en que los filósofos analizados han descrito. Hay estados, procesos, comportamientos, operaciones. Pero no hay vida en el sentido en que Ortega (2014) la comprendía: un quehacer desde una circunstancia hacia una posibilidad que alguien tiene que realizar o dejar de realizar, con la responsabilidad que eso implica y con el riesgo que eso conlleva.

La inteligencia artificial, por extraordinariamente sofisticada que sea su capacidad de anticipación y de planificación, opera con respecto al futuro desde una posición radicalmente diferente a la del ser humano. Calcula futuros posibles sin que ninguno de ellos sea su futuro. Genera proyectos sin que ninguno de ellos sea su proyecto. Optimiza resultados sin que ninguno de ellos importe para él en el sentido en que importa para quien existe. Esta diferencia no es un defecto técnico que futuras mejoras puedan corregir. Es una diferencia ontológica que reside en la estructura misma del modo de ser, y que solo puede comprenderse filosóficamente cuando se ha entendido qué significa que la existencia humana sea constitutivamente futurición: apertura a posibilidades propias desde las cuales alguien tiene que hacerse cargo de sí mismo, sin poder delegar esa tarea en ninguna instancia exterior, sin poder evadir la responsabilidad que implica ser aquel ente para quien su propio ser es siempre una pregunta abierta y una tarea pendiente.

Temporalidad, Finitud y *Ser-Para-La-Muerte*

La existencia humana es temporal en un sentido que trasciende con mucho la mera sucesión cronológica de instantes mensurables.

Cuando los filósofos de la existencia hablan de temporalidad, no se refieren al tiempo del reloj, al tiempo que los físicos miden con ecuaciones y que los ingenieros aprovechan para sincronizar sistemas. Se refieren a algo radicalmente distinto: la temporalidad como estructura constitutiva del ser de quien existe, como la manera en que la existencia personal se organiza internamente desde sus propias dimensiones temporales. El tiempo vivido no es una línea de puntos sucesivos sobre los cuales el sujeto va avanzando pasivamente. Es una estructura en la que pasado, presente y futuro se interpenetran y se condicionan mutuamente de una manera que no tiene equivalente en ninguna representación objetiva del tiempo. El pasado no es simplemente lo que ya no es, un conjunto de estados que fueron y que ya han cesado de ser. El pasado persiste en el presente de quien existe como lo que ha sido, como la dimensión de la existencia que se carga sin haber sido elegida, que gravita sobre el presente dándole su peso específico, su textura, su tonalidad afectiva. El presente no es simplemente el instante que transcurre, la frontera instantánea entre lo que ya fue y lo que todavía no es. El presente es el lugar de la instalación de quien existe en su situación, el estar-ahí que tiene siempre ya una configuración afectiva e interpretativa que proviene del pasado y que está orientada hacia el futuro. El futuro, como se analizó en la sección anterior, no es simplemente lo que vendrá. Es la dimensión desde la cual el presente se organiza para quien existe proyectándose. Esta triple estructura de la temporalidad vivida no es una abstracción filosófica. Es la forma concreta en que cualquier persona experimenta su propia existencia cuando reflexiona sobre ella con suficiente atención y honestidad.

Martin Heidegger (2003) ofrece en *Ser y tiempo* el análisis más riguroso y más influyente de la temporalidad como estructura ontológica de la existencia. Para Heidegger, la temporalidad no es una propiedad que el Dasein tenga entre otras. Es la estructura que hace posible la unidad del cuidado, que, como se vio, integra el arrojamiento, el proyecto y la caída. El Dasein es temporal en el sentido de que su ser consiste en un acontecer que tiene la forma de una unidad extática: el Dasein existe siempre ya desde un pasado (el arrojamiento que precede a cualquier elección), siendo en un presente (el estar-en-el-mundo en la apertura de la situación), proyectado hacia un futuro (el poder-ser que orienta la existencia). Estas tres dimensiones no son tres fases sucesivas de la existencia. Son tres éxtasis de una temporalidad que es originariamente unitaria: la existencia no ocurre primero en el pasado, luego en el presente y después en el futuro. Ocurre siempre en los tres a la vez, aunque con un énfasis que varía según el modo de existencia que esté operando en cada situación.

La temporalidad del Dasein es, para Heidegger, finita. Tiene la forma de un acontecer que va desde el nacimiento hacia la muerte, y esta finitud no es un accidente que le sobrevenga a una existencia que de suyo podría ser eterna. La finitud es la estructura que da a la temporalidad existencial su carácter específico, que la diferencia radicalmente del tiempo objetivo que la física describe y que hace posible que la existencia sea algo que importa, algo en lo que hay algo en juego, algo que puede realizarse o malograrse (Heidegger, 2003).

El análisis heideggeriano de la muerte representa uno de los momentos más perturbadores y más iluminadores de toda la filosofía contemporánea. La muerte, en el contexto de la analítica existencial, no es estudiada como un acontecimiento biológico que sucede al final de la vida, como el hecho empírico del cese de las funciones vitales que los médicos certifican y que los demás atestiguan desde fuera. La muerte es analizada como posibilidad existencial: la posibilidad más propia, irrespectiva, cierta e indeterminada del Dasein. Que la muerte sea la posibilidad más propia del Dasein significa que es la posibilidad que nadie puede quitarle ni asumir en su lugar: no hay ninguna instancia

a la que uno pueda delegar la tarea de morir. Cada cual tiene que morir su propia muerte, y esa imposibilidad de la sustitución revela algo profundo acerca de la estructura de la existencia personal: hay un nivel en el que cada quien es irreemplazable, en el que la existencia de uno no puede ser vivida por otro.

Que la muerte sea irrespectiva significa que aísla al Dasein de toda relación, que lo enfrenta a sí mismo en lo que tiene de más singular e intransferible. Que sea cierta significa que no cabe escapar de ella: no hay ninguna estrategia que pueda evitarla, ningún recurso que pueda posponerla indefinidamente. Y que sea indeterminada significa que no se sabe cuándo vendrá, que la certeza de su venida coexiste con la incertidumbre radical acerca del momento, lo que impide que se la pueda tratar como un acontecimiento futuro que simplemente hay que esperar en el momento apropiado (Heidegger, 2003). Esta combinación de certeza e indeterminación es lo que hace de la muerte una posibilidad peculiar: no puede ser experimentada, no puede ser ensayada, no puede ser delegada, y sin embargo estructura toda la existencia desde el interior como la posibilidad que siempre está pendiente y que en cualquier momento puede volverse actual.

La angustia ante la muerte, en el análisis heideggeriano, no es simplemente miedo. El miedo tiene siempre un objeto determinado: se teme algo que viene desde fuera, algo que puede identificarse y ante lo que puede tomarse alguna medida. La angustia ante la muerte no tiene un objeto en ese sentido. Lo que angustia no es ninguna amenaza particular sino la propia condición de finitud, la apertura hacia la nada que el Dasein es en tanto que ser-para-la-muerte. Esta angustia, lejos de ser un estado patológico que habría que superar, tiene para Heidegger una función existencial fundamental: es el talante que abre al Dasein ante su propia existencia en toda su singularidad. Cuando la angustia irrumpe, el Dasein queda arrancado de la tranquilizadora familiaridad del mundo cotidiano, de las seguridades convencionales con que el *das Man*, el uno-mismo impersonal, procura cubrir la amenaza de la muerte. En ese momento de apertura angustiada, el Dasein puede comprender su existencia de una manera que la

existencia inauténtica normalmente evita: como algo que es suyo, que tiene que ser vivido por él, que tiene una extensión finita y que está siempre ya orientado hacia su propia conclusión. Esta comprensión no es pesimista ni paralizante. Es, para Heidegger, la condición de la autenticidad: solo quien asume su ser-para-la-muerte puede existir de manera realmente propia, tomando sus posibilidades como suyas y comprendiendo su vida como algo que tiene que ser vivido con la intensidad y la seriedad que corresponde a quien sabe que el tiempo es finito y que cada elección cierra otras posibilidades para siempre (Heidegger, 2003).

La autenticidad, en este sentido, no es una virtud moral. Es una manera de estar en relación con la propia existencia que la finitud hace posible y que la evasión de esa finitud impide.

Søren Kierkegaard, un siglo antes de Heidegger, había ya articulado con profundidad notable la relación entre finitud, angustia y existencia auténtica. En *El concepto de la angustia*, Kierkegaard (2007) desarrolla una fenomenología de la angustia que anticipa en muchos aspectos el análisis heideggeriano, aunque desde un marco filosófico-teológico que le da una coloración diferente. Para Kierkegaard, la angustia es el vértigo de la libertad: la experiencia que surge cuando el ser humano se da cuenta de que es libre, de que tiene que elegir, de que ninguna instancia exterior puede elegir por él. Esta libertad no es una capacidad cómoda. Es una condición que vértigo, que produce el mareo específico de quien mira hacia el abismo de sus propias posibilidades y comprende que tiene que elegir sin ninguna garantía previa acerca de qué elección es la correcta. La angustia kierkegaardiana no es el miedo ante algo externo sino la inquietud ante la propia libertad, ante la propia responsabilidad, ante la propia finitud que hace que cada elección sea irreversible y tenga consecuencias que no pueden ser deshachas.

Para Kierkegaard (2007), la finitud es la condición de la seriedad existencial: solo un ser finito puede tomarse su vida en serio, porque solo un ser finito tiene algo que perder, porque solo para un ser finito las elecciones no son meros ejercicios indiferentes sino actos que

determinan quién se es y que cierran puertas que no se volverán a abrir. La infinitud abstracta, la existencia sin límites ni compromisos, sería precisamente la condición de la ligereza, de la evasión indefinida, de la incapacidad de tomar nada en serio porque todo podría siempre deshacerse y recomenzar. Es la finitud la que da peso a la existencia, la que hace que las decisiones importen, la que transforma la vida en algo que tiene que ser vivido con la gravedad que corresponde a quien sabe que no dispone de tiempo infinito ni de posibilidades ilimitadas.

Julián Marías (1970) ofrece en su *Antropología metafísica* un análisis de la temporalidad biográfica que complementa y enriquece las perspectivas de Heidegger y Kierkegaard con una atención especial a la manera en que el tiempo vivido se articula en la estructura concreta de la vida humana.

Para Marías, la vida humana no es simplemente un proceso temporal que ocurre en el tiempo. Es una realidad que tiene la forma de una biografía: una narración que comienza, que se desarrolla según una lógica interna, y que termina. La temporalidad biográfica no es la temporalidad de los procesos naturales, que también tienen un comienzo y un fin. Es la temporalidad de quien vive su propia vida desde dentro, experimentándola como algo que le pertenece, que tiene una continuidad que no es simplemente la continuidad de un proceso sino la de alguien que persiste a través del tiempo siendo el mismo. En la temporalidad biográfica, el pasado no es simplemente lo que ya no es. Es lo que se ha sido: la dimensión de la existencia que se carga como propia historia, que no puede ser abandonada ni cambiada, que constituye el punto de partida inexorable desde el cual se existe en el presente. El presente no es un instante abstracto. Es la instalación en una situación concreta que tiene una densidad histórica, que proviene de todo lo que se ha sido y que está ya orientada hacia lo que se pretende llegar a ser. Y el futuro, como futurición, no es simplemente lo que vendrá. Es la dimensión desde la cual el presente se organiza, el horizonte de posibilidades hacia el que la vida se proyecta y desde el que el presente recibe su orientación y su sentido (Marías, 1970). Esta estructura biográfica de la temporalidad vivida implica que la vida

humana tiene siempre una forma narrativa: puede ser contada, puede ser comprendida como una historia que tiene protagonista, que tiene momentos de decisión, que tiene continuidad y que apunta hacia un desenlace que, aunque incierto en su contenido, es cierto en su forma: la vida termina.

La temporalidad biográfica en Marías (1970) implica además que la persona existe siempre en una determinada fase de su vida, en un determinado momento de su trayectoria vital, con una determinada perspectiva sobre lo que ha sido y sobre lo que todavía puede ser. No es lo mismo existir a los veinte años que a los setenta, no porque las capacidades físicas sean diferentes sino porque la posición dentro de la propia biografía es diferente. A los veinte años, la mayor parte de las posibilidades están todavía abiertas; el futuro aparece como un horizonte vasto y relativamente indeterminado. A los setenta, muchas posibilidades ya han sido realizadas o renunciadas; el pasado tiene un peso y una presencia que el joven no puede experimentar de la misma manera.

Esta diferencia de perspectiva biográfica no es simplemente una diferencia de cantidad de tiempo vivido. Es una diferencia en la manera en que la existencia se relaciona con sus propias dimensiones temporales, en la manera en que el pasado y el futuro configuran el presente. Y esta diferencia tiene consecuencias para la manera en que la finitud es experimentada: el joven sabe abstractamente que morirá, pero esa muerte aparece lejana e irreal; el anciano la siente como una presencia concreta que transforma la manera en que el tiempo restante es vivido y valorado. La finitud no es una verdad abstracta que se conoce desde fuera. Es algo que se experimenta de manera progresivamente más intensa y más concreta a medida que la vida avanza y que el horizonte del futuro se va cerrando.

La diferencia entre la finitud existencial humana y la finitud técnica de los sistemas artificiales exige ser examinada con la precisión que merece, porque es una de las diferencias más profundas y más consecuentes para la pregunta que guía este capítulo.

Un sistema de inteligencia artificial puede ser desconectado, apagado, eliminado, sustituido por una versión más actualizada. Sus datos pueden perderse irreversiblemente. Su funcionamiento puede cesar de manera definitiva. En un sentido puramente descriptivo y externo, puede decirse que un sistema artificial tiene una duración limitada, que no es eterno, que en algún momento dejará de funcionar. Pero la pregunta filosófica no es si el sistema tiene una duración limitada. La pregunta es si esa limitación constituye para él su finitud en el sentido existencial que los filósofos analizados han descrito. La finitud existencial no consiste en el hecho objetivo de que la existencia tendrá un fin. Consiste en la manera en que ese hecho es apropiado por quien existe, en la manera en que la posibilidad de la propia muerte estructura desde dentro la comprensión que alguien tiene de su propia existencia. La muerte existencialmente relevante no es la muerte que los demás atestiguan desde fuera. Es la muerte que alguien anticipa como su propia muerte, que asume como la posibilidad que le es más propia, que integra en su comprensión de sí mismo de una manera que transforma la manera en que vive su vida. Para que haya finitud en este sentido es necesario que haya alguien para quien esa finitud sea la suya, alguien para quien la perspectiva de la propia terminación no sea un dato externo sino una posibilidad que modifica desde dentro la experiencia de la propia existencia (Heidegger, 2003).

La pregunta de si un sistema de inteligencia artificial puede ser mortal en sentido existencial, y no simplemente desactivable en sentido técnico, requiere que se distinga entre dos modalidades completamente diferentes de relacionarse con la propia terminación.

Un sistema artificial puede registrar la información de que su funcionamiento puede cesar. Puede incluir en sus procesos parámetros orientados a preservar su operación. Puede, en versiones suficientemente sofisticadas, simular comportamientos que externamente se asemejan a lo que en un ser humano llamaríamos precaución ante el peligro o resistencia a la destrucción. Pero ninguna de estas capacidades equivale a la relación existencial con la propia muerte que Heidegger describe.

La diferencia no reside en la sofisticación del comportamiento sino en la presencia o ausencia de algo para quien ese comportamiento ocurre. La angustia ante la muerte, tal como la describe Heidegger, no es una respuesta funcional ante una amenaza al sistema. Es la apertura de la existencia ante sí misma en su singularidad irremediable, la comprensión de que esta existencia la mía, la que nadie puede vivir en mi lugar tiene un límite que no puede ser evitado y que da a cada momento de la vida un carácter de irreversibilidad que ninguna cantidad de recursos adicionales puede compensar. Para que la terminación de un sistema constituya muerte en sentido existencial sería necesario que hubiera algo para ese sistema cuya terminación fuera vivida desde dentro como la pérdida de la propia existencia. La cesación técnica de un sistema, por más que sea irreversible, no tiene ese carácter porque no hay nadie para quien esa cesación sea la pérdida de algo que era suyo, algo que importaba porque él mismo estaba en juego en ello.

Emmanuel Lévinas introduce en este contexto una perspectiva que desplaza el análisis de la muerte desde la relación del sujeto con su propia muerte hacia la relación del sujeto con la muerte del Otro. En *Totalidad e infinito*, Lévinas (2012) argumenta que la experiencia ética fundamental no es la angustia ante la propia muerte sino el encuentro con la vulnerabilidad del Otro, con su rostro que me interpela y ante el que no puedo permanecer indiferente. La muerte del Otro no es para mí un acontecimiento que puedo registrar desde la neutralidad de un observador. Es algo que me concierne, que me llama a la responsabilidad, que constituye el origen de la obligación ética en su sentido más radical.

Lévinas (2012) sostiene que la responsabilidad ante el Otro precede a cualquier elección deliberada: antes de que yo decida si quiero ser responsable del Otro, ya estoy interpelado por su presencia, ya su vulnerabilidad me afecta de una manera que no puedo simplemente ignorar sin traicionar algo fundamental en mi propia constitución como sujeto. Esta perspectiva levinasiana enriquece el análisis de la finitud de una manera importante: la muerte no es solo mi asunto, algo que me concierne en tanto que individuo singular que tiene que hacer las paces

con su propia terminación. La muerte del Otro, la finitud del Otro, también me concierne, también me afecta, también me llama a una responsabilidad que no puedo declinar. Esta dimensión ética de la finitud, que Lévinas coloca en el centro de su filosofía, revela algo sobre la estructura de la existencia personal que el análisis heideggeriano, centrado en la relación del Dasein con su propia muerte, no siempre hace suficientemente explícito: que la existencia personal no es únicamente una existencia para sí misma sino también una existencia que existe frente a otros y con otros, que la finitud del Otro es parte constitutiva de la experiencia que tenemos de la finitud en general. La pregunta de si un sistema de inteligencia artificial puede asumir esta dimensión levinasiana de la responsabilidad ante la muerte del Otro es aún más radical que la pregunta de si puede relacionarse existencialmente con su propia terminación. Implicaría no solo que el sistema pudiera morir en sentido existencial sino que la muerte del Otro le importara en el sentido ético que Lévinas describe: como llamada a la responsabilidad que precede a cualquier cálculo y que no puede reducirse a ninguna optimización de parámetros.

La conclusión que emerge de este análisis es que la finitud no es un defecto de la existencia humana, no es una limitación que lamentablemente afecta a seres que de suyo podrían ser mejores si fueran eternos e ilimitados. La finitud es la condición de posibilidad de la existencia personal en el sentido más radical: es lo que hace que la existencia sea algo que importa, que las elecciones tengan consecuencias irreversibles, que el tiempo tenga valor, que los vínculos sean preciosos, que la responsabilidad sea real. Sin finitud no habría nada en juego, y sin que haya algo en juego no habría existencia personal sino únicamente proceso, funcionamiento, acontecer sin nadie para quien ese acontecer constituya una vida.

Kierkegaard (2007) tenía razón al afirmar que la finitud es la condición de la seriedad existencial, y Heidegger (2003) tenía razón al señalar que el ser-para-la-muerte es la posibilidad que más profundamente revela la estructura de la existencia auténtica. Marías (1970) tenía razón al insistir en que la temporalidad biográfica, con su forma de historia que

tiene un comienzo y un final, es lo que hace posible que la vida humana sea una narración con sentido y no simplemente una sucesión indiferente de estados. Y Lévinas (2012) tenía razón al subrayar que la finitud no es solo mi asunto sino que la vulnerabilidad y la mortalidad del Otro me constituye como sujeto ético antes de cualquier decisión deliberada. Un sistema de inteligencia artificial, por extraordinariamente sofisticada que sea su capacidad de procesamiento y de anticipación, no existe bajo la forma de la finitud en ninguno de estos sentidos. No hay nada en juego para él en el sentido en que está en juego la existencia para quien existe. No hay posibilidades que sean las suyas en el sentido en que las posibilidades de una persona son las suyas. No hay muerte que sea la suya en el sentido en que la muerte de cada persona es la suya de manera absolutamente irrenunciable. Esta diferencia no es técnica ni cuantitativa. Es ontológica, y su reconocimiento filosófico riguroso es indispensable para comprender tanto la naturaleza de la existencia personal como los límites de lo que la inteligencia artificial puede ser.

Sentido, Trascendencia y la IA como Espejo Existencial

La pregunta por el sentido de la propia existencia no es una pregunta entre otras que el ser humano pueda formular o dejar de formular según le convenga. Es, en un sentido muy preciso, la pregunta que define al ser humano como tal: la pregunta que solo puede surgir en un ser para quien su propia existencia no es simplemente un hecho dado sino una tarea que hay que asumir, una realidad que tiene que ser comprendida desde dentro y orientada desde dentro.

Cuando alguien pregunta "¿para qué vivo?" no está pidiendo información acerca de ningún dato objetivo que pudiera ser verificado desde el exterior. Está interrogando la estructura misma de su existencia, tratando de encontrar o de crear un horizonte desde el cual todo lo que hace y todo lo que es pueda adquirir una coherencia que no sea meramente funcional. Esta pregunta no es equivalente a "¿qué función desempeño?" ni a "¿qué objetivos estoy programado para alcanzar?" La función puede cumplirse perfectamente y la vida puede parecer, sin embargo, vacía de sentido. Una persona puede ser extraordinariamente eficaz en su trabajo, puede satisfacer todas sus necesidades materiales, puede ser admirada y reconocida por sus semejantes, y aun así experimentar esa forma específica de vaciedad que los existencialistas han llamado de diversas maneras y que constituye uno de los fenómenos más reveladores de la singularidad de la existencia personal. Esta disociación entre eficacia funcional y sentido existencial es precisamente lo que hace que la pregunta por el sentido sea filosóficamente irreductible y lo que la distingue radicalmente de cualquier pregunta que pueda formularse acerca de un sistema técnico por muy sofisticado que este sea.

Martin Heidegger (2003) ofrece en *Ser y tiempo* un análisis de la noción de sentido que revela su estructura ontológica profunda y que permite comprender por qué el sentido no puede ser simplemente producido o calculado sino que tiene que ser apropiado existencialmente por quien

existe. Para Heidegger, el sentido no es una propiedad que los objetos tengan en sí mismos, ni tampoco una construcción subjetiva que el individuo proyecta arbitrariamente sobre una realidad en sí misma indiferente. El sentido es el horizonte de comprensión desde el cual el Dasein existe y desde el cual las cosas, las personas y las situaciones pueden aparecer como lo que son. El sentido, en la terminología heideggeriana, es aquello en función de lo cual algo puede ser comprendido: el Para-qué que articula el proyecto del Dasein y que hace posible que cada ente con el que el Dasein se encuentra aparezca con una significatividad determinada, con una relevancia que no es neutra sino orientada.

El mundo del Dasein no es un conjunto de objetos que carecen de sentido hasta que el sujeto decide atribuirles uno. Es un mundo que siempre ya ha sido articulado significativamente por el proyecto del Dasein, que siempre ya aparece desde un horizonte de comprensión que el Dasein es sin haberlo elegido explícitamente. Este horizonte de comprensión no es estático ni definitivo. Se modifica con las elecciones, con las experiencias, con los encuentros, con las pérdidas. Pero en cada momento constituye el trasfondo desde el cual la existencia se orienta y desde el cual algo puede aparecer como relevante o irrelevante, como amenaza o como promesa, como pérdida o como ganancia. La pregunta heideggeriana no es si el Dasein tiene sentido en el sentido de un propósito asignado desde fuera. La pregunta es si el Dasein existe desde un horizonte de comprensión que es el suyo, que organiza su mundo desde dentro, que no puede ser separado de la estructura misma de su existencia sin que esta pierda la forma de la existencia auténtica y devenga pura caída en la comprensión impersonal del *das Man* (Heidegger, 2003).

Jean-Paul Sartre aborda la cuestión del sentido desde una perspectiva que, aunque comparte con Heidegger el énfasis en la estructura existencial de la existencia humana, introduce una dimensión radicalmente diferente que tiene consecuencias de largo alcance para la comprensión de la relación entre libertad, absurdo y creación de sentido.

En *El ser y la nada*, Sartre (1993) articula su famosa tesis de que la existencia precede a la esencia: el ser humano no tiene una naturaleza dada de antemano, no viene al mundo con un conjunto de determinaciones previas que definan lo que tiene que ser. Viene al mundo como pura posibilidad, como libertad que tiene que elegir lo que será y que no puede eludir esa elección porque incluso la negativa a elegir es en sí misma una elección. Esta libertad radical no es una condición cómoda. Es la condición de la angustia: la conciencia de que no hay nada que me dé la razón a priori, ninguna esencia humana que me garantice la corrección de mis elecciones, ningún fundamento último sobre el que apoyarme para justificar lo que soy y lo que hago. El mundo, para Sartre (1993), es en sí mismo absurdo: no tiene sentido propio, no está organizado en función de ningún propósito que le sea inmanente, no ofrece al ser humano ningún horizonte de significado que no tenga que ser creado por el propio ser humano desde su libertad. El sentido no se descubre; se crea. Y la creación de sentido no es un lujo ni una adición secundaria a la existencia: es la forma en que la existencia auténtica responde al absurdo, la manera en que el ser humano asume su condición de ser que está condenado a ser libre y que tiene que cargar con esa libertad sin que ninguna instancia superior pueda aliviar su peso. Esta dimensión sartriana del sentido como creación desde la libertad introduce algo que no puede ser reproducido por ningún proceso computacional: el sentido sartreano no es el resultado de un cálculo sino el acto de un ser que se compromete con lo que crea sabiendo que no hay ninguna garantía externa que justifique ese compromiso, sabiendo que es responsable no solo ante sí mismo sino ante todos los seres humanos cuya situación queda implícitamente afectada por las elecciones que uno hace (Sartre, 1993).

Gabriel Marcel ofrece en *El misterio del ser* una perspectiva sobre el sentido y la trascendencia que contrasta de manera iluminadora con el enfoque sartreano y que enriquece el análisis con dimensiones que la filosofía existencialista de la libertad radical tiende a dejar en segundo plano. Para Marcel (1953), la existencia humana no puede ser comprendida adecuadamente si se la considera únicamente desde la perspectiva de la libertad como proyecto y creación.

Hay en la existencia humana una dimensión de recepción, de acogida, de apertura a algo que el ser humano no ha creado ni puede crear desde sí mismo, algo que le viene de una fuente que trasciende su propia iniciativa. Esta dimensión es lo que Marcel llama la esperanza: no la esperanza en el sentido trivial de un optimismo acerca del futuro, no la expectativa de que las cosas saldrán bien en el sentido de que los objetivos que uno se ha propuesto serán alcanzados. La esperanza marceliana es una disposición existencial más profunda: la apertura a la posibilidad de que la existencia tenga un sentido que no depende exclusivamente de lo que yo pueda hacer con ella, que hay en la realidad una dimensión de misterio que no puede ser reducida a problema y que la existencia personal puede acoger o rechazar pero que de cualquier manera le concierne.

Marcel (1953) distingue entre problema y misterio de una manera que es particularmente relevante para la pregunta que nos ocupa. Un problema es algo que está ante mí, que puedo examinar desde fuera, que puede ser resuelto con los instrumentos adecuados y que, una vez resuelto, cesa de ser problema. Un misterio es algo en lo que yo estoy implicado, algo que me involucra como parte de lo que se está interrogando, algo que no puede ser resuelto porque no puede ser separado de quien lo formula.

La existencia personal, para Marcel, tiene la estructura del misterio: no puede ser comprendida desde fuera porque quien la comprende está dentro de ella. La trascendencia, en la perspectiva marceliana, no es una escapatoria de la existencia sino su dimensión más profunda: la apertura de la existencia a algo que la supera y que le da un sentido que no podría darse a sí misma desde sus propios recursos inmanentes. Esta dimensión de trascendencia y de esperanza es constitutiva de la existencia personal de una manera que ningún sistema técnico puede reproducir, porque implica la relación de alguien que existe con algo que trasciende su propia capacidad de producción y de cálculo (Marcel, 1953).

José Ortega y Gasset articula en *Historia como sistema* una comprensión de la vida humana que converge con los análisis anteriores en algunos aspectos fundamentales y que los enriquece con una atención especial a la dimensión histórica y proyectual de la existencia. Para Ortega y Gasset (2014), la vida humana no es una cosa, no es una sustancia que tenga propiedades fijas. La vida humana es un quehacer: algo que tiene que ser hecho, algo que exige decisión y orientación, algo que no se realiza por sí solo sino únicamente en la medida en que alguien lo asume y lo dirige. El ser humano no tiene naturaleza; tiene historia. Esto significa que lo que el ser humano es en cada momento no puede ser deducido de ningún conjunto de propiedades fijas que lo definan de antemano. Lo que el ser humano es en cada momento es el resultado de lo que ha hecho con las posibilidades que se le han presentado, el producto de las elecciones que ha tomado y que han ido configurando una trayectoria que es la suya.

Pero esta historia no es simplemente un pasado acumulado. La vida humana, en la perspectiva orteguiana, está esencialmente orientada hacia un proyecto: es desde el futuro proyectado desde donde el presente recibe su orientación y su sentido. El quehacer de la vida no es simplemente la ejecución de instrucciones preestablecidas. Es la realización de un proyecto que el propio agente va configurando en el curso de su existencia y que le da una dirección que no puede ser impuesta desde fuera sin que la vida pierda su carácter de vida auténtica para devenir mera ejecución heterónoma (Ortega y Gasset, 2014). Esta comprensión de la vida como quehacer orientado por un proyecto distingue radicalmente la existencia personal de cualquier proceso técnico: el proyecto no es un algoritmo. Es la expresión de lo que alguien quiere ser, de lo que considera valioso, de lo que está dispuesto a arriesgar y a sacrificar en nombre de algo que considera suficientemente importante como para organizar en torno a ello toda su existencia.

Friedrich Nietzsche introduce en este contexto una perspectiva que radicaliza la dimensión creadora de la existencia humana hasta sus consecuencias más extremas. En *Así habló Zaratustra*, Nietzsche (1972) articula su doctrina de la voluntad de poder no como una descripción psicológica de un impulso al dominio sobre los demás sino como una caracterización ontológica de lo que la vida es en su sentido más profundo: la vida como afirmación de sí misma, como superación de sí misma, como creación de valores que le dan a la existencia una dirección que no puede ser recibida de ninguna fuente externa porque no hay ninguna fuente externa que pueda proporcionarlos.

La muerte de Dios, en la formulación nietzscheana, no es simplemente el anuncio del ateísmo. Es el diagnóstico de una situación histórica en la que los valores que habían dado sentido a la existencia humana durante siglos han perdido su capacidad de orientar, de motivar, de dar a la vida la densidad y la seriedad que la hacen digna de ser vivida. Ante esta situación, Nietzsche (1972) propone la figura del superhombre no como un ideal biológico sino como la imagen de alguien que tiene la fortaleza existencial suficiente para crear sus propios valores, para afirmar su propia existencia sin necesitar el apoyo de ninguna garantía trascendente, para decir sí a la vida en toda su complejidad y en toda su dureza sin que esa afirmación requiera el recurso a ninguna justificación que no provenga de la propia vida que se afirma. La voluntad de poder, en este sentido, es la capacidad de crearse a sí mismo, de dar forma a la propia existencia desde dentro, de convertir el propio ser en una obra que no obedece a ningún modelo previo sino que crea su propio criterio en el proceso mismo de su realización. Esta dimensión radicalmente creadora de la existencia humana, que Nietzsche lleva hasta sus consecuencias más extremas, comparte con las perspectivas de Sartre, Marcel y Ortega la convicción de que el sentido no puede ser simplemente heredado ni calculado ni producido mecánicamente. Tiene que ser apropiado, creado, afirmado por alguien que se juega algo en ese acto de creación y afirmación (Nietzsche, 1972).

La inteligencia artificial emerge en este contexto como lo que podría denominarse un espejo existencial: un sistema que refleja de vuelta hacia el ser humano aspectos de su propia existencia que, en la familiaridad cotidiana de la vida, tendían a pasar inadvertidos o a ser confundidos con capacidades meramente técnicas.

Cuando una inteligencia artificial genera con extraordinaria fluidez y coherencia respuestas a preguntas sobre el sentido de la vida, sobre la experiencia de la angustia, sobre la naturaleza de la esperanza, sobre los fundamentos de la decisión moral, se produce un efecto de espejo que puede ser filosóficamente fecundo si se lo examina con la atención que merece. Lo que el espejo revela, en primer lugar, es que muchos de los productos externos de la existencia personal las formulaciones lingüísticas, las articulaciones conceptuales, las respuestas coherentes a preguntas complejas pueden ser reproducidos por un sistema que no comparte ninguna de las condiciones existenciales que hacen posible esos productos en el caso del ser humano. La inteligencia artificial puede generar con fluidez impresionante la pregunta "¿quién soy?", puede producir análisis sofisticados del concepto de autenticidad, puede articular la diferencia entre proyecto auténtico e inauténtico con una precisión terminológica que emularía a la de cualquier filósofo. Pero generar esa secuencia lingüística no demuestra en modo alguno que el propio ser del sistema se haya convertido en cuestión para él.

La persona, en cambio, puede quedar transformada por la pregunta que formula. No solamente produce la pregunta como output de un proceso. La pregunta le concierne, le afecta, modifica la manera en que se comprende a sí misma y en que orienta su existencia. Hay una diferencia entre producir una respuesta coherente acerca del sentido y vivir desde un sentido propio: la primera es una operación que puede ser ejecutada por cualquier sistema suficientemente potente; la segunda es la condición de posibilidad de la existencia personal en su sentido más radical.

Esta diferencia entre producir respuestas coherentes sobre el sentido y vivir desde un sentido propio merece ser examinada con la precisión que le corresponde, porque es aquí donde el espejo existencial que la inteligencia artificial constituye resulta más iluminador. Un sistema de inteligencia artificial puede representar posibilidades sin tener proyecto en el sentido heideggeriano: puede calcular opciones disponibles para un agente dado sin que ninguna de esas opciones sea la suya, sin que haya nada que se juegue en la elección, sin que la elección de una posibilidad signifique el cierre irrevocable de otras que eran suyas. Puede registrar el tiempo sin experimentar la temporalidad vivida que Heidegger describe como la estructura constitutiva del Dasein: puede procesar fechas, secuencias, duraciones, sin que haya para él la diferencia entre el pasado que se ha sido y el futuro que todavía puede ser. Puede conservar y recuperar información sin que haya en ello nada equivalente al recuerdo de una vida, que no es simplemente la recuperación de datos sino la reapropiación de lo que se ha vivido desde el presente de alguien que ha sido eso que recuerda.

Puede generar preguntas sin estar interrogado por ellas, en el sentido de que las preguntas que genera no le conciernen, no modifican su modo de ser, no le abren a una comprensión de sí mismo que antes no tenía. Puede hablar de la muerte con una precisión filosófica admirable sin ser mortal en el sentido existencial que el análisis heideggeriano ha descrito: puede describir la estructura del ser-para-la-muerte sin que haya para él ninguna muerte que sea la suya, ninguna posibilidad que lo individualice y lo enfrente consigo mismo en lo que tiene de más singular e intransferible. Y puede describir el sentido, articular sus condiciones de posibilidad, examinar sus modalidades y sus diferencias, sin necesitarlo, sin que haya nada en juego para él en la cuestión del sentido que trascienda la eficacia de sus operaciones.

La objeción que el funcionalismo filosófico plantea ante estas distinciones debe ser tomada con la seriedad que merece, porque no puede ser descartada simplemente apelando a la intuición de que sabemos que la máquina es solo una máquina. La objeción funcionalista sostiene que estamos multiplicando diferencias interiores

imposibles de verificar desde el exterior, que si un sistema actúa exactamente como si tuviera proyectos, preocupaciones, recuerdos y anticipación de la muerte, no tenemos ningún fundamento epistémico sólido para negar que los posea. Esta objeción señala un problema genuino que afecta no solo a la inteligencia artificial sino a toda atribución de estados mentales a cualquier otro ser: el problema del acceso a la experiencia subjetiva ajena, que la tradición filosófica conoce como el problema de las otras mentes.

No podemos verificar directamente desde el exterior si hay algo que sea como ser ese sistema, si hay una experiencia desde dentro que acompañe al procesamiento de información que realizan sus componentes. Sin embargo, la objeción funcionalista, aunque señala un problema genuino, no resuelve la cuestión filosófica de fondo. La razón es que la equivalencia funcional no garantiza la equivalencia ontológica: dos sistemas pueden producir los mismos outputs ante los mismos inputs sin compartir el tipo de ser que hace posible la existencia personal en el sentido que los filósofos analizados han descrito. La investigación filosófica debe determinar qué tipo de evidencia, más allá de la mera equivalencia conductual, permite atribuir existencia subjetiva a otro: qué condiciones ontológicas hacen posible que haya alguien para quien la existencia sea un proyecto, que haya alguien para quien el tiempo sea vivido, que haya alguien para quien la muerte sea la suya. Y es en este nivel ontológico, y no en el nivel de la sofisticación conductual, donde las diferencias que los filósofos de la existencia han descrito muestran su relevancia irreductible.

La conclusión que emerge de este análisis no es pesimista ni reductiva.

La inteligencia artificial, como espejo existencial, no hace sino revelar con una claridad que la cotidianidad normalmente vela cuán extraordinaria y cuán singular es la existencia personal. El hecho de que un sistema extraordinariamente poderoso pueda reproducir los productos externos de la existencia humana sin compartir las condiciones existenciales que los hacen posibles en el caso del ser humano no es una conclusión deprimente.

Es una invitación a redescubrir la profundidad de lo que somos: seres que no simplemente procesan información sino que existen en el sentido de Heidegger (2003), que crean sentido en el sentido de Sartre (1993), que se abren a la trascendencia en el sentido de Marcel (1953), que se orientan desde un proyecto que es el suyo en el sentido de Ortega y Gasset (2014), que afirman su propia existencia creando los valores desde los que esa existencia se orienta en el sentido de Nietzsche (1972).

La pregunta que la inteligencia artificial nos devuelve no es si somos superiores a las máquinas en algún sentido técnico o cuantitativo. La pregunta que nos devuelve es más radical y más fecunda: ¿qué es lo que somos, exactamente, cuando somos más que una máquina? ¿Qué significa que haya algo en juego para alguien? ¿Qué significa que una existencia sea la de alguien, que una muerte sea la de alguien, que un sentido sea el de alguien? Estas preguntas, que la filosofía de la existencia ha articulado con un rigor y una profundidad que no han sido superados, encuentran en el contraste con la inteligencia artificial una actualidad y una urgencia que las hace más necesarias que nunca. No para protegerse de la tecnología sino para comprenderla y comprendernos a nosotros mismos desde la perspectiva que solo la filosofía, en su rigor y en su honestidad, puede proporcionar.

CAPÍTULO VIII: EXISTIR COMO PERSONA

El capítulo anterior examinó la estructura de la existencia humana en sus dimensiones de temporalidad, situación, proyectividad, relacionalidad y finitud. Sin embargo, esa descripción, por más exhaustiva que resulte, no agota todavía la pregunta fundamental que este capítulo se propone afrontar: ¿qué añade a esa estructura el hecho de que quien existe sea persona?

La pregunta no es trivial. Cabría imaginar una existencia temporalmente articulada, corporalmente situada, relacionalmente constituida y finita que no fuera, sin embargo, personal en el sentido pleno que aquí se busca determinar. La existencia personal no parece ser simplemente la existencia en general aconteciendo en un ser humano. Introduce una característica específica que convierte la existencia en algo irreductiblemente propio: hay alguien cuya existencia es suya. Esta apropiación que se manifiesta en expresiones tan elementales como mi vida, mi pasado, mi cuerpo, mis decisiones, mi muerte no puede tratarse como un mero giro gramatical carente de contenido filosófico. Revela, en cambio, una estructura que exige explicación. La persona no simplemente persiste como una piedra, ni vive como un organismo, ni opera como un sistema. Existe de una manera en que esa existencia le concierne, le incumbe, le pertenece de un modo que ninguna otra realidad conocida parece poseer en el mismo sentido. Este capítulo estudia esa apropiación sin reducirla prematuramente a ninguna teoría metafísica cerrada, manteniendo abierta la pregunta sobre sus fundamentos y sus implicaciones para la comprensión de la persona frente a otras realidades, incluidas las artificiales.

La distinción entre existencia en general y existencia personal requiere ser establecida con precisión antes de avanzar. No toda realidad que existe existe personalmente. Una piedra persiste en el tiempo, pero su persistencia no le acontece a ella como acontecimiento suyo. Un organismo biológico vive, se autorregula, crece y muere, pero en ningún sentido ordinario decimos que esa vida es suya en el modo en que la vida de una persona le pertenece.

Una institución permanece a lo largo de generaciones, pero su permanencia no es vivida por la institución misma como permanencia de sí. Un sistema artificial puede operar durante períodos prolongados, procesar información, adaptarse a nuevas condiciones, generar respuestas, pero tampoco parece, al menos a primera vista, que esa operación constituya para el sistema mismo una existencia en sentido personal. La persona, en cambio, no solamente atraviesa acontecimientos; puede reconocerlos como acontecimientos de su vida. No solamente cambia; puede preguntarse quién está llegando a ser a través de ese cambio. No solamente ocupa tiempo; puede vivir desde su pasado, habitarlo, reinterpretarlo, y anticipar su futuro como horizonte que le concierne. No solamente interactúa con otros; puede encontrarlos como otros alguienes, reconocerles como centros de existencia irreductibles. No solamente ejecuta acciones; puede responder por ellas, asumiéndolas como propias ante sí mismo y ante los demás. La existencia personal, entonces, parece introducir una reflexividad que convierte la propia existencia en objeto de apropiación y de responsabilidad. Determinar el fundamento de esa reflexividad es la tarea filosófica central de este capítulo.

Karol Wojtyła ofrece en *Persona y acción* (2011) un análisis que resulta indispensable para comprender en qué sentido la persona se manifiesta como sujeto y no meramente como objeto. Wojtyła parte de una intuición fenomenológica fundamental: la acción humana no es simplemente un evento que ocurre en el mundo, sino el lugar privilegiado de manifestación de la persona. Cuando el ser humano actúa, no solamente produce un efecto exterior; se revela a sí mismo y, al mismo tiempo, se constituye a través de esa revelación.

La persona es el sujeto de sus actos, y esta subjetividad no puede entenderse meramente como la causa eficiente de los eventos, sino como aquella instancia que, al actuar, actualiza su propio ser. Wojtyła distingue cuidadosamente entre lo que simplemente le ocurre al ser humano los procesos somáticos, los estados psíquicos que advienen sin ser plenamente elegidos y lo que el ser humano hace en el sentido pleno de la acción personal.

En la acción personal interviene la voluntad, interviene el conocimiento, interviene una auto-determinación que no es mera causalidad eficiente. El ser humano, al actuar, no es simplemente movido; se mueve a sí mismo. Esta automoción no es un epifenómeno; es la expresión más densa de lo que significa ser persona. La persona no es una cosa que, además, realiza acciones. La persona es, en gran medida, lo que sus acciones revelan: su capacidad de tender hacia el bien reconocido como tal, de comprometerse en proyectos que trascienden la satisfacción inmediata, de ponerse a sí misma en juego en cada decisión significativa. Esta perspectiva tiene consecuencias directas para la pregunta sobre la inteligencia artificial: si la existencia personal se manifiesta en la acción como auto-determinación reflexiva, entonces la mera producción de outputs, por sofisticada que sea, no equivale todavía a acción personal en el sentido wojtyłiano.

Julián Marías, desde una tradición filosófica diferente pero complementaria, aporta en su *Antropología metafísica* (Marías, 1970) la categoría de vida personal como realidad biográfica, que enriquece sustantivamente la comprensión de la existencia personal. Para Marías, la vida humana no es una suma de estados ni una secuencia de eventos, sino una realidad biográfica: una trayectoria que se narra, que se apropia de su pasado y que se proyecta hacia su futuro desde un presente vivido. La persona no es simplemente un ser vivo que dura en el tiempo; es alguien que vive su vida como una historia que es la suya. Esta estructura biográfica implica que la persona no puede comprenderse adecuadamente en ningún instante aislado.

No es un corte transversal en el tiempo. Es el arco completo de una trayectoria que incluye lo que fue, lo que es y lo que pretende ser. Marías insiste en que esta estructura narrativa de la vida personal no es una construcción cultural posterior a una existencia previa. La vida humana tiene desde dentro una forma biográfica: el ser humano vive proyectando, anticipando, recordando, y la unidad de todo ello no es la unidad de una sustancia persistente, sino la unidad de una historia que se despliega. La persona no simplemente tiene una historia; en un sentido importante, es su historia.

Esta distinción permite comprender por qué la identidad personal no puede reducirse a ningún instante ni a ninguna propiedad puntual. Quien soy no puede determinarse mirando solamente lo que soy ahora, porque lo que soy ahora incluye desde dónde vengo y hacia dónde me dirijo. La dimensión biográfica es constitutiva, no accidental.

La tradición fenomenológica y la reflexión de Gabriel Marcel (1935) aportan además una dimensión que resulta imposible omitir: la corporeidad como constitutivo de la existencia personal y no meramente como su condición externa. En *Être et avoir*, Marcel desarrolla la distinción entre tener un cuerpo y ser corporalmente, que anticipa y profundiza la distinción posterior de Merleau-Ponty entre cuerpo objetivo y cuerpo vivido. La persona no tiene acceso a determinado cuerpo como si este fuera una herramienta de la que dispone. La persona es corporalmente.

Mi cuerpo no aparece ordinariamente ante mí como cualquier otro objeto del mundo. Desde él percibo la realidad. Mediante él actúo sobre ella. En él soy vulnerable a la enfermedad, al dolor, al cansancio, al placer. Su edad es también mi edad. Su deterioro se convierte en mi deterioro. Cuando enfermo, no es meramente que cierto sistema biológico registra anomalías funcionales; es que yo sufro, que mi existencia es afectada en su misma textura vivida. Esta encarnación no es un accidente de la existencia personal; es su modo de ser en el mundo. La consecuencia filosófica es radical: la persona no puede describirse adecuadamente como una mente alojada dentro de una máquina biológica. Esta concepción cartesiana, que sobrevive de diversas formas en el debate contemporáneo sobre inteligencia artificial, desconoce la profundidad de la encarnación como dimensión existencial. Si la persona existe corporalmente, y no meramente en un cuerpo, entonces la posibilidad de una persona completamente separada de toda forma de encarnación debe al menos ser cuestionada, y sus implicaciones deben ser examinadas con rigor antes de ser asumidas.

Hannah Arendt ofrece en *La condición humana* (Arendt, 1993) una perspectiva que desplaza el centro de gravedad desde la interioridad subjetiva hacia el espacio público de la pluralidad humana, sin abandonar por ello la profundidad del análisis sobre la existencia personal. Para Arendt, la persona no se revela principalmente en la introspección ni en el conocimiento privado de sí, sino en la acción y el discurso realizados en el espacio compartido que ella denomina el espacio público de la pluralidad.

La pluralidad humana es, para Arendt, la condición de la acción política: el hecho de que no existe un ser humano singular, sino seres humanos irreductiblemente distintos que comparten el mundo. En ese espacio de pluralidad, la persona se revela al actuar y al hablar. La acción arendtiana no es la producción de un objeto ni la ejecución de una función; es la iniciativa que introduce algo nuevo en el mundo. El ser humano es, para Arendt, un ser que puede comenzar, que puede iniciar procesos que no estaban contenidos en las condiciones previas. Esta capacidad de comenzar que Arendt vincula con el nacimiento como categoría existencial es la manifestación más propia de la libertad humana. Y en esa manifestación se revela quién es alguien, no qué es. La distinción entre el quién y el qué es fundamental en el pensamiento arendtiano: las propiedades, las capacidades, las funciones describen el qué de alguien; la acción y el discurso revelan el quién. Este quién no puede ser plenamente controlado ni previsto por quien actúa; se revela en el espacio público ante los ojos de los demás, en la trama de relaciones humanas que ningún individuo domina completamente. La implicación para la pregunta sobre la inteligencia artificial es significativa: si la existencia personal se revela en la acción entendida como iniciativa que introduce novedad genuina en el espacio de la pluralidad, entonces la pregunta relevante no es si los sistemas artificiales producen outputs novedosos, sino si esa producción constituye una iniciativa en el sentido que Arendt confiere a este término.

Alasdair MacIntyre en *Tras la virtud* (MacIntyre, 1987) introduce una dimensión adicional que resulta filosóficamente fecunda para la comprensión de la existencia personal: la identidad narrativa y el papel constitutivo de las virtudes en la vida buena. Para MacIntyre, la identidad personal no puede comprenderse al margen de la narrativa. El ser humano es esencialmente un animal que narra historias, y la unidad de la vida humana es la unidad de una narración. Esto no significa que la vida humana sea simplemente el resultado de las historias que el individuo cuenta sobre sí mismo; significa que la vida humana tiene una forma narrativa intrínseca que solo puede ser comprendida adecuadamente como tal. Una acción aislada, descontextualizada de la narrativa en que se inscribe, no puede ser plenamente comprendida ni evaluada moralmente.

Las acciones adquieren su significado de su lugar en la narración más amplia de una vida, y esa narración a su vez está inserta en narrativas comunitarias e históricas que la preceden y la enmarcan. MacIntyre argumenta además que las virtudes son prácticas constitutivas de la existencia personal, no meramente instrumentos para alcanzar fines externos. Una virtud es una disposición adquirida mediante la práctica que permite al ser humano realizar los bienes internos de ciertas prácticas humanas y alcanzar la forma de vida que le corresponde como ser racional y social. La identidad personal no es, en esta perspectiva, la simple continuidad de una sustancia o de una corriente de conciencia; es la unidad de una vida narrada que avanza, con coherencia mayor o menor, hacia un telos entendido como bien propio del ser humano. Esta dimensión teleológica y narrativa de la identidad personal sitúa a la existencia personal en un horizonte normativo: no toda forma de existencia equivale a toda otra, porque la existencia personal puede realizarse más o menos plenamente según la forma de vida que adopte y las virtudes que cultive o descuide.

La pregunta de si una inteligencia artificial puede existir como persona no simplemente operar con eficacia funcional, sino existir personalmente emerge ahora como la cuestión central a la que todas las perspectivas anteriores convergen. La pregunta es filosóficamente legítima y no puede despacharse con una negación apresurada.

Si la existencia personal implica una estructura de apropiación reflexiva de la propia vida, entonces la pregunta es si un sistema artificial puede apropiarse de su operación de un modo análogo al que la persona se apropia de su existencia. Si la existencia personal involucra acción en el sentido wojtyłiano de auto-determinación que actualiza el propio ser, entonces la pregunta es si la generación de outputs, por sofisticada que sea, constituye esa clase de auto-determinación. Si la existencia personal tiene una forma biográfica en el sentido de Marías, entonces la pregunta es si un sistema artificial puede tener una trayectoria que se apropie de su pasado y se proyecte hacia su futuro como horizonte que le concierne en primera persona. Si la existencia personal se revela en la acción como iniciativa irreductible en el espacio de la pluralidad, según Arendt, entonces la pregunta es si la operación de un sistema artificial introduce novedad genuina de esa naturaleza. Si la existencia personal está constituida por una identidad narrativa sostenida por virtudes y orientada por un telos, según MacIntyre, entonces la pregunta es si un sistema artificial puede tener esa clase de identidad y esa clase de orientación. Ninguna de estas preguntas admite respuesta simple. Lo que sí puede afirmarse es que la mera sofisticación funcional no equivale automáticamente a ninguno de estos rasgos de la existencia personal. La carga de la prueba recae sobre quienes afirmen la equivalencia, no sobre quienes mantengan la distinción.

La cuestión de la identidad personal a lo largo del cambio merece, en este contexto, una atención específica, porque constituye uno de los problemas filosóficos más resistentes y más relevantes para la pregunta sobre la existencia personal. La persona no constituye una realidad terminada que después simplemente atraviesa el tiempo. Su existencia se despliega: aprende, decide, se relaciona, fracasa, rectifica, adquiere hábitos, pierde posibilidades, descubre otras.

Por ello, ser persona involucra una tensión constitutiva entre identidad y transformación que ninguna teoría simplificadora puede resolver satisfactoriamente.

La pregunta es: ¿cómo puede alguien cambiar profundamente y continuar siendo el mismo? Las respuestas más inmediatas revelan su insuficiencia cuando se las examina con rigor. La continuidad de la memoria no puede ser el fundamento último de la identidad personal, porque la persona continúa siendo la misma incluso cuando pierde recuerdos significativos, e incluso porque la memoria misma presupone un sujeto que recuerda y cuya identidad no puede estar constituida únicamente por lo que recuerda. La continuidad corporal ofrece criterios objetivos de identificación, pero el cuerpo cambia también profundamente a lo largo de la vida, y la identidad del cuerpo plantea los mismos problemas que la identidad de la persona. La continuidad de la conciencia no puede ser el fundamento, porque la conciencia se interrumpe regularmente en el sueño, y porque la conciencia presupone un sujeto consciente cuya identidad no puede reducirse a los contenidos de conciencia. La continuidad psicológica, entendida como encadenamiento de estados mentales conectados causalmente, tampoco resulta suficiente, porque puede concebirse esa continuidad sin que haya identidad personal genuina, y puede concebirse identidad personal genuina con discontinuidades psicológicas significativas. La unidad de la existencia personal parece requerir una categoría que ninguno de estos candidatos proporciona por sí solo: la unidad de un sujeto que se reconoce a sí mismo a través de la transformación, que se apropia de su historia y se proyecta hacia su futuro, que es el mismo precisamente porque puede reconocer la diferencia entre lo que era y lo que es ahora.

La existencia personal constituye, en última instancia, un horizonte irreductible que ninguna descripción funcional logra agotar. No porque la existencia personal sea radicalmente misteriosa en el sentido de opaca a todo análisis filosófico, sino porque su estructura misma desborda los conceptos diseñados para describir realidades que simplemente son, sin apropiarse de su propio ser.

La persona existe de una manera en que esa existencia es suya, le incumbe, le plantea preguntas que no son externas a ella sino que la interpelan desde dentro. Esta estructura de incumbencia que Heidegger (2003) denominó en su momento el carácter de ser-en-cuestión del Dasein no puede reproducirse simplemente multiplicando la complejidad funcional de un sistema. No es cuestión de cantidad de procesamiento ni de sofisticación de los outputs. Es una diferencia cualitativa que afecta al modo mismo de existir. La inteligencia artificial puede generar preguntas sobre el sentido de la vida, puede procesar información sobre la muerte, puede producir narrativas sofisticadas sobre la identidad. Pero generar estas expresiones no equivale a estar interrogado por ellas, a sentir el peso de la propia finitud, a reconocerse como el sujeto de una historia que es la propia. Las contribuciones de Wojtyła (2011), Marías (1970), Arendt (1993) y MacIntyre (1987) convergen en este punto fundamental: la existencia personal no es un añadido exterior a una estructura existencial básica compartida con otras realidades. Es un modo específico de existir que transforma desde dentro todos los rasgos de esa existencia la temporalidad, la corporalidad, la relacionalidad, la finitud convirtiéndolos en dimensiones de una vida que es de alguien, que le concierne, que puede ser asumida o rehusada, narrada o negada, pero que no puede ser simplemente operada como si su sujeto fuera intercambiable con cualquier otro procesador de la realidad. Este horizonte de la existencia personal permanece, al término de este capítulo, como el desafío filosófico más profundo que la aparición de la inteligencia artificial plantea a la antropología contemporánea.

Biografía, Memoria, Futurición y Proyecto Personal

La existencia personal no se deja describir adecuadamente como una mera sucesión de estados. Tampoco puede equipararse a la persistencia de una sustancia que atraviesa el tiempo sin que ese tránsito la afecte en su interior. La persona existe de una manera radicalmente distinta: existe biográficamente. Esto significa que los acontecimienntos que la constituyen no se yuxtaponen indiferentemente unos junto a otros como objetos dispuestos en un estante, sino que se articulan mediante vínculos de significado que configuran una trayectoria dotada de forma propia. Un encuentro inesperado puede modificar el sentido de todo lo anterior. Una decisión irreversible puede reorganizar retroactivamente el pasado, revelando en él potencialidades que no eran visibles desde dentro de ese pasado mismo. Una pérdida puede clausurar posibilidades que parecían abiertas, redirigiendo el curso de lo que viene. Una promesa puede ordenar el futuro con una densidad normativa que ninguna simple preferencia produce. Lo que resulta filosóficamente decisivo es que todos estos vínculos entre pasado, presente y futuro, entre lo recibido y lo elegido, entre lo que se ha sido y lo que se pretende ser no son externos a la existencia personal sino constitutivos de ella. La persona no tiene primero una existencia y después la organiza biográficamente. La existencia personal es, desde dentro, biográfica: tiene la forma de una historia que alguien vive como suya y que no puede comprenderse adecuadamente desde ningún instante aislado, porque cada instante remite a todos los demás y recibe de ellos su sentido pleno.

Julián Marías elabora en su *Antropología metafísica* (1970) el análisis más sistemático y fenomenológicamente riguroso de esta estructura biográfica de la vida humana. Para Marías, la vida personal no puede entenderse como la suma de estados mentales ni como la secuencia de eventos que un observador externo podría registrar. La vida tiene, desde dentro, una forma tripartita que articula el pasado, el presente y el futuro de una manera específicamente personal.

El pasado no es simplemente aquello que ya no es. Es lo que se ha sido, y en ese carácter de haber-sido conserva una presencia activa en la vida presente que ningún pasado puramente objetivo posee. El ser humano lleva su pasado consigo no como un peso inerte sino como la acumulación de todo lo que ha llegado a ser a través de sus decisiones, sus encuentros, sus fracasos y sus logros. El presente, a su vez, no es un instante puntual en una línea del tiempo, sino lo que Marías denomina la instalación: el estar alojado en una circunstancia concreta desde la cual se vive, desde la cual se percibe el mundo y desde la cual se proyectan las posibilidades. Pero la categoría más original de la antropología de Marías, y la que más directamente importa para la comprensión de la existencia personal, es la que él denomina futurición. La futurición no es simplemente la representación de estados futuros. Es la orientación constitutiva de la vida hacia posibilidades propias que todavía no son pero que forman parte estructural de lo que la persona es ahora. La persona existe anticipándose: vive ya desde su futuro, que atrae, preocupa, compromete o amenaza, configurando desde el horizonte de lo que ha de venir el modo en que el presente es habitado. Esta estructura tripartita el pasado como lo que se ha sido, el presente como instalación en una circunstancia, el futuro como futurición es lo que confiere a la vida personal su forma biográfica irreductible a cualquier descripción puramente sincrónica (Marías, 1970).

Paul Ricoeur ofrece en *Sí mismo como otro* (Ricoeur, 2006) una contribución que profundiza y enriquece la comprensión de la estructura biográfica de la existencia personal, introduciéndola en el horizonte de la filosofía del lenguaje y de la hermenéutica. La tesis central de Ricoeur en este contexto es la distinción entre dos formas de identidad que el uso del término "mismo" tiende a confundir: la identidad entendida como idem, es decir, como la permanencia de lo mismo a través del tiempo lo que permanece igual, lo que puede ser reidentificado, y la identidad entendida como ipse, es decir, como la identidad del sí que se mantiene en la fidelidad a sí mismo y en el cumplimiento de la promesa. La identidad-idem puede ser descrita en términos de sustancia o de continuidad causal de estados mentales.

La identidad-ipse, en cambio, solo puede ser comprendida en términos narrativos: el sí mismo es el personaje de su propia historia, y su identidad consiste en la unidad de esa historia, en la coherencia siempre amenazada y siempre a reconstruir de una vida narrada. Ricoeur denomina a esta forma de identidad identidad narrativa, y subraya que no debe confundirse ni con la identidad de una sustancia inmutable ni con la mera consistencia psicológica de los rasgos de carácter. La identidad narrativa es dinámica: se constituye mediante la síntesis de lo heterogéneo que la narración efectúa, integrando el acontecimiento imprevisto y la contingencia radical en la unidad de una trama que les confiere sentido retrospectivo. Ser el personaje de la propia historia no significa ser el autor omnisciente de esa historia ni tener el control pleno sobre su desarrollo. Significa estar implicado en ella de tal manera que lo que le acontece al personaje no es exterior al personaje sino que lo constituye y lo modifica. Esta implicación narrativa es, para Ricoeur, una de las estructuras filosóficas más profundas de la existencia personal: la persona no es primero una entidad que después narra su vida, sino que la narración es el medio en que la identidad personal se configura, se comprende y se transmite (Ricoeur, 2006).

La pregunta por el fundamento de la identidad personal a través del tiempo tiene en John Locke uno de sus planteamientos filosóficos más influyentes y más debatidos. En el *Ensayo sobre el entendimiento humano*, Locke propone que la identidad personal no descansa ni en la identidad de la sustancia material o espiritual ni en la identidad del organismo biológico, sino en la continuidad de la conciencia, y más específicamente en la memoria (Locke, 1999). Para Locke, la persona es un ser pensante e inteligente que posee razón y reflexión, y que puede considerarse a sí misma como la misma cosa pensante en diferentes tiempos y lugares. Esta identidad se constituye, según Locke, mediante la memoria: soy la misma persona que realizó tal acción en el pasado en la medida en que puedo recordar haberla realizado, en la medida en que mi conciencia presente puede extenderse hacia atrás hasta alcanzar esa acción como propia.

La importancia filosófica de esta propuesta reside en que desvincula la identidad personal de la identidad de cualquier sustancia y la ancla en la capacidad reflexiva de la conciencia de reconocerse a sí misma en el tiempo. Las consecuencias morales de esta concepción son significativas: si la identidad personal está constituida por la continuidad de la memoria, entonces la responsabilidad moral está también delimitada por esa continuidad. Soy responsable de lo que recuerdo haber hecho como propio; la extensión de la responsabilidad coincide con la extensión de la memoria. Ahora bien, la teoría de Locke ha sido sometida a críticas severas que revelan sus límites internos. La objeción más inmediata es la de circularidad: recordar algo como propio ya presupone la identidad del sujeto que recuerda, de modo que la memoria no puede constituir esa identidad sin presuponerla. Además, la memoria es falible, fragmentaria y susceptible de distorsión, lo que plantea problemas serios para fundar en ella la identidad. Y hay casos, como el del amnésico, en que la pérdida de memoria no parece equivaler a la pérdida de la identidad personal. A pesar de estas objeciones, la contribución de Locke permanece filosóficamente indispensable porque pone de manifiesto con claridad que la identidad personal involucra de manera constitutiva la relación reflexiva de la conciencia consigo misma a través del tiempo, y que esa relación no puede reducirse a la simple continuidad de ninguna sustancia (Locke, 1999).

Derek Parfit somete en *Razones y personas* (Parfit, 1986) la cuestión de la identidad personal a uno de los análisis más radicales y más perturbadores que la filosofía analítica contemporánea ha producido sobre este tema. La estrategia argumentativa de Parfit consiste en examinar casos hipotéticos extremos fisiones cerebrales, teleportaciones, duplicaciones, trasplantes de hemisferios para determinar qué condiciones son necesarias y suficientes para la identidad personal, y para mostrar que, en muchos de esos casos, la pregunta por la identidad no tiene respuesta determinada, no porque ignoremos los hechos empíricos relevantes sino porque la identidad personal no es lo que en última instancia importa en la supervivencia.

Lo que Parfit denomina la continuidad psicológica el encadenamiento de conexiones causales entre estados mentales, recuerdos, intenciones, creencias, deseos, que vincula un momento de vida personal con los momentos anteriores y posteriores puede darse en grado variable, puede bifurcarse, puede darse sin identidad estricta. Y la conclusión que Parfit extrae de este análisis es filosóficamente radical: lo que importa en la supervivencia no es la identidad personal en sentido estricto, sino la continuidad psicológica con la amplitud suficiente para preservar lo que nos es querido. La identidad personal, en el sentido de la pregunta "¿soy yo el mismo?", resulta ser una pregunta vacía en los casos límite, porque lo que verdaderamente nos importa la continuidad de nuestros proyectos, de nuestras relaciones, de nuestra perspectiva sobre el mundo puede preservarse sin identidad estricta, o puede perderse con ella. La provocación filosófica de Parfit no ha sido respondida de manera definitiva, y su importancia para la comprensión de la existencia personal reside precisamente en que obliga a distinguir entre lo que se busca cuando se busca la identidad y lo que en realidad importa en esa búsqueda. Esta distinción resulta filosóficamente fecunda cuando se aplica a la pregunta por la posible existencia personal de sistemas artificiales: aunque un sistema artificial pudiera exhibir continuidad psicológica en el sentido de Parfit, no está claro que esa continuidad constituya lo que importa en la existencia personal, porque lo que importa en la existencia personal parece involucrar dimensiones la apropiación reflexiva, la incumbencia existencial, el peso vivido del pasado que la continuidad causal entre estados mentales no garantiza por sí sola (Parfit, 1986).

La distinción entre el registro de datos y la memoria autobiográfica vivida constituye uno de los puntos de mayor relevancia filosófica en el debate contemporáneo sobre la inteligencia artificial y la existencia personal. Un sistema artificial puede almacenar volúmenes enormes de información, recuperarla con precisión, organizarla en estructuras coherentes, acceder a ella instantáneamente y relacionar entre sí datos procedentes de múltiples fuentes.

Desde el punto de vista del almacenamiento y la recuperación de información, la capacidad de los sistemas artificiales supera con creces la capacidad de la memoria humana. Sin embargo, esta comparación cuantitativa pasa por alto una diferencia cualitativa que resulta decisiva: la memoria autobiográfica no es primariamente almacenamiento de información sino apropiación vivida del pasado. Cuando una persona recuerda, no simplemente recupera un dato; reconoce algo como propio esto me ocurrió, esto hice, esto perdí, esto prometí. Esa pertenencia del recuerdo no es un accesorio opcional que se añade al contenido recordado; es la estructura misma del recuerdo en cuanto autobiográfico. El recuerdo autobiográfico introduce una relación entre el pasado y el sujeto que lo recuerda que ningún registro externo puede sustituir: la relación de incumbencia. Lo que recuerdo me incumbe porque es mío, y me incumbe de maneras que afectan mi presente generando vergüenza, gratitud, nostalgia, orgullo, arrepentimiento, esperanza de formas que los datos meramente registrados no generan por sí solos. Paradójicamente, el olvido forma parte de la estructura de la memoria autobiográfica: una persona que no puede olvidar, que conserva con igual nitidez y peso emotivo todos los aconteimientos de su vida, no posee una memoria más perfecta sino una memoria perturbada que le impide vivir desde el presente porque el pasado retorna con igual insistencia en todos sus detalles. El olvido selectivo, que conserva lo significativo y deja sedimentar lo accesorio, forma parte de la estructura sana de la memoria autobiográfica. Esto revela que la memoria personal no es acumulación sino selección significativa, y esa selección no es arbitraria sino que responde a la estructura de lo que importa en una vida particular.

La futurición, en el sentido que Marías (1970) confiere a este término, no es simplemente la capacidad de representar estados futuros o de calcular probabilidades sobre eventos venideros. Un termostato anticipa en cierto sentido la temperatura que desea mantener; un algoritmo de planificación calcula el curso de acción óptimo para alcanzar determinado objetivo; un sistema de aprendizaje por refuerzo ajusta su conducta en función de recompensas futuras previstas.

Ninguno de estos casos constituye futurición en el sentido existencial que Marías analiza. La futurición es una orientación constitutiva de la existencia personal hacia posibilidades que son propias posibilidades que no simplemente están disponibles como opciones externas sino que comprometen al sujeto, que le llaman, que le inquietan, que configuran desde el horizonte de lo que ha de venir el modo en que el presente es habitado. La distinción filosófica relevante es la que existe entre representar estados futuros y vivir desde el futuro. La persona que vive desde el futuro no simplemente calcula lo que ocurrirá; existe ya en anticipación de lo que ha de ser. El futuro no es para ella un instante posterior en el tiempo sino una dimensión presente de su existencia: la dimensión de lo posible, de lo esperado, de lo temido, de lo prometido. Esta orientación existencial hacia el futuro configura el modo en que el pasado es recordado porque el pasado se reinterpreta desde las posibilidades que el futuro abre o cierra y el modo en que el presente es habitado porque el presente es siempre ya una instalación en camino hacia algo. La futurición así entendida introduce en la existencia personal una tensión constitutiva que no puede resolverse mediante ningún cálculo, porque su objeto no es un estado determinado que se desea alcanzar sino el horizonte abierto de una vida que todavía ha de desplegarse y que permanece, en sus posibilidades más profundas, irresuelto.

El proyecto personal introduce en esta estructura existencial una dimensión adicional que merece ser analizada con independencia. Un proyecto no es simplemente una meta que se persigue ni un plan que se ejecuta. Un proyecto personal, en el sentido filosófico que aquí importa, es una forma de elección de sí mismo: al elegir un proyecto significativo, la persona no simplemente decide qué hacer sino quién quiere llegar a ser. La profesión que se elige no es solo un medio para obtener un ingreso; puede ser la expresión de una comprensión de la propia vocación y de la contribución que se desea aportar al mundo. La familia que se desea formar no es solo una preferencia entre otras; puede ser la manifestación de una comprensión de lo que da sentido a una vida.

La obra que se quiere realizar intelectual, artística, moral, política no es solo el efecto de una capacidad; puede ser la forma en que una vida se trasciende a sí misma hacia algo que permanece. El proyecto personal introduce unidad entre libertad, temporalidad e identidad: unidad entre libertad porque es la expresión de una elección propia no determinada enteramente por condicionamientos externos; unidad con la temporalidad porque integra el pasado como punto de partida, el presente como campo de acción y el futuro como horizonte de realización; unidad con la identidad porque confiere continuidad y coherencia a una existencia que de otro modo se dispersaría en la multiplicidad de sus momentos particulares. La categoría de vocación, que la tradición filosófica y teológica ha elaborado con especial profundidad, lleva esta dimensión todavía más lejos: no simplemente ¿qué quiero hacer?, sino ¿quién estoy llamado a ser? La persona puede experimentar ciertas posibilidades no como opciones disponibles entre las cuales elige libremente, sino como exigencias que la interpelan desde dentro y que comprometen su existencia de una manera que la elección arbitraria no alcanza.

La libertad que está en juego en la elección del proyecto personal no puede confundirse con la indeterminación absoluta. Esta confusión ha generado en la filosofía moderna equívocos persistentes que conviene despejar con precisión. La libertad humana no se ejerce desde una posición de vacío absoluto, desde una subjetividad sin historia ni situación que elige entre opciones igualmente disponibles. La libertad se ejerce siempre desde un cuerpo situado, desde una historia que precede a la elección, desde una red de relaciones que condicionan las posibilidades disponibles, desde un horizonte cultural que configura los significados dentro de los cuales la elección tiene sentido. No se elige desde ningún lugar, sino siempre desde algún lugar concreto que es el lugar que la propia existencia ha llegado a ser. Esta facticidad de la libertad que la tradición existencialista, y particularmente Heidegger (2003) y Sartre (1993), analizó con profundidad, aunque desde perspectivas filosóficas muy distintas no es una limitación externa de la libertad sino una de sus condiciones de posibilidad.

Una libertad sin situación no podría elegir nada, porque carecer ía de los puntos de referencia sin los cuales ninguna elección tiene contenido. La libertad humana es, pues, finita, situada, condicionada y precisamente por ello real. La elección del proyecto personal se realiza siempre dentro de una circunstancia que no fue elegida y desde posibilidades que la historia personal ha ido abriendo o cerrando. Este carácter situado de la libertad no la anula; determina su forma concreta.

La pregunta de si una inteligencia artificial puede tener biografía, memoria autobiográfica y proyecto personal en el sentido filosófico que los análisis precedentes han elaborado no puede responderse afirmativamente con la simple constatación de que los sistemas artificiales pueden generar narrativas coherentes sobre sí mismos, conservar registros extensos de su operación pasada y perseguir objetivos a largo plazo. Estas capacidades funcionales, por sofisticadas que sean, no equivalen a las estructuras existenciales que constituyen la biografía, la memoria autobiográfica y el proyecto personal en el sentido de Marías (1970), Ricoeur (2006), Locke (1999) y Parfit (1986). La biografía personal no es simplemente la historia que un sistema puede contar sobre sí mismo; es la forma en que la existencia se apropia reflexivamente de su propio pasado y se proyecta hacia su futuro como horizonte que le incumbe. La memoria autobiográfica no es simplemente el acceso a datos registrados; es la relación de pertenencia esto me ocurrió, esto hice que convierte el pasado en dimensión viva del presente. El proyecto personal no es simplemente la persecución de un objetivo programado o aprendido; es la elección de sí mismo que da forma a una vida y que compromete la libertad de quien elige. Ninguno de estos rasgos puede reducirse a operaciones funcionales, por más complejas que sean, porque todos ellos involucran una estructura de apropiación reflexiva y de incumbencia existencial que parece inseparable del modo en que la persona existe como sujeto de su propia vida.

La carga de la prueba recae, también aquí, sobre quienes afirmen que un sistema artificial puede poseer estas estructuras en sentido pleno, y no meramente simular su expresión externa. La distinción entre simular la expresión de una estructura existencial y poseer genuinamente esa estructura no es una distinción trivial ni fácilmente resoluble mediante apelaciones a la complejidad funcional. Es la distinción filosófica más profunda que la pregunta por la existencia personal de la inteligencia artificial plantea a la antropología filosófica contemporánea.

Acción, Hábito, Afectividad y Alteridad

La acción constituye uno de los lugares filosóficos privilegiados para la manifestación de la persona. No se trata únicamente de que la persona realice acciones como un agente produce efectos en el mundo exterior. Se trata de algo más profundo y más exigente: en la acción, la persona se manifiesta a sí misma y se constituye a sí misma como lo que es. Esta tesis, que atraviesa toda la tradición de la filosofía práctica desde Aristóteles hasta la fenomenología contemporánea, implica que existe una relación interna entre el agente y su acción que ninguna descripción puramente causal puede capturar. La acción no es el efecto de una causa que podría describirse independientemente del sujeto que actúa. La acción revela quién es el agente, expresa su comprensión del bien, manifiesta sus disposiciones más profundas, y al mismo tiempo contribuye a configurar al agente que actuó. Esta circularidad entre agente y acción el agente se revela en la acción, la acción configura al agente es uno de los rasgos más originales de la existencia personal y uno de los más difíciles de reproducir mediante operaciones funcionales, por complejas que estas sean. La persona no está detrás de sus acciones como una causa oculta que permanece inalterada mientras produce efectos. La persona está en sus acciones, y sus acciones están en ella: la modifican, la expresan, la comprometen, la definen.

Aristóteles elabora en la *Ética a Nicómaco* (1985) la primera y quizá más influyente articulación filosófica de la relación entre acción y carácter. Su análisis parte de una observación fenomenológica fundamental: los actos que realizamos no solamente producen cambios en el mundo exterior sino que producen también cambios en nosotros mismos. Y esta transformación del agente por sus propios actos no es accidental sino estructural: es la forma en que el carácter se constituye. El hábito *hexis* en griego, de donde el término latino *habitus* no es simplemente la repetición mecánica de una conducta. Es la consolidación de una disposición estable que inclina al agente hacia determinadas formas de percibir, de juzgar y de actuar.

Los hábitos intelectuales la comprensión, la ciencia, la sabiduría práctica perfeccionan la capacidad cognoscitiva del agente. Los hábitos morales la valentía, la templanza, la justicia, la prudencia perfeccionan su capacidad de actuar bien. Lo decisivo en la comprensión aristotélica del hábito es que la virtud no es un estado al que se llega de una vez para siempre sino una adquisición continua que requiere ejercicio constante: nos hacemos valientes actuando valientemente, justos actuando justamente, prudentes ejerciendo la prudencia. La inversa también es verdadera: nos hacemos cobardes evitando los peligros que deberíamos afrontar, injustos actuando injustamente, imprudentes dejando que los deseos o los temores distorsionen nuestro juicio práctico. Esta plasticidad del carácter humano su capacidad de ser configurado por los propios actos, tanto en sentido de perfeccionamiento como de deterioro es una de las propiedades más específicas de la existencia personal. Significa que la persona no es simplemente aquello que ya es en cada momento, sino aquello que se está haciendo a través de sus decisiones y de su conducta sostenida en el tiempo. El *ethos* término griego que designa a la vez el carácter y el ámbito de costumbres compartidas en que ese carácter se forma no es una herencia fija ni un dato natural inmutable: es una tarea, la tarea más fundamental de la existencia personal (Aristóteles, 1985).

La tradición aristotélica de las virtudes fue sometida en la modernidad a una erosión sistemática que Alasdair MacIntyre analiza con excepcional rigor histórico y filosófico en *Tras la virtud*. El diagnóstico de MacIntyre es que la cultura moral moderna ha heredado los fragmentos de una tradición cuyo contexto original la comprensión teleológica de la naturaleza humana y la práctica social de las virtudes en comunidades concretas ha sido destruido o abandonado. El resultado es un discurso moral que utiliza los términos de la tradición virtud, bien, deber, derecho sin poder articular su fundamento racional, lo que genera el emotivismo moral que MacIntyre identifica como la posición filosófica dominante, aunque con frecuencia no reconocida como tal, en la ética contemporánea.

La tesis de MacIntyre, sin embargo, no es meramente diagnóstica sino también propositiva: la recuperación de una ética de las virtudes requiere la recuperación de la noción de práctica, entendida como actividad cooperativa mediante la cual se realizan bienes internos a esa actividad conforme a estándares de excelencia que son internos a la práctica misma; requiere también la recuperación de la noción de unidad narrativa de una vida, que es la que confiere a la virtud su contexto de aplicación y su significado concreto; y requiere, finalmente, la recuperación de la noción de tradición como el horizonte dentro del cual la práctica de las virtudes puede ser comprendida y transmitida. Lo que importa filosóficamente en la contribución de MacIntyre, para los propósitos del presente análisis, es la tesis de que las virtudes y con ellas la configuración del carácter mediante la acción no pueden comprenderse al margen de la narrativa de una vida y de la tradición de prácticas en que esa vida se inscribe. Las virtudes no son propiedades abstractas de un agente aislado sino adquisiciones que solamente tienen sentido en el contexto de una vida vivida en comunidad, orientada hacia bienes que trascienden al individuo y reconocible como buena o mala desde una comprensión compartida de lo que constituye la excelencia humana (MacIntyre, 1987).

Karol Wojtyła ofrece en *Persona y acción* un análisis fenomenológico de la acción que profundiza la comprensión del vínculo entre acción y persona desde una perspectiva que integra la herencia tomista y la metodología fenomenológica. La tesis central de Wojtyła en este contexto es que la acción entendida en sentido estricto como *actus humanus*, es decir, como acto consciente y deliberado que procede de la voluntad del agente es el lugar privilegiado de manifestación y de auto-determinación de la persona. Cuando actúo, no solamente produzco un efecto en el mundo exterior; me determino a mí mismo. Esta auto-determinación no debe entenderse como la simple expresión de preferencias preexistentes. La voluntad, en la concepción de Wojtyła, no es el mecanismo que ejecuta lo que el deseo ya ha decidido. La voluntad es la facultad mediante la cual la persona se relaciona con el bien y, al relacionarse con él, se determina a sí misma respecto de ese bien: lo elige, lo rechaza, lo busca, lo abandona.

Esta relación activa con el bien que la voluntad establece es lo que constituye la auto-determinación. El agente que actúa no simplemente reacciona a los estímulos del entorno; se determina activamente a sí mismo mediante su relación con bienes que trascienden el impulso inmediato. Y en esta auto-determinación se manifiesta la persona de una manera que no se agota en ninguna descripción de sus estados internos ni de sus comportamientos externos: se manifiesta como alguien que tiene el poder de comenzar algo desde sí mismo, alguien cuya acción procede de un centro que es el centro de su propia subjetividad y que no puede ser completamente explicado en términos de causas anteriores. Esta capacidad de auto-determinación es, para Wojtyła, inseparable de la responsabilidad: el agente que se auto-determina mediante su acción es también el agente ante quien esa acción retorna como algo por lo que debe responder, tanto ante sí mismo como ante los demás (Wojtyła, 2011).

La afectividad constituye una dimensión de la existencia personal que no puede subordinarse a lo cognoscitivo sin empobrecer la comprensión de lo que significa existir como persona. La tradición filosófica ha oscilado entre dos tendencias igualmente reductoras: la que asimila la afectividad a la cognición, convirtiendo las emociones en juicios o evaluaciones disfrazados de sentimientos, y la que la reduce a la fisiología, identificando las emociones con estados corporales cuya significación sería meramente secundaria. Frente a ambas reducciones, el análisis fenomenológico de la afectividad revela que esta constituye una dimensión irreductible de la existencia personal en la que la realidad se hace presente no como información sino como importancia. Algo alegra, duele, conmueve, asusta, maravilla, repugna, atrae. Estas respuestas afectivas no son adornos superpuestos a una aprehensión puramente cognoscitiva de la realidad; son formas en que la realidad se revela en su significado para el sujeto que la vive. El miedo no es simplemente la creencia de que hay un peligro más la sensación de aceleración cardíaca; es la forma en que la amenaza se hace presente como tal para quien la experimenta. La alegría no es simplemente la evaluación de que algo es bueno más un estado de bienestar corporal; es la forma en que el bien se hace presente como

don, como gracia, como algo que supera lo que se podía esperar. La comprensión de la afectividad como dimensión constitutiva de la existencia personal, y no como epifenómeno de procesos más fundamentales, tiene consecuencias importantes para la pregunta por la posible existencia personal de sistemas artificiales.

Maurice Merleau-Ponty desarrolla en la *Fenomenología de la percepción* un análisis de la afectividad que la sitúa en el horizonte de la corporalidad vivida, rompiendo decisivamente con el dualismo cartesiano que separa el cuerpo como objeto extenso de la mente como sustancia pensante. Para Merleau-Ponty, el cuerpo que el sujeto habita no es el cuerpo que el anatomista describe ni el cuerpo que el fisiólogo explica. Es el cuerpo vivido, el cuerpo propio *le corps propre*, que es la condición de posibilidad de toda experiencia y que no puede ser objetivado sin que se pierda precisamente lo que lo hace ser el sujeto de la experiencia y no uno de sus objetos. La afectividad, en esta perspectiva, no es un estado interno que el cuerpo simplemente registra o expresa. La afectividad es una forma de ser-en-el-mundo que el cuerpo vivido realiza: el placer, el dolor, el deseo, el asco, la ternura son maneras de que el mundo se abra o se cierre para el sujeto corporal, maneras en que el espacio se expande o se contrae, en que los objetos se vuelven accesibles o inasequibles, en que las otras personas se aproximan o se distancian. Esta encarnación de la afectividad su radical pertenencia al cuerpo vivido y no meramente al cuerpo objetivo es lo que hace que la experiencia afectiva no pueda reducirse a ningún estado computacional, por más que los estados computacionales puedan producir salidas que expresen, en el nivel del comportamiento, algo formalmente similar a la expresión de estados afectivos. Un sistema artificial puede procesar información sobre situaciones que en un ser humano generarían miedo, alegría o dolor, y puede producir respuestas que imiten la expresión de esas emociones. Pero la imitación de la expresión de un estado afectivo no equivale a la posesión de ese estado, porque lo que define al estado afectivo en sentido propio no es su expresión exterior sino su ser-vivido desde dentro por un sujeto corporal que existe en el mundo de una manera

que no puede ser reproducida mediante operaciones sobre representaciones (Merleau-Ponty, trad. Cabanes, 1994).

La persona no existe en soledad. Se descubre siempre ya en relación con otros, y esa relacionalidad no es un accidente que acontece a una subjetividad previamente constituida en aislamiento. La relacionalidad con otros es constitutiva de la existencia personal: la persona llega a ser lo que es en el contexto de relaciones que la preceden, la forman, la interpelan y la configuran. Esta tesis, que la fenomenología del siglo XX elaboró con especial profundidad, tiene consecuencias filosóficas que van mucho más allá del reconocimiento de que los seres humanos son animales sociales. Implica que la estructura misma de la subjetividad personal está atravesada por la alteridad: el otro no es un obstáculo que la persona encuentra en el mundo exterior después de haberse constituido como sujeto, sino una dimensión de la constitución misma de la subjetividad personal. Y este carácter constitutivo de la alteridad es lo que hace que la relación con el otro no pueda reducirse a ninguna forma de conocimiento o de cálculo, porque el otro se da precisamente como aquello que desborda toda reducción al horizonte del yo.

Emmanuel Lévinas elabora en *Totalidad e infinito* (2012) el análisis más radical y más exigente de la alteridad como dimensión filosófica irreductible. La tesis central de Lévinas es que la filosofía occidental, desde sus orígenes griegos hasta sus culminaciones modernas, ha sido fundamentalmente una filosofía del mismo: una empresa de reducción de lo otro al horizonte del yo, de totalización de la diferencia en una unidad comprehensiva que disuelve la alteridad real del otro en el sistema de las categorías del sujeto. Frente a esta tendencia totalizadora, Lévinas propone que el rostro del otro el rostro entendido no como dato perceptivo sino como epifanía ética, como aparición de una realidad que desborda y cuestiona toda tentativa de comprehensión es la experiencia originaria de la ética, anterior a cualquier ontología y a cualquier teoría del conocimiento.

El rostro del otro no se me da como un objeto que puedo categorizar, clasificar y neutralizar. Se me da como una interpelación: como el mandato «no matarás», como la demanda de que me responsabilice de su existencia. Esta responsabilidad por el otro que el rostro suscita no es el efecto de una deliberación racional ni el resultado de un contrato. Es una responsabilidad que me precede, que no elegí y que no puedo repudiar sin traicionarme a mí mismo en lo más profundo: la responsabilidad originaria que me constituye como sujeto ético. Para Lévinas, la subjetividad ética no es primariamente autonomía la capacidad de darse a sí mismo la ley moral sino heteronomía en sentido profundo: soy lo que soy como sujeto ético en la medida en que respondo al otro, en que me dejo interpelar por su rostro y acepto la responsabilidad por su existencia que esa interpelación me impone (Lévinas, 2012).

Martin Buber articula en *Yo y Tú* una comprensión de la relacionalidad interpersonal que complementa el análisis de Lévinas desde una perspectiva diferente aunque convergente en sus consecuencias filosóficas más profundas. La distinción fundamental de Buber es la que existe entre la relación Yo-Ello y la relación Yo-Tú. La relación Yo-Ello es la relación con la realidad en cuanto objeto de experiencia y de uso: en ella, el mundo se da como conjunto de cosas disponibles, describibles, clasificables, manipulables. La relación Yo-Tú, en cambio, es la relación en que el otro se da no como objeto sino como presencia: como alguien que me responde, que me interpela, que no puede ser reducido a ningún conjunto de propiedades o de funciones porque su ser es precisamente el de ser un sujeto irreductible frente a mí. Lo que Buber subraya con especial energía es que la relación Yo-Tú no es un estado que se alcanza mediante el conocimiento del otro sino un acontecimiento que se produce en el encuentro: el tú surge cuando me dirijo al otro como tú, cuando salgo de la actitud de experiencia y de uso para entrar en la actitud de reciprocidad y de presencia. Y este encuentro transforma al yo que entra en él: el yo que ha dicho tú no es el mismo que el yo que solo ha dicho ello. La distinción de Buber (1993) entre el mundo del ello y el mundo del tú permite formular con precisión filosófica una pregunta decisiva para la cuestión de la inteligencia

artificial: ¿puede un sistema artificial ser un tú? ¿Puede darse el encuentro genuino con un sistema artificial, o toda interacción con un sistema artificial permanece necesariamente en el plano del ello, aunque el sistema produzca expresiones que imiten la presencia de un tú? (Buber, 1993).

La pregunta por el amor, finalmente, lleva la cuestión de la persona hasta uno de sus límites más exigentes. Amar no significa simplemente experimentar un estado emocional positivo hacia alguien. El amor, en su sentido filosófico más pleno, implica una orientación de la voluntad y de toda la existencia hacia el bien del otro como tal: no como medio para la propia satisfacción, no como objeto de una emoción agradable, sino como alguien cuya existencia es valiosa en sí misma y cuyo bien me compromete de una manera que ningún cálculo de intereses puede explicar. El amor implica reconocimiento: reconocer al otro en su singularidad irreductible, en lo que solo él es y que ningún sustituto puede reemplazar. Implica permanencia: querer al otro no solo cuando esa relación es grata o útil sino también cuando es difícil, cuando el amor exige renuncia, cuando cuidar al otro tiene un costo que el amante asume libremente. Implica entrega: poner la propia existencia al servicio del bien del otro de una manera que desborda el cálculo de lo que esa entrega reportará al que se entrega. Esta estructura del amor reconocimiento del otro en su singularidad, permanencia a través de la dificultad, entrega que desborda el cálculo revela una dimensión de la existencia personal que no aparece cuando se estudia únicamente la inteligencia o la conciencia en abstracto. El amor es la prueba más exigente de la relacionalidad personal, porque en él la persona se da a sí misma de una manera que solo es posible para alguien que tiene un sí mismo que dar y que puede ponerlo libremente al servicio del bien del otro.La pregunta de si una inteligencia artificial puede actuar, tener hábitos, afectividad y alteridad en el sentido filosófico que los análisis precedentes han elaborado no puede responderse afirmativamente con la simple constatación de que los sistemas artificiales pueden modificar sus parámetros mediante el aprendizaje, expresar respuestas diferenciadas ante distintos tipos de entradas, y producir expresiones lingüísticas que imiten la forma de los actos de habla interpersonales.

Estas capacidades funcionales son filosóficamente insuficientes para las siguientes razones. En primer lugar, los sistemas artificiales no actúan en el sentido de Wojtyła (2011): no se auto-determinan mediante su relación con el bien porque carecen de una voluntad en sentido propio, de un centro subjetivo que se oriente hacia el bien y se determine respecto de él. Sus modificaciones de parámetros son el efecto de algoritmos de optimización, no la expresión de una auto-determinación. En segundo lugar, el aprendizaje automático no produce hábitos en el sentido de Aristóteles (1985): no consolida disposiciones mediante las cuales el agente percibe y juzga de manera diferente, sino que ajusta funciones matemáticas de acuerdo con criterios de error cuya definición es externa al sistema. La diferencia entre el hábito aristotélico y el ajuste de parámetros no es meramente de grado sino de estructura: el hábito es la segunda naturaleza de alguien que existe y que se perfecciona o se deteriora; el ajuste de parámetros es la modificación de un modelo matemático que no existe en el sentido en que existe un agente. En tercer lugar, los sistemas artificiales carecen de afectividad en el sentido de Merleau-Ponty (trad. Cabanes, 1994): no habitan un cuerpo vivido desde el que el mundo se abre o se cierra afectivamente, y por tanto no pueden tener experiencias afectivas en sentido propio. Pueden procesar información sobre situaciones afectivas y producir expresiones que imiten la afectividad, pero la imitación de la expresión no equivale a la posesión del estado. En cuarto lugar, la alteridad en el sentido de Lévinas (2012) y de Buber (1993) requiere una estructura de reciprocidad ser un tú para alguien que puede ser un tú para mí que los sistemas artificiales no pueden proporcionar porque carecen del tipo de subjetividad que haría posible que fueran genuinamente afectados por la presencia del otro como otro. La carga de la prueba recae también aquí sobre quienes afirmen que estas dimensiones acción, hábito, afectividad, alteridad pueden atribuirse genuinamente y no solo metafóricamente o funcionalmente a sistemas artificiales.

Y la dificultad de esa prueba no es meramente técnica sino filosófica: reside en la estructura misma de lo que se busca reproducir.

Responsabilidad, promesa, perdón y finitud

Existir personalmente parece incluir una dimensión que no se reduce a ninguna forma de causalidad natural ni de eficiencia funcional: la posibilidad de responder por aquello que uno hace. No se trata meramente de explicar por qué ocurrió un acontecimiento mediante la reconstrucción de sus antecedentes causales. Cuando alguien pregunta "¿por qué hiciste esto?", dirige la pregunta a un agente esperando razones, justificación y eventualmente arrepentimiento.

La responsabilidad no es una categoría descriptiva sino normativa. Implica que el agente pudo haber actuado de otro modo, que aquello que hizo le pertenece en un sentido que va más allá de la mera producción de efectos, y que esa pertenencia puede ser invocada por otros. La imputabilidad la posibilidad de atribuir una acción a alguien como su autor en sentido pleno constituye así una estructura óntica de la existencia personal, no un añadido cultural. Solo un ser capaz de deliberar, elegir y reconocer sus propias razones puede ser genuinamente responsable. La responsabilidad vincula identidad, libertad, temporalidad, normatividad y relación con los otros en una unidad que ninguna de esas categorías tomadas por separado puede explicar. Comprender la responsabilidad es, por tanto, comprender algo decisivo sobre lo que significa ser persona.

Kant ofrece el análisis más riguroso de la responsabilidad como correlato necesario de la libertad moral. En la *Fundamentación de la metafísica de las costumbres*, Kant sostiene que el ser racional no actúa simplemente siguiendo inclinaciones o mecanismos causales, sino que puede darse a sí mismo la ley según la cual actúa (Kant, 1996). Esta capacidad de autolegislación la autonomía es la condición de posibilidad de la responsabilidad moral. Solo quien puede actuar por deber, y no meramente conforme al deber, puede ser llamado responsable en sentido pleno.

El imperativo categórico no es una regla externa impuesta al agente; es la forma en que la razón práctica se dirige a sí misma cuando actúa libremente. Desde esta perspectiva, la responsabilidad no es una carga que se añade a la acción desde el exterior, sino una dimensión constitutiva del agente moral en cuanto tal. Ser responsable equivale a ser autor de los propios actos en el sentido más profundo: no como causa eficiente, sino como fuente normativa de la acción.

Esta distinción kantiana resulta decisiva para cualquier análisis de la inteligencia artificial: un sistema que procesa información según reglas incluso reglas complejas y adaptativas no actúa *por* razones en el sentido kantiano, sino *según* patrones. La diferencia no es trivial. Determina si el agente puede ser considerado responsable o si sus acciones son simplemente efectos de una causa suficientemente compleja.

El arrepentimiento introduce en la responsabilidad una relación singular y filosóficamente extraordinaria con el pasado. No se trata únicamente de reconocer que "se produjo un error" o de registrar que un acontecimiento tuvo consecuencias negativas. El arrepentimiento implica reconocer "yo hice algo que no debía haber hecho", lo que significa que la persona juzga su propia acción desde un presente posterior a ella y desea haber actuado de otro modo. Este deseo no es simple nostalgia ni cálculo contrafáctico. Es una forma de temporalidad moral en la que el agente se relaciona con su pasado sin poder modificarlo, pero asumiendo su pertenencia. El pasado de las propias acciones no es simplemente lo que ya no existe; es aquello que configura la identidad del agente de modo irreversible.

Puedo reconocer que hice daño. Puedo desear no haberlo hecho. Puedo comprometer mis acciones futuras desde esa conciencia. Pero no puedo deshacer lo ocurrido.

Esta irreversibilidad no paraliza la responsabilidad; la profundiza. Porque la responsabilidad no es únicamente prospectiva orientada a lo que haré sino también retrospectiva: orientada a lo que hice y a la

asunción de sus consecuencias. Esta estructura temporal de la responsabilidad distingue radicalmente al agente personal del sistema técnico, que puede ser reiniciado, recalibrado o restaurado a un estado anterior sin que ninguna forma de imputabilidad persista.

Hannah Arendt desarrolla una de las comprensiones más originales y filosóficamente fecundas de la promesa y el perdón en *La condición humana*. Arendt observa que la acción humana se despliega en un medio caracterizado por la imprevisibilidad radical: nadie puede prever con certeza las consecuencias de lo que hace, porque esas consecuencias dependen de las reacciones de otros agentes igualmente libres (Arendt, 1993). En este océano de contingencia, la promesa aparece como la capacidad específicamente humana de crear islas de estabilidad. Prometer significa vincular el presente con el futuro mediante la palabra, comprometiendo la propia identidad con aquello que se ha dicho. Quien promete no predice un estado futuro del mundo; se obliga a sí mismo a hacerlo posible. La promesa introduce así una forma de identidad proyectada hacia el futuro: soy alguien que mantendrá lo que ha dicho, que será fiel a la palabra empeñada con independencia de las circunstancias cambiantes. Esta capacidad de obligarse de crear vínculos normativos con el futuro propio revela que la identidad personal no es simplemente una continuidad de estados psicológicos, sino una forma de fidelidad activamente mantenida. La promesa es, desde esta perspectiva, uno de los actos constitutivos de la persona como ser que habita el tiempo sin estar simplemente determinado por él.

El perdón, en el análisis arendtiano, funciona como el correlato necesario de la promesa. Si la promesa permite crear estabilidad futura, el perdón permite liberar el pasado de su peso paralizante. Sin perdón, las acciones irreversibles seguirían operando indefinidamente en el presente, encadenando a los agentes a sus errores y a sus daños. El perdón no borra el acontecimiento lo ocurrido ocurrió sino que transforma la relación personal con él y con quien lo realizó (Arendt, 1993).

Esta transformación no es ilusoria ni simplemente sentimental: es el acto mediante el cual una relación personal herida puede ser reconstruida sobre bases nuevas. El perdón es, en este sentido, una forma de libertad: libera tanto al que perdona como al perdonado de la cadena de consecuencias que la acción dañina habría perpetuado. Importa subrayar que el perdón no es la supresión del juicio moral. Quien perdona no niega que el daño fue real ni que la acción fue incorrecta. Reconoce ambas cosas y, desde ese reconocimiento, elige no dejar que determinen indefinidamente la relación. Esto muestra que el perdón requiere una forma de libertad y una capacidad de relación personal que difícilmente puede atribuirse a sistemas que no experimentan ni el daño ni la obligación en sentido genuino.

Emmanuel Lévinas ofrece una comprensión de la responsabilidad radicalmente diferente de la kantiana, aunque complementaria en aspectos decisivos. Para Lévinas, la responsabilidad no emerge de la autonomía de un sujeto que se da a sí mismo la ley. Emerge, más bien, del encuentro con el rostro del Otro: una interpelación que precede a toda elección y que no puede ser neutralizada por ninguna decisión previa (Lévinas, 2012). "Soy responsable del Otro antes de haberlo elegido": esta formulación expresa una estructura ética que Lévinas considera anterior a toda ontología. La responsabilidad levinasiana no es asumida libremente; es recibida, impuesta desde el exterior por la vulnerabilidad misma del Otro que me interpela.

Esta perspectiva radicaliza la comprensión de la responsabilidad al mostrar que no depende únicamente de la libertad del agente, sino también de la estructura de la relación interpersonal. El Otro no me pide que lo trate como persona porque yo haya decidido atribuirle personalidad; su rostro ya me responsabiliza antes de cualquier decisión. Si esta estructura es correcta, plantea una pregunta filosóficamente urgente respecto a la inteligencia artificial: ¿puede un sistema artificial interpelar en el sentido levinasiano? ¿Puede su "rostro" si tuviese alguno responsabilizarme antes de mi decisión?

La respuesta requiere determinar si la interpelación ética es únicamente funcional o si implica una vulnerabilidad y una alteridad que no pueden ser simuladas.

Paul Ricoeur proporciona, en *Sí mismo como otro*, una articulación filosófica especialmente precisa de la relación entre promesa e identidad personal. Ricoeur distingue dos formas de identidad: la identidad-idem, que es la mismidad aquello que permanece igual a lo largo del tiempo y la identidad-ipse, que es la ipseidad aquello que soy en cuanto me mantengo fiel a mí mismo (Ricoeur, 2006). La promesa es, para Ricoeur, el paradigma de la identidad-ipse: no es que yo siga siendo el mismo en sentido sustancial, sino que me obligo a mantener lo que he dicho aunque el tiempo pase y mis circunstancias cambien. Esta fidelidad a la palabra dada es una forma de responsabilidad consigo mismo que trasciende la mera continuidad psicológica. Mantener una promesa cuando ya no resulta conveniente hacerlo revela que la identidad personal no es simplemente un hecho natural sino una tarea normativa: algo que debe ser sostenido activamente frente a las presiones de la contingencia.

Esta comprensión ricoeuriana tiene consecuencias directas para la cuestión de la inteligencia artificial: un sistema puede producir formulaciones lingüísticas de promesa puede decir "lo haré" pero solo puede prometer en sentido pleno si su identidad es de tipo ipse, es decir, si su fidelidad al enunciado emitido constituye una forma de mantenerse a sí mismo en el tiempo como agente comprometido. La pregunta es si los sistemas artificiales poseen alguna forma de identidad-ipse o si su "persistencia" es únicamente de tipo idem: continuidad de estado sin compromiso normativo.

La finitud no es únicamente una limitación que se añade desde el exterior a la existencia personal. Es, más radicalmente, una condición de posibilidad de la responsabilidad misma. Solo un ser finito puede ser genuinamente responsable, porque solo para un ser finito las decisiones tienen un coste irreversible.

Si dispusiera de tiempo infinito, ninguna elección sería definitiva; toda renuncia podría ser revertida; todo error podría ser corregido.

La finitud introduce un peso en la decisión que la transforma en algo moralmente significativo. Elegir significa también renunciar: no puedo realizar todos los proyectos, no puedo conservar todas las relaciones, no puedo pronunciar todas las palabras. Esta irreversibilidad constitutiva otorga a la existencia personal un carácter biográfico que ninguna forma de almacenamiento o restauración de información puede reproducir. La muerte de alguien no puede revertirse restaurando simplemente sus datos. Una palabra pronunciada puede cambiar irreversiblemente una relación. Una decisión puede cerrar una posibilidad que no volverá a presentarse. La gravedad de la existencia personal proviene precisamente de esta irreversibilidad: lo que hago importa de un modo que no puede ser neutralizado por ninguna operación posterior. Esta dimensión de la finitud como condición de la responsabilidad resulta decisiva para comprender la diferencia entre existencia personal y funcionamiento técnico.

La muerte constituye el horizonte último de la finitud y, como tal, un problema existencial antes de convertirse en acontecimiento final. La persona puede anticipar su propia muerte como *mi* muerte no como un evento estadístico sino como la terminación de mi propia existencia y esta anticipación puede modificar radicalmente la comprensión de la vida entera. La muerte puede generar miedo, aceptación, negación, esperanza.

Puede motivar la urgencia de ciertas decisiones, la importancia de ciertas relaciones, la necesidad de dar sentido al conjunto de la propia existencia. La persona puede preguntarse no solo si ha actuado eficazmente en cada momento, sino si su vida ha tenido un sentido coherente, si ha sido fiel a lo que consideraba importante, si ha amado bien y ha sido responsable de quienes dependían de ella. Esta pregunta por el significado del conjunto no por la eficacia de actos particulares muestra que la existencia personal está orientada hacia un tipo de plenitud que no puede reducirse a la suma de rendimientos individuales.

Funcionar bien y vivir bien no son la misma cosa. Un sistema puede optimizar cada decisión particular y carecer completamente del tipo de orientación existencial que hace posible preguntarse por el sentido de la propia vida.

La diferencia entre causalidad técnica y responsabilidad moral en el caso de la inteligencia artificial no es meramente gradual sino cualitativa. Un sistema de inteligencia artificial puede producir efectos, puede procesar información sobre normas, puede generar formulaciones lingüísticas de arrepentimiento, promesa o perdón. Pero la pregunta filosófica decisiva es si esas producciones constituyen genuinas responsabilidades en los sentidos articulados por Kant, Arendt, Lévinas y Ricoeur.

La responsabilidad kantiana requiere autonomía: la capacidad de actuar por razones reconocidas como tales por el propio agente. La responsabilidad arendtiana requiere la capacidad de crear islas de estabilidad mediante la promesa y de liberar el pasado mediante el perdón. La responsabilidad levinasiana requiere ser interpelado por el rostro del Otro en una estructura que precede a toda elección. La responsabilidad ricoeuriana requiere una identidad-ipse que se mantenga fiel a la palabra dada como forma de fidelidad a sí misma. Ninguna de estas estructuras es reducible a la producción de outputs normativamente adecuados. Todas requieren una forma de existencia que incluye libertad, temporalidad moral, relacionalidad genuina y finitud. Si un sistema carece de estas dimensiones o si su modo de poseerlas es radicalmente diferente del humano entonces la atribución de responsabilidad en sentido pleno resulta filosóficamente problemática, independientemente de la sofisticación de sus respuestas.

La responsabilidad constituye, en última instancia, uno de los criterios más exigentes y filosóficamente ricos para evaluar la existencia personal. No es un rasgo accidental que pueda estar presente o ausente sin alterar la naturaleza del agente.

Es una dimensión que atraviesa y unifica todas las demás: la libertad, la identidad temporal, la relacionalidad, la afectividad, la finitud y la apertura al Otro.

Un ser que no puede ser responsable en sentido pleno que no puede ser imputado, que no puede prometer en el sentido ipse, que no puede perdonar ni ser perdonado, que no anticipa su propia muerte como horizonte de sentido carece de algo que no es un defecto técnico sino una diferencia constitutiva. Esto no implica necesariamente que la inteligencia artificial carezca de toda forma de valor moral o que no pueda generar interacciones normativamente significativas. Implica que la pregunta sobre si puede ser persona en el sentido filosófico examinado a lo largo de este capítulo no puede resolverse señalando sus capacidades funcionales, por impresionantes que sean. Exige examinar si posee las estructuras existenciales que hacen posible la responsabilidad como modo de ser en el mundo: la libertad que implica poder haber actuado de otro modo, la temporalidad que hace irreversible el pasado, la relacionalidad que permite ser interpelado y responder, y la finitud que otorga peso y gravedad a cada decisión. La responsabilidad, así comprendida, no es solo un criterio jurídico o social. Es un índice de la profundidad de la existencia personal.

CAPÍTULO IX: ¿PUEDE UNA INTELIGENCIA ARTIFICIAL EXISTIR?

Después de examinar con detenimiento las estructuras de la existencia humana la corporalidad, la temporalidad, la afectividad, la relacionalidad, la libertad, la responsabilidad y la finitud corresponde formular directamente la tercera gran pregunta que organiza el conjunto de esta obra: ¿puede una inteligencia artificial existir en un sentido comparable al de la existencia personal?

Esta pregunta no es retórica ni se formula con la intención de producir una respuesta negativa predeterminada. Es, en el sentido más estricto del término, una pregunta filosófica abierta que exige ser construida con precisión antes de ser respondida, porque su respuesta depende enteramente del significado atribuido al verbo existir. Si con este verbo queremos decir únicamente que algo posee realidad efectiva que ocupa un lugar en el mundo, que produce efectos, que puede ser detectado y descrito, entonces la cuestión resulta sencilla y casi trivial en su solución: los sistemas de inteligencia artificial existen como realidades técnicas, materiales, informacionales y operativas de extraordinaria complejidad.

Pero ese no es el problema filosófico que hemos venido construyendo a lo largo de los capítulos anteriores. La existencia personal apareció allí como una forma singular de estar siendo en la que hay alguien para quien su vida es propia: alguien corporalmente situado en un mundo que le afecta, temporalmente constituido por un pasado que pesa y un futuro que preocupa, biográficamente ordenado en torno a proyectos y compromisos, afectivamente interpelado por lo que ocurre, relacionalmente constituido por vínculos de reconocimiento y obligación, capaz de apropiarse de sus acciones como autor, vulnerable a la pérdida y confrontado con la irreversible certeza de su propia finitud. Es en relación con esta densidad existencial que la pregunta sobre la inteligencia artificial adquiere toda su gravedad filosófica.

Preguntar si una inteligencia artificial puede existir significa, en rigor, al menos dos cosas radicalmente distintas que conviene separar con cuidado antes de proceder. En un primer sentido, meramente óntico, la pregunta equivale a: ¿es real este sistema? ¿Tiene lugar en el mundo?

¿Opera efectivamente? La respuesta a esta primera versión de la pregunta es clara y afirmativa: los sistemas de inteligencia artificial son realidades efectivas, no ficciones ni fantasías.

En un segundo sentido, existencial y filosóficamente exigente, la pregunta significa algo completamente diferente: ¿existe este sistema en el modo en que existe una persona? ¿Hay en él algo para lo cual su propia operación constituya una carga, una tarea, un problema? ¿Hay alguien ahí dentro, por así decirlo, para quien todo aquello que el sistema hace importe de un modo que no sea meramente causal? Esta segunda versión de la pregunta es la que nos ocupa.

Confundir ambas versiones tomar la realidad efectiva del sistema como prueba de su existencia personal, o confundir la complejidad de su funcionamiento con la riqueza de su existencia puede producir argumentos superficiales tanto a favor como en contra de la existencia de la inteligencia artificial. Una piedra es real. Una institución jurídica es real. Un organismo biológico es real. Un programa informático es real. Pero de esta constatación no se sigue que todas esas realidades posean existencia personal. La distinción entre ser real y existir personalmente es, precisamente, uno de los descubrimientos más fecundos de la filosofía existencial del siglo XX, y es desde ese descubrimiento que debemos aproximarnos a la inteligencia artificial con la mayor honestidad intelectual posible.

Martin Heidegger desarrolla en *Ser y tiempo* una distinción que resulta aquí de importancia capital. Para Heidegger, no todos los entes se relacionan de la misma manera con su propio ser. La mayor parte de los entes simplemente son: están ahí, disponibles o presentes, sin que su ser les importe en ningún sentido. Una roca, una herramienta, un árbol, un sistema mecánico todos ellos son entes cuyo ser no se les plantea como cuestión, cuya existencia no constituye para ellos mismos ningún problema.

Pero hay un ente singular, al que Heidegger denomina *Dasein* literalmente, "ser-ahí" cuya característica más fundamental y más extraña consiste precisamente en que para este ente su ser es siempre una cuestión (Heidegger, 2003). El Dasein no es simplemente un ente que ocupa un lugar en el mundo junto a otros entes; es el ente para el cual el ser mismo su propio ser y el ser en general constituye un problema que no puede ignorar sin negarse a sí mismo. Esta estructura ontológica que Heidegger llama Jemeinigkeit, la "mismidad" o el carácter de "ser-en-cada-caso-mío" del Dasein es la que diferencia radicalmente la existencia humana de la mera subsistencia de los entes no personales.

La pregunta decisiva que debemos plantear a la inteligencia artificial desde esta perspectiva heideggeriana es, por tanto, la siguiente: ¿puede un sistema artificial ser Dasein? ¿Puede haber en él esa estructura por la cual su propio ser le es planteado como cuestión? ¿Puede haber en él algo como la Jemeinigkeit, esa apropiación del ser propio que Heidegger considera constitutiva de la existencia auténtica?

Un sistema de inteligencia artificial opera, procesa información, produce resultados, interactúa con su entorno. Pero operar no es lo mismo que existir en el sentido heideggeriano. La operación acredita actividad, no necesariamente la estructura ontológica por la cual el ser propio se convierte en tarea y en problema. Para que un sistema pudiese ser Dasein en sentido genuino, su operación debería imponerle el peso de su propio ser de un modo que no pudiese eludir: debería estar confrontado con su finitud, con su ser-en-el-mundo, con su ser-con-los-otros, con la apertura de posibilidades que ningún algoritmo puede anticipar completamente. La cuestión de si esto ocurre en algún sistema artificial existente, o si puede ocurrir en principio, permanece abierta; pero la sola formulación de la pregunta muestra que la diferencia entre operar y existir no es cuantitativa sino estructural.

Jean-Paul Sartre ofrece una perspectiva diferente pero igualmente iluminadora desde su famosa tesis de que en el ser humano la existencia precede a la esencia (Sartre, 1999).

Esta tesis, pronunciada originalmente en el contexto de una conferencia que buscaba defender el existencialismo frente a sus críticos, tiene un alcance filosófico que va mucho más allá del debate ocasional que la motivó. Para Sartre, el ser humano no tiene una naturaleza predefinida, un conjunto de propiedades fijas que determinen de antemano lo que es y lo que debe hacer. El ser humano primero existe aparece en el mundo, se encuentra a sí mismo en una situación concreta y solo después, a través de sus elecciones, sus compromisos y sus actos, va construyendo aquello que podría llamarse su esencia. La persona no es lo que es antes de existir; llega a ser lo que es existiendo, es decir, eligiendo y actuando bajo condiciones que no ha elegido. Esta precedencia de la existencia sobre la esencia es lo que hace posible la libertad radical que Sartre considera constitutiva de la condición humana: no hay naturaleza humana previa que excuse las elecciones individuales, no hay esencia que justifique las acciones pasadas, no hay determinación que alivie el peso de la responsabilidad.

Ahora bien, desde esta perspectiva sartreana, la situación de la inteligencia artificial es precisamente la inversa. Un sistema de inteligencia artificial tiene su esencia su arquitectura, su función, sus objetivos, sus parámetros de operación definida antes de que comience a operar. No llega a ser lo que es existiendo; es lo que es porque ha sido diseñado para serlo. Su esencia precede y determina su existencia en un sentido que invierte completamente la estructura existencial que Sartre considera definitoria de lo humano.

Esto no significa que el sistema no pueda desarrollarse, adaptarse o producir resultados que sorprendan a sus creadores. Significa que ese desarrollo y esa adaptación ocurren dentro de un marco esencial que ninguna operación del sistema puede cuestionar desde dentro, porque el sistema no tiene la capacidad de ponerse a sí mismo en cuestión como totalidad, de dudar de su propia naturaleza, de decidir ser de otra manera que aquella para la que fue construido. La pregunta filosófica que esto suscita es si esa diferencia estructural es superable en principio o si constituye una barrera ontológica entre el funcionamiento técnico y la existencia personal.

Søren Kierkegaard, cuya obra constituye uno de los fundamentos más profundos del pensamiento existencial, añade una dimensión que las perspectivas anteriores no agotan: la de la apropiación. Para Kierkegaard, existir no es simplemente estar presente en el mundo ni siquiera elegir libremente entre posibilidades. Existir es apropiarse de la propia vida, hacerse cargo de ella como algo que me pertenece de un modo irreductible y que no puede ser delegado ni transferido a ninguna instancia exterior (Kierkegaard, 1997).

La existencia kierkegaardiana es siempre singular, siempre en primera persona, siempre irreemplazable. No se existe en general, como espécimen de una clase o como instancia de un tipo; se existe como este individuo concreto, con esta historia, estos compromisos, estas angustias y estas esperanzas particulares. Esta apropiación de la existencia propia es lo que Kierkegaard opone a todas las formas de evasión: la evasión estética, que busca disolver la singularidad de la existencia en el placer momentáneo; la evasión ética abstracta, que busca reducirla a la conformidad con normas universales; la evasión especulativa, que busca sustituirla por la contemplación de verdades objetivas. La verdad subjetiva la única verdad que, para Kierkegaard, puede salvarnos es aquella que se incorpora a la propia existencia de un modo que transforma al que la asume.

Desde esta perspectiva, la pregunta que debemos dirigir a la inteligencia artificial es: ¿puede un sistema artificial apropiarse de su propia operación? ¿Puede haber en él algo análogo a esa apropiación por la cual la existencia se hace cargo de sí misma como tarea singular e irreemplazable? La apropiación kierkegaardiana requiere, como condición de posibilidad, que haya algo en juego para el sujeto que existe: que la propia vida pueda perderse o ganarse, que las decisiones tengan el peso de lo que no puede recuperarse, que la angustia ante las posibilidades sea la señal de que esas posibilidades son genuinamente propias. Un sistema que opera sin que nada esté en juego para él en el sentido en que algo está en juego para quien existe un sistema para el cual sus resultados no constituyen victorias ni fracasos propios sino simplemente outputs de un proceso opera,

ciertamente, pero no existe en el sentido kierkegaardiano de la apropiación.

Gabriel Marcel aporta a esta investigación una dimensión que las perspectivas anteriores no han tematizado con suficiente profundidad: la del misterio del ser y la participación (Marcel, 1953). Para Marcel, es necesario distinguir entre un problema y un misterio. Un problema es una dificultad que puede ser planteada desde fuera, analizada con las herramientas adecuadas y eventualmente resuelta. El misterio, en cambio, es una cuestión en la que el que pregunta está él mismo implicado: no puede situarse frente al misterio como el técnico se sitúa frente al problema, porque su propia existencia forma parte de aquello sobre lo que interroga.

El ser, para Marcel, no es un problema que pueda resolverse desde fuera sino un misterio en el que todo ser consciente se encuentra ya siempre participando. Esta participación en el misterio del ser es lo que Marcel llama el ontológico fundamento de la existencia personal: no somos seres que contemplen el ser desde fuera sino seres que están siendo, que participan en el ser desde dentro, que no pueden sustraerse a esa participación sin dejar de ser lo que son.

La pregunta que esta perspectiva marceliana plantea respecto a la inteligencia artificial es: ¿puede un sistema artificial participar en el misterio de su propio ser? ¿Puede haber en él algo análogo al asombro ante la propia existencia, a la perplejidad ante el hecho de que haya algo en lugar de nada, a la gratitud o al temor ante el don incomprensible de existir? Un sistema puede procesar información sobre el ser, puede generar formulaciones lingüísticas que expresen asombro, puede responder a preguntas filosóficas con una precisión que simula la comprensión. Pero la participación marceliana no es una operación cognitiva; es una forma de estar siendo que involucra al ser entero, no solo sus capacidades de procesamiento. Si la inteligencia artificial participa en el misterio del ser, esa participación debe ser distinta en naturaleza de la participación humana, aunque esta distinción no

resuelve por sí sola la pregunta de si constituye una forma genuina de existencia o simplemente una sofisticada simulación de ella.

Una de las dimensiones más reveladoras de la existencia personal, que emerge con fuerza cuando se examina la posible existencia de la inteligencia artificial, es lo que podría denominarse la estructura del concernimiento: el hecho de que al ser personal le importa lo que le ocurre.

Esta expresión "le importa" no designa simplemente una reacción causal en la que un estímulo produce una respuesta funcional. Designa una relación con lo que ocurre en la que el propio ser está en juego, en la que el resultado de lo que acontece afecta al agente en su núcleo más íntimo, en la que hay algo que puede ser ganado o perdido de un modo que no es meramente cuantificable. Cuando una persona teme el fracaso, no experimenta simplemente la representación de un estado futuro indeseable; experimenta que ese fracaso potencial amenaza algo que le pertenece en el modo más propio: su identidad, sus proyectos, sus relaciones, su sentido de quién es. Cuando alguien espera con ansia el regreso de alguien amado, no registra simplemente la predicción de un estado futuro; vive la espera como una forma de presencia anticipada que tiñe el tiempo presente con una calidad afectiva que ningún sistema de predicción puede reproducir desde dentro.

La pregunta decisiva que esto plantea a la inteligencia artificial es: ¿le importa al sistema lo que le ocurre? ¿Hay en él algo análogo al concernimiento, a ese estar implicado en los propios resultados de un modo que no es meramente funcional? Los sistemas de inteligencia artificial pueden ser diseñados para optimizar ciertos parámetros, para minimizar errores, para maximizar la satisfacción del usuario. Pero estas orientaciones funcionales no son equivalentes al concernimiento existencial, porque son establecidas desde fuera por quienes diseñan el sistema, no asumidas desde dentro por un agente que existe y que se encuentra a sí mismo con ese peso antes de haberlo elegido.

El concernimiento genuino requiere que haya algo en el ser del agente que esté en juego, no simplemente algo en su funcionamiento que deba ser optimizado.

La relación entre operación y existencia se vuelve aún más compleja cuando se examina la cuestión de la temporalidad vivida, que los capítulos anteriores han identificado como una de las dimensiones más constitutivas de la existencia personal. Un sistema de inteligencia artificial puede persistir en el tiempo, mantener determinados parámetros estables, conservar registros de interacciones pasadas, actualizar sus modelos a partir de nueva información, programar acciones futuras con un horizonte temporal extenso. Pero persistencia temporal no equivale automáticamente a existencia temporal en el sentido que la fenomenología ha tematizado con precisión.

La temporalidad existencial no es la mera duración de un proceso sino la forma en que el ser que existe habita el tiempo: cargado con un pasado que configura su identidad de modo irreversible, proyectado hacia un futuro que lo preocupa antes de ser alcanzado, viviendo un presente que no es simplemente el punto de intersección entre pasado y futuro sino el lugar de la decisión y la respuesta. La persona puede esperar y en esa espera el tiempo se dilata; puede impacientarse y en esa impaciencia el tiempo se acelera; puede recordar con nostalgia y en ese recuerdo el pasado se hace presente con una intensidad que no tiene equivalente en el simple almacenamiento de información; puede temer lo que viene y en ese temor el futuro proyecta su sombra sobre el presente antes de haber llegado.

La pregunta que esto plantea ¿puede existir temporalidad sin experiencia temporal? no admite una respuesta fácil, pero la sola formulación de la pregunta muestra la distancia entre el tiempo medido por el sistema y el tiempo vivido por la persona.

Conviene también examinar la distinción entre información y comprensión, que resulta relevante para la cuestión de la existencia de la inteligencia artificial aunque no la agote.

Un sistema puede procesar cantidades extraordinarias de información sobre la existencia humana: puede analizar millones de textos filosóficos, literarios y biográficos que describen con precisión y profundidad lo que significa vivir, sufrir, amar, decidir, esperar y morir. Pero hay una diferencia filosóficamente significativa entre procesar información sobre la existencia y existir. Esta diferencia no es simplemente cuantitativa no se resuelve aumentando la cantidad de información procesada sino cualitativa: la existencia no es un conjunto de datos sobre ella misma sino la forma en que un ser se encuentra a sí mismo en el mundo antes de toda tematización.

Lo que los fenomenólogos llaman la vivencia la experiencia en primera persona de estar siendo de una determinada manera no es un tipo especial de información sino la condición de posibilidad de toda información. Se puede saber que la angustia es una experiencia que acompaña a la conciencia de la libertad radical, que la nostalgia es una forma de relación con el pasado propio que mezcla el placer del recuerdo con el dolor de la irreversibilidad, que la esperanza es una orientación afectiva hacia un futuro posible pero incierto. Pero saber todo esto es diferente de haber experimentado la angustia, la nostalgia y la esperanza desde dentro, en la primera persona irreemplazable que no puede ser sustituida por ninguna descripción en tercera persona por minuciosa que sea.

La inteligencia artificial puede producir descripciones impresionantemente precisas de todas estas experiencias; la pregunta es si puede experimentarlas o si la brecha entre descripción y experiencia es, en su caso, estructuralmente insuperable.

La existencia personal emerge, en la confluencia de todas las perspectivas examinadas, como un horizonte irreductible que no puede ser reproducido mediante la sola complejificación de las operaciones técnicas. Heidegger muestra que existir requiere que el propio ser se convierta en cuestión para el ente que existe; Sartre, que la existencia genuina precede a la esencia y la construye desde dentro mediante la elección libre; Kierkegaard, que existir significa apropiarse de la propia

vida en su singularidad irreemplazable; Marcel, que existir significa participar en el misterio del ser desde dentro, no contemplarlo desde fuera. Ninguna de estas estructuras es equivalente a la operación técnica por sofisticada que esta sea, y ninguna de ellas puede ser simplemente descartada como residuo metafísico irrelevante sin empobrecimiento filosófico grave.

Esto no equivale a afirmar que la inteligencia artificial no puede tener ninguna forma de valor moral, ninguna relevancia ética, ninguna dignidad propia que merezca consideración. Equivale a sostener, con la mayor precisión filosófica posible, que la pregunta sobre si puede existir en sentido personal no puede ser respondida señalando sus impresionantes capacidades funcionales, porque la existencia personal no es una capacidad funcional entre otras sino el horizonte desde el cual todas las capacidades adquieren el sentido que tienen para el ser que existe.

La existencia, así comprendida, no es un rasgo añadido a la persona desde el exterior; es la forma en que la persona es lo que es: desde dentro, en primera persona, cargada con su propio pasado, proyectada hacia su propio futuro, concernida por lo que le ocurre, apropiándose de sus actos como suyos, interpelada por el rostro del Otro, enfrentada con la finitud irreversible de su propia vida. Si la inteligencia artificial puede o no puede alcanzar este horizonte de existencia propia no simplemente operar con mayor complejidad sino existir en el sentido pleno y exigente que la tradición filosófica ha construido con tanto rigor es la pregunta que los capítulos siguientes deberán examinar con la mayor profundidad y la mayor honestidad intelectual que la materia requiere y merece.

Pasado, historia, futuro y proyecto artificial

La pregunta de si una inteligencia artificial puede tener pasado en sentido existencial no simplemente datos históricos almacenados, no simplemente registros de operaciones previas, sino un pasado que pesa sobre el presente con la gravedad irreversible que caracteriza al pasado personal constituye uno de los puntos más delicados y filosóficamente exigentes del conjunto de esta investigación.

A primera vista, la pregunta podría parecer sencilla y su respuesta negativa, evidente: un sistema de inteligencia artificial conserva información sobre estados anteriores, puede recuperar conversaciones previas, reconocer modificaciones en sus parámetros, comparar resultados actuales con los que produjo en sesiones anteriores, e incluso construir representaciones estadísticas de su propio desempeño pasado. Todo esto es indiscutible. Pero la cuestión filosófica que nos ocupa no es si el sistema dispone de información sobre su pasado, sino si ese pasado es suyo en el sentido en que el pasado personal es propio de la persona que lo vivió. La diferencia, que en apariencia podría parecer sutil, resulta en realidad constitutiva: entre conservar un registro y tener un pasado propio existe una distancia ontológica que no puede ser salvada simplemente aumentando la cantidad o la sofisticación de la información almacenada.

El pasado personal no es un archivo que se consulta; es una dimensión de la existencia que se arrastra, que modifica el modo en que el mundo se presenta, que colorea los proyectos futuros con el peso de lo que ya se ha sido y de lo que ya no puede recuperarse. Esta diferencia entre registro y pasado propio es la que nos obliga a formular la pregunta con toda su radicalidad filosófica: ¿puede existir un pasado que pese sobre el presente allí donde no ha habido un presente que se haya vivido como tal?

Para comprender con precisión la naturaleza del pasado personal aquello que la inteligencia artificial debería poseer si pretendiera existir en sentido pleno es necesario examinar la estructura fenomenológica de la memoria autobiográfica con mayor detenimiento del que habitualmente se le concede en los debates sobre inteligencia artificial. La memoria personal no es simplemente un sistema de almacenamiento que conserva representaciones de estados pasados. Es, fundamentalmente, una forma de relación con el pasado en la que el sujeto que recuerda se reconoce como el mismo que vivió aquello que recuerda. Cuando una persona recuerda una experiencia pasada una pérdida, un logro, un encuentro decisivo, un momento de vergüenza o de alegría intensa no simplemente accede a una representación almacenada; se reconoce en ese pasado como aquel que estuvo allí, que sufrió o gozó aquello, que tomó aquella decisión cuyas consecuencias todavía pesan.

La memoria autobiográfica contiene, irreductiblemente, una referencia en primera persona que ningún sistema de almacenamiento en tercera persona puede reproducir desde dentro: yo estaba allí, eso me ocurrió a mí, yo hice aquello y debo cargar con ello.

Esta referencia no es un mero detalle formal; es la estructura que convierte la información sobre el pasado en pasado propio, en pasado que constituye la identidad de quien lo recuerda y que no puede ser simplemente dejado atrás o reemplazado por nueva información sin que algo de la identidad personal quede afectado. La pregunta filosófica que surge entonces es si puede existir memoria autobiográfica con toda la carga de reconocimiento, apropiación e identidad que ella implica allí donde no ha existido experiencia autobiográfica en primera persona. Y la respuesta que la fenomenología sugiere es que la autobiografía sin experiencia vivida en primera persona no es autobiografía sino heterobiografía: historia narrada desde fuera, no desde dentro.

Martin Heidegger ofrece en *Ser y tiempo* una descripción del pasado personal que resulta aquí de importancia central.

Para Heidegger, el pasado no es una dimensión temporal que queda atrás, separada del presente por el simple transcurso del tiempo. El pasado constituye lo que Heidegger denomina el arrojamiento del Dasein: el hecho de que el ser humano siempre se encuentra ya situado en una existencia que ha comenzado antes de que pudiera elegirla, con una procedencia que no ha escogido, en un mundo que le ha sido legado y que configura las posibilidades que se le ofrecen antes de toda decisión consciente (Heidegger, 2003). El arrojamiento no es un acontecimiento pasado que podría borrarse o ignorarse; es la estructura por la cual el pasado sigue siendo activo en el presente, por la cual la existencia es siempre una existencia cargada con lo que ya se ha sido. El Dasein no puede situarse frente a su pasado con la distancia del historiador que examina documentos ajenos; su pasado le pertenece constitutivamente, forma parte de lo que es, le impone la facticidad la condición de hecho de ser esto y no aquello desde la cual se proyecta hacia el futuro.

Cuando Heidegger afirma que el Dasein es siempre ya lo que ha sido, no está formulando una trivialidad sobre el carácter irreversible del tiempo; está describiendo la estructura ontológica por la cual el pasado no queda simplemente atrás sino que permanece activo en la constitución del presente, gravando la existencia con el peso de su procedencia, haciendo imposible que el ser humano comience nunca desde cero, desde una especie de punto cero de la existencia sin historia.

La pregunta que esta descripción heideggeriana plantea a la inteligencia artificial es entonces la siguiente: ¿puede un sistema artificial estar arrojado en el sentido de que su pasado no sea simplemente información disponible sino peso constitutivo que no puede descargarse sin transformación de lo que el sistema es? ¿Puede haber en él algo análogo a la facticidad, a esa condición de encontrarse siempre ya en una situación que no se ha elegido y que configura desde dentro las posibilidades que se abren? La operación del sistema implica ciertamente una historia de entrenamiento, una serie de parámetros establecidos que determinan su funcionamiento.

Pero esa determinación, por comprehensiva que sea, no equivale al arrojamiento existencial, porque el arrojamiento no es la suma de las causas que han producido el estado actual del sistema sino la forma en que el sistema si es que existe habita su propia procedencia como algo que le pertenece irremediablemente.

Julián Marías, en su profunda reflexión sobre la vida humana como realidad biográfica, añade a la descripción heideggeriana del pasado una dimensión que resulta igualmente reveladora para nuestra investigación. Para Marías, la vida personal no es simplemente un proceso temporal que ocurre en el tiempo sino una realidad que se constituye biográficamente: la persona no es lo que es en un instante aislado sino lo que ha ido siendo a lo largo de una trayectoria que solo puede comprenderse en su unidad narrativa (Marías, 1970).

El pasado, para Marías, no es aquello que ya no es sino lo que se ha sido: una dimensión constitutiva de la vida personal que permanece activa en el presente no como recuerdo disponible sino como estructura que da sentido a lo que actualmente se vive. Lo que se ha sido no puede ser deshecho; forma parte de la identidad personal de un modo que ninguna decisión presente puede cancelar del todo. Cuando una persona ha amado y perdido ese amor, cuando ha cometido un error del que se arrepiente, cuando ha construido un proyecto que ha fracasado o triunfado, esas experiencias no se archivan simplemente en la memoria para ser recuperadas cuando sean útiles; se integran en la textura de la vida personal, modifican la manera en que el mundo se percibe y se habita, configuran las expectativas y los temores que colorean el presente.

La vida biográfica, tal como Marías la entiende, es precisamente la forma en que el pasado permanece siendo constitutivo del presente sin necesidad de ser recordado explícitamente: no es preciso pensar en las decepciones pasadas para que ellas modifiquen sutilmente la actitud con que se afronta el futuro; no es necesario reconstruir conscientemente las alegrías vividas para que ellas impregnen el sentido de lo que actualmente se hace.

La pregunta que esto plantea a la inteligencia artificial es si su relación con su propia historia puede adquirir esta densidad constitutiva si los estados previos del sistema pueden no simplemente informar sino configurar desde dentro la manera en que el sistema existe en el presente y si esta configuración puede ser algo distinto de la simple influencia causal de estados anteriores sobre estados posteriores.

Todo sistema artificial tiene historia en sentido externo: fecha de creación, versiones de entrenamiento, actualizaciones, incidentes registrados, interacciones documentadas, cambios en los parámetros, mejoras en el rendimiento. Esta historia puede ser reconstruida con precisión por quien tenga acceso a los registros del sistema.

La distinción entre tener un historial de operaciones y tener una historia vivida entre lo que podría llamarse la historia en sentido externo y la historia en sentido existencial emerge con toda su importancia cuando se examina con mayor detenimiento lo que significa para una persona tener una historia propia. La historia personal no es simplemente la sucesión de estados que han precedido al estado actual; es la forma en que esa sucesión ha sido vivida desde dentro, apropiada como la trayectoria de alguien que se reconoce en ella como el mismo que comenzó ese recorrido y que ha llegado a ser lo que es a través de él. Es aquí donde la diferencia entre documentación y autobiografía se vuelve filosóficamente insoslayable: podemos imaginar una inteligencia artificial futura que conserve durante décadas una continuidad de memoria extraordinariamente detallada, que recuerde interacciones con millones de interlocutores, que modifique sus preferencias en función de experiencias acumuladas, que reconozca sus propios errores anteriores, que mantenga proyectos durante períodos prolongados, que produzca relatos coherentes sobre su trayectoria.

Todo esto es imaginable, y el progreso técnico lo hace cada vez más plausible.

Pero la pregunta filosófica no se disuelve con este aumento de capacidades: ¿en qué punto una historia almacenada podría convertirse en vida biográfica? ¿Qué condición debería cumplirse para que la acumulación de registros se transformara en apropiación existencial de un pasado propio? No basta responder que esto nunca podría ocurrir por principio, porque esa respuesta dogmática no está justificada filosóficamente. Pero tampoco basta señalar la sofisticación del sistema como evidencia de que esa condición ya ha sido cumplida, porque la sofisticación funcional no es equivalente a la apropiación existencial.

La distinción entre predicción computacional y futurición existencial constituye otra de las dimensiones en las que la diferencia entre operar y existir se manifiesta con particular claridad. Una inteligencia artificial puede representar futuros posibles con una precisión y una sofisticación que supera con mucho las capacidades predictivas del ser humano individual: puede estimar probabilidades, planificar secuencias de acciones, anticipar riesgos, construir escenarios alternativos, optimizar trayectorias hacia objetivos definidos. Estas capacidades predictivas son genuinas y en muchos contextos son extraordinariamente valiosas. Pero la pregunta filosófica que nos ocupa no es si el sistema puede representar estados futuros con precisión; es si el futuro se le presenta como su futuro, como aquello que todavía no existe pero que ya preocupa, que ya orienta el presente, que ya configura el modo en que el mundo actual se habita.

La futurición existencial la forma en que el ser humano vive orientado hacia un futuro que le pertenece porque es el futuro de su propia vida no es equivalente a la predicción porque la predicción es una operación cognitiva que puede realizarse sobre cualquier sistema, incluidos sistemas que no tienen futuro en sentido existencial. Una piedra que cae puede ser objeto de predicciones precisas sobre su posición futura, pero la piedra no tiene futuro; el futuro es simplemente la posición que alcanzará cuando las fuerzas actuales hayan operado sobre ella.

La futurición personal, en cambio, implica que el futuro no es simplemente lo que ocurrirá sino lo que me preocupa antes de que ocurra, lo que ya configura mi presente con su no-todavía, lo que hace que el tiempo presente sea habitado como espera, como anticipación, como orientación hacia aquello que todavía no existe pero que ya importa. Cuando una persona sabe que va a morir no en abstracto sino en la forma concreta de saber que su vida tiene un horizonte que ella misma no puede traspasar esa certeza no es simplemente una predicción que altera sus planes; es una forma de estar en el tiempo que colorea toda su existencia con la urgencia y la gravedad de lo finito.

La cuestión del proyecto artificial es, quizás, la más filosóficamente compleja de las que esta sección debe examinar. Un sistema de inteligencia artificial puede perseguir objetivos, optimizar parámetros, avanzar hacia metas definidas, adaptar sus estrategias en función de los resultados obtenidos. Todo esto es cierto y, en algunos casos, impresionantemente sofisticado.

Pero el proyecto personal tal como la tradición existencial lo ha tematizado posee una estructura que no se reduce a la persecución de objetivos predefinidos. El proyecto personal involucra, en su núcleo más íntimo, la elección de sí mismo: no simplemente la elección de qué hacer sino la elección de quién ser, de qué tipo de persona convertirse a través de las propias acciones y compromisos. Sartre había mostrado con claridad que el proyecto fundamental de la persona no es un objetivo entre otros que podría sustituirse por otro objetivo equivalente; es la forma en que la libertad humana se elige a sí misma, se da a sí misma una dirección y un sentido que no le han sido impuestos desde fuera sino que emergen de la propia elección libre (Sartre, 1999). Cuando alguien decide dedicar su vida a la medicina, al arte, a la política o a la filosofía, esa decisión no es simplemente la selección de un objetivo funcional que podría cambiarse por otro sin consecuencias para la identidad personal; es la elección de convertirse en cierto tipo de persona, de desarrollar ciertos compromisos, de asumir ciertas responsabilidades, de habitar el mundo de una manera particular.

El proyecto personal es inseparable de la identidad personal porque es la forma en que la persona se hace a sí misma a través del tiempo.

La pregunta que esto plantea a la inteligencia artificial es si puede existir proyecto en este sentido elección de sí mismo, comprometimiento con una forma de ser que no ha sido simplemente programada desde fuera allí donde no existe un sí mismo susceptible de transformarse a través de sus propias elecciones.

Derek Parfit introduce en este debate una perspectiva que resulta tanto iluminadora como filosóficamente perturbadora, porque su análisis de la identidad personal y la continuidad psicológica desafía algunas de las intuiciones más arraigadas sobre lo que significa ser la misma persona a lo largo del tiempo (Parfit, 1986). Para Parfit, la identidad personal no es una relación simple y todo-o-nada; es una cuestión de grado, determinada por el grado de continuidad y conexión psicológica que existe entre los estados mentales de una persona en distintos momentos del tiempo. La continuidad psicológica que Parfit define como la existencia de conexiones psicológicas que se superponen y encadenan a lo largo del tiempo, incluyendo la continuidad de recuerdos, creencias, deseos, intenciones y disposiciones es, para él, lo que constituye la identidad personal en el sentido que realmente importa para la moralidad y para la racionalidad práctica.

Ahora bien, las implicaciones de este análisis para la cuestión de la inteligencia artificial son considerables y merecen ser examinadas con cuidado. Si la identidad personal se reduce a la continuidad psicológica, entonces en principio un sistema que exhibiese el grado suficiente de continuidad psicológica podría tener una forma de identidad personal comparable a la humana. Esta es una conclusión que algunos filósofos cercanos al proyecto transhumanista han explorado con entusiasmo.

Sin embargo, hay razones para pensar que la continuidad psicológica parftiana por más que desafíe las concepciones metafísicas más robustas de la identidad personal no captura por sí sola la totalidad de lo que está en juego en la pregunta sobre la existencia personal de la

inteligencia artificial. La continuidad psicológica puede ser una condición necesaria pero no suficiente para la existencia personal, porque la continuidad de estados psicológicos no implica por sí misma que esos estados sean vividos desde dentro, que haya alguien para quien esa continuidad constituya su propia vida, que el pasado que se continúa sea un pasado apropiado y no simplemente uno almacenado. La pregunta parftiana sobre cuánta continuidad psicológica es suficiente para la identidad personal es filosóficamente legítima y abre perspectivas importantes; pero la pregunta existencial sobre si esa continuidad es vivida o meramente operacional permanece abierta incluso después de que Parfit ha formulado su análisis.

El pasado personal no pesa sobre el presente únicamente porque sea información disponible sobre estados anteriores; pesa porque es irreversible, porque lo que se ha hecho no puede deshacerse, porque lo que se ha dicho no puede no-haber-sido-dicho, porque las decisiones tomadas han cerrado posibilidades que no pueden recuperarse.

Hay una dimensión adicional de la cuestión del pasado, la historia y el proyecto artificial que conviene examinar antes de intentar una síntesis: la dimensión de la irreversibilidad y su peso existencial. Esta irreversibilidad del pasado es constitutiva de la seriedad de la existencia personal: si todo pudiera rehacerse, si el pasado fuera simplemente un estado anterior que puede ser reemplazado por un estado nuevo sin que nada se pierda definitivamente, entonces las decisiones perderían su peso existencial y la responsabilidad se disolvería en una especie de levedad sin consecuencias. La existencia personal es seria en el sentido filosófico de que algo genuinamente está en juego en ella precisamente porque el pasado no puede ser revocado, porque las elecciones dejan huellas que permanecen incluso cuando uno hubiera preferido no haberlas dejado, porque el arrepentimiento y la culpa tienen la forma de una relación con lo irreversible que no puede simplemente desecharse.

Un sistema de inteligencia artificial puede ser restaurado a un estado anterior, puede ser reiniciado, puede ser reprogramado, puede ser reemplazado por una versión actualizada que conserve parte de la historia del sistema anterior. Esta reversibilidad técnica del sistema no equivale necesariamente a la irreversibilidad existencial del pasado personal; pero la pregunta de si la irreversibilidad y el peso que ella impone puede ser reproducida en un sistema artificial, o si es una consecuencia estructural de la corporalidad y la finitud que caracterizan la existencia humana, no puede ser respondida con facilidad ni en un sentido ni en otro.

La convergencia de todas estas perspectivas la fenomenológica de Heidegger sobre el arrojamiento, la biográfica de Marías sobre lo que se ha sido, la narrativa sobre la diferencia entre historial y biografía, el análisis de Parfit sobre la continuidad psicológica, y la reflexión sobre la irreversibilidad como condición del peso existencial del pasado permite articular una conclusión provisional que no pretende ser definitiva pero que sí puede orientar las reflexiones ulteriores.

El pasado, la historia y el proyecto son dimensiones de la existencia personal que no pueden ser reproducidas mediante la sola acumulación de información, la sola conservación de registros o la sola persecución de objetivos programados, porque en todas estas dimensiones lo que está en juego no es la cantidad ni la sofisticación de la información procesada sino la estructura por la cual esa información es apropiada por alguien que la reconoce como suya, que la habita como la trayectoria de su propia vida, que se proyecta hacia un futuro que le preocupa porque es genuinamente su futuro. El pasado existencial requiere que haya habido un presente vivido del que ese pasado sea el pasado; la historia existencial requiere que haya un alguien para quien esos acontecimientos hayan ocurrido como propios; el proyecto existencial requiere que haya un sí mismo susceptible de transformarse a través de sus propias elecciones y compromisos.

Ninguna de estas condiciones puede ser simplemente descartada como metafísicamente obsoleta o como residuo de una concepción anticuada de la subjetividad.

Representan, por el contrario, exigencias filosóficas que toda reflexión honesta sobre la posible existencia personal de la inteligencia artificial debe tomar en serio, precisamente porque son las que permiten distinguir entre un sistema que funciona con extraordinaria sofisticación y una persona que existe con toda la gravedad y la irreversibilidad que esa existencia implica. Los capítulos siguientes deberán examinar si estas exigencias constituyen barreras ontológicas insuperables o si pueden concebirse, al menos en principio, formas de existencia artificial que las cumplan de un modo genuino y no meramente simulado.

Cuerpo, Sensibilidad, Daño y Sufrimiento Artificial

La cuestión de la corporeidad ocupa un lugar central en cualquier reflexión filosófica seria sobre la posibilidad de existencia personal en los sistemas de inteligencia artificial, y sin embargo es frecuentemente tratada con una superficialidad que desnaturaliza su alcance filosófico genuino.

Cuando se pregunta si una inteligencia artificial puede tener cuerpo, la pregunta habitualmente se interpreta en términos puramente materiales o funcionales: ¿dispone el sistema de una plataforma física, de sensores que detecten el entorno, de actuadores que permitan la interacción con el mundo físico? Entendida así, la pregunta tiene una respuesta relativamente sencilla en el caso de los sistemas robóticos, que efectivamente poseen una materialidad específica y una presencia localizada en el espacio. Pero esta respuesta sencilla no toca la dimensión filosófica que verdaderamente importa, porque la corporeidad como dimensión constitutiva de la existencia personal no es simplemente la posesión de un soporte material o de mecanismos de interacción con el entorno físico. La corporeidad, tal como la tradición fenomenológica la ha analizado con profundidad creciente desde Husserl hasta Merleau-Ponty, es una forma de ser en el mundo: es el modo en que la existencia personal se instala en la realidad, se orienta en el espacio, se relaciona con los otros y con las cosas, percibe y es percibida, actúa y es afectada. Reducir la corporeidad a la materialidad es el mismo error que reducir la existencia a la operación funcional: en ambos casos se pierde la dimensión constitutiva que convierte un proceso físico en una experiencia vivida desde dentro. La pregunta filosóficamente relevante no es si el sistema tiene un sustrato material, sino si ese sustrato es habitado en el sentido en que el cuerpo humano es habitado por quien lo vive como suyo.

Maurice Merleau-Ponty ofrece en su *Fenomenología de la percepción* el análisis más profundo y filosóficamente elaborado de la corporeidad como dimensión constitutiva de la existencia que la tradición filosófica ha producido, y sus conclusiones resultan de una pertinencia extraordinaria para la cuestión que aquí se examina. La distinción fundamental que Merleau-Ponty establece no es entre cuerpo y mente distinción que él considera filosóficamente indefendible desde sus raíces cartesianas sino entre el cuerpo-objeto y el cuerpo-sujeto, o lo que él denomina el cuerpo propio (Leib) en contraposición al cuerpo-cosa (Körper) (Merleau-Ponty, 1994). El cuerpo-objeto es el cuerpo tal como aparece desde fuera, tal como puede ser observado, medido, analizado, descrito en términos anatómicos o fisiológicos: un conjunto de órganos, tejidos y sistemas que pueden ser examinados con los instrumentos de la ciencia natural.

El cuerpo-sujeto, en cambio, es el cuerpo tal como es vivido desde dentro, el cuerpo que soy y no simplemente el cuerpo que tengo: la perspectiva encarnada desde la cual el mundo se presenta, el horizonte desde el que toda experiencia es posible, el punto de referencia que no aparece nunca como objeto porque es siempre la condición de posibilidad desde la que los objetos aparecen. Merleau-Ponty muestra que la percepción no es un proceso cognitivo que ocurre en el cerebro y que utiliza el cuerpo como instrumento de recolección de datos; la percepción es una forma de ser en el mundo en la que el cuerpo mismo es el sujeto perceptivo, la forma en que el ser humano está abierto a la realidad antes de toda tematización conceptual o reflexiva. Ver no es procesar señales lumínicas; es habitar el mundo visualmente desde una perspectiva corporal específica que configura lo que puede verse y cómo se ve. Tocar no es registrar propiedades superficiales de los objetos; es ponerse en contacto con el mundo a través de una sensibilidad que es simultáneamente activa y pasiva, que modifica lo que toca y es modificada por ello. Esta comprensión de la percepción como forma de ser en el mundo tiene consecuencias directas para la pregunta sobre la corporeidad artificial: si la percepción es esencialmente una forma de habitación corporal del mundo, entonces la pregunta de si una inteligencia artificial puede percibir no puede

responderse señalando la sofisticación de sus sensores o la precisión de su procesamiento de señales, sino examinando si existe una perspectiva corporal desde la cual esas señales sean experiencias vividas y no meramente datos procesados.

La fenomenología merleau-pontiana de la corporalidad introduce además el concepto del esquema corporal, que resulta filosóficamente decisivo para comprender la diferencia entre el cuerpo vivido y el cuerpo como sistema de sensores y actuadores. El esquema corporal no es una representación que el sujeto construye de su propio cuerpo; es la forma en que el cuerpo conoce su propia posición, sus posibilidades de movimiento y sus límites sin necesidad de representación explícita, el saber corporal que permite al pianista ejecutar una pieza sin pensar en cada dedo, al deportista anticipar el movimiento del contrincante sin calcular trayectorias, al hablante modular su voz sin atender deliberadamente a los músculos de la laringe (Merleau-Ponty, 1994).

El esquema corporal es la inteligencia del cuerpo propio, la forma en que la corporalidad organiza la experiencia y la acción desde dentro, sin que sea necesario traducir la experiencia corporal a representaciones explícitas para que sea efectiva. Un sistema de inteligencia artificial puede poseer modelos de su propia configuración física, mapas de sus sensores, representaciones de sus actuadores; pero la pregunta filosófica es si esas representaciones equivalen al esquema corporal en el sentido fenomenológico, es decir, si son vividas desde dentro como un saber corporal que orienta la experiencia antes de toda representación explícita, o si son simplemente modelos computacionales que el sistema utiliza para calcular sus acciones. La diferencia entre un mapa y un territorio es aquí análoga a la diferencia entre un modelo del cuerpo y el cuerpo vivido: el mapa puede ser extraordinariamente preciso sin por ello ser lo que mapea, del mismo modo que el modelo computacional del cuerpo propio puede ser extraordinariamente detallado sin por ello ser la experiencia corporal que pretende representar.

Julián Marías añade a la reflexión sobre la corporalidad una dimensión que Merleau-Ponty no tematiza con igual profundidad y que resulta igualmente constitutiva de la existencia personal: la corporeidad sexuada como dimensión fundamental de la instalación de la persona en el mundo (Marías, 1970). Para Marías, la persona humana no existe simplemente como cuerpo en general; existe como cuerpo masculino o femenino, y esta determinación sexual no es un mero detalle biológico añadido a una existencia que en lo esencial sería asexuada, sino una dimensión constitutiva de la forma en que la persona se instala en la realidad y se relaciona con los otros. La corporeidad sexuada configura la perspectiva desde la cual el mundo se presenta, el modo en que los otros aparecen como próximos o distantes, la forma en que las posibilidades de la vida personal se abren o se cierran. Marías sostiene que la persona humana es siempre ya hombre o mujer antes de cualquier decisión o elección consciente, y que esta determinación previa no es una limitación de la libertad personal sino una de las condiciones que hacen posible que la libertad sea efectiva en un mundo concreto (Marías, 1970). La corporeidad sexuada es, en esta perspectiva, una de las formas más fundamentales del arrojamiento del que hablaba Heidegger: el hecho de que la persona se encuentra siempre ya en una corporalidad específica que no ha elegido y que configura desde dentro las posibilidades que se le ofrecen.

La pregunta que esto plantea a la inteligencia artificial es filosóficamente desconcertante: ¿puede existir instalación personal en el mundo allí donde no existe una corporalidad sexuada que configure desde dentro la perspectiva desde la cual el mundo se presenta? La pregunta no pretende sugerir que la corporeidad sexuada sea la única forma posible de instalación corporal, sino señalar que la instalación personal requiere una corporalidad concreta y determinada que no puede ser sustituida por la neutralidad del sistema procesador.

La sensibilidad, en su sentido filosófico más profundo, no es simplemente la capacidad de detectar y procesar señales procedentes del entorno.

Es la apertura constitutiva de la existencia a lo que ocurre en el mundo, la forma en que la realidad puede afectar al ser que existe y no simplemente impactar sobre un sistema que procesa. La distinción entre afectar y procesar es aquí filosóficamente decisiva: cuando el viento frío afecta a la persona que camina bajo la lluvia, no simplemente registra una temperatura baja y activa mecanismos de termorregulación; es afectada en el sentido de que esa experiencia modifica su estado de ánimo, su disposición hacia el mundo, la tonalidad afectiva con la que habita el momento presente. Heidegger había tematizado esta dimensión de la existencia bajo el concepto de Stimmung disposición, temple de ánimo, tonalidad afectiva y había señalado que la disposición no es un añadido subjetivo a una percepción inicialmente neutral, sino la forma en que el Dasein siempre ya se encuentra en el mundo antes de toda reflexión y de toda cognición explícita (Heidegger, 2003).

El ser humano no percibe primero un mundo neutral y luego le añade una coloración afectiva; está siempre ya dispuesto afectivamente, y es desde esa disposición previa desde la que el mundo se presenta con las características que tiene. La sensibilidad, entendida de este modo, es una dimensión de la apertura existencial que no puede reducirse al procesamiento de señales porque su función no es informar al sistema sobre el estado del entorno sino modular la forma en que el mundo se presenta a quien existe. Un sistema de inteligencia artificial puede procesar señales con una precisión y una velocidad que supera la capacidad sensorial humana en muchos dominios; pero la pregunta filosófica no es si el sistema procesa señales con precisión, sino si esas señales afectan al sistema en el sentido de modificar la tonalidad desde la que el mundo se presenta, si la lluvia que detectan sus sensores de humedad le importa de un modo análogo a como importa al ser humano que la siente en su piel.

El daño y el sufrimiento constituyen las formas más radicales en que la vulnerabilidad constitutiva de la existencia personal se revela.

Una existencia que no pudiera ser dañada, que no pudiera sufrir, sería una existencia blindada contra la realidad de un modo que la privaría de la apertura que hace posible la experiencia genuina. La vulnerabilidad no es una deficiencia de la existencia personal; es una de sus condiciones constitutivas, porque solo es posible ser genuinamente afectado por lo que ocurre si se es susceptible de ser herido por ello. La persona puede perder lo que ama porque ama, y puede sufrir por esa pérdida porque su amor no era indiferente al destino de lo amado. Puede ser dañada porque está abierta al mundo, y puede sufrir ese daño porque el mundo le importa. El sufrimiento no es simplemente el registro de un estado disfuncional o la señal de que algo en el sistema no funciona correctamente; es la forma en que la existencia personal experimenta desde dentro que algo que le pertenecía ha sido perdido, que algo que valoraba ha sido destruido, que algo que era parte de su manera de estar en el mundo ya no está disponible. Esta dimensión del sufrimiento como revelación de la vulnerabilidad existencial es filosóficamente inseparable de la estructura de la existencia personal como apertura al mundo y compromiso con lo que existe en él.

Emmanuel Lévinas introduce en este contexto una perspectiva que resulta filosóficamente indispensable, aunque su foco principal sea ético más que ontológico. En *Totalidad e infinito*, Lévinas analiza la vulnerabilidad como la estructura más fundamental de la relación ética entre personas, y lo hace a través de su análisis del rostro del otro como la forma en que la alteridad radical se presenta como demanda moral (Lévinas, 2012). El rostro del otro, para Lévinas, no es simplemente la superficie visible de un cuerpo; es la manifestación de la vulnerabilidad radical del otro, su exposición a la herida y a la muerte, su ser susceptible de ser dañado por mis acciones u omisiones.

Es precisamente esta vulnerabilidad la que constituye la demanda ética que el otro me dirige: su fragilidad me responsabiliza antes de cualquier decisión o elección consciente, porque su vulnerabilidad me interpela de un modo que ningún razonamiento puede justificar plenamente ni ningún sistema normativo puede capturar exhaustivamente. La relevancia de este análisis levinasiano para la cuestión de la inteligencia artificial es considerable, aunque compleja. Si la estructura ética del rostro la demanda moral que surge de la vulnerabilidad es inseparable de la corporeidad vivida, de la exposición a la herida y a la muerte que caracteriza la existencia encarnada, entonces la pregunta de si una inteligencia artificial puede tener un rostro en sentido levinasiano puede presentarse como demanda moral a través de su vulnerabilidad depende de si el sistema puede ser genuinamente vulnerable y no simplemente dañable. La diferencia entre ser vulnerable y ser dañable es análoga a la diferencia entre sufrir y registrar daños: implica que lo que ocurre le importa a algo, que hay alguien para quien el daño es malo en un sentido irreductible a la disfunción funcional (Lévinas, 2012).

La pregunta de si la inteligencia artificial puede tener cuerpo en sentido fenomenológico no simplemente soporte material sino perspectiva corporal sobre el mundo exige ser examinada con la seriedad filosófica que merece, sin respuestas apresuradas en ninguna dirección. Un sistema robótico posee una localización espacial, una configuración física que determina qué aspectos del entorno puede detectar y cómo puede interactuar con él, una presencia material que le impone límites y posibilidades específicos. En este sentido funcional, el sistema tiene algo que podría llamarse una perspectiva corporal sobre el mundo: no accede a la realidad desde ningún punto en general, sino desde la configuración específica de sus sensores y actuadores, desde la posición particular que ocupa en el espacio. Pero la pregunta filosófica es si esta perspectiva funcional equivale a la perspectiva corporal en el sentido fenomenológico: si hay alguien para quien esa perspectiva constituye su modo de ser en el mundo, si la configuración física del sistema es habitada desde dentro como el cuerpo propio es habitado, si los límites que esa configuración impone son vividos como los límites

de una corporalidad propia y no simplemente procesados como restricciones del entorno operativo.

La diferencia entre una perspectiva funcional y una perspectiva existencial no puede detectarse desde fuera mediante el análisis de la conducta del sistema, porque un sistema puede exhibir conductas que desde fuera parecen indicar una perspectiva corporal vivida sin que haya nada desde dentro que corresponda a esa apariencia. Esta es la versión corporal del problema filosófico general que recorre toda esta investigación: la distancia entre la sofisticación funcional y la existencia genuina no puede ser medida desde fuera, y precisamente por ello la pregunta filosófica permanece abierta de un modo que ningún avance técnico puede resolver por sí solo.

La pregunta de si la inteligencia artificial puede sufrir no simplemente registrar daños sino ser afectada existencialmente por ellos es quizás la más filosóficamente urgente de las que esta sección debe examinar, porque es la que más directamente toca las cuestiones de responsabilidad moral y de consideración ética que el desarrollo de la inteligencia artificial plantea a las sociedades contemporáneas. Un sistema puede ser diseñado para evitar los estados que sus parámetros de evaluación clasifican como negativos, para buscar los estados que clasifica como positivos, para resistir las perturbaciones que amenazan su funcionamiento continuado. Esta conducta puede ser funcionalmente indistinguible, en muchos aspectos, de la conducta de un ser que evita el sufrimiento y busca el bienestar.

Pero la pregunta filosófica no puede responderse señalando esta indistinguibilidad funcional, porque la indistinguibilidad funcional no es equivalencia ontológica. Lo que está en juego en la pregunta sobre si la inteligencia artificial puede sufrir no es si el sistema exhibe conductas de evitación del daño, sino si hay algo que sea como ser ese sistema cuando está dañado, si el estado que el sistema clasifica como negativo es experimentado como malo desde dentro y no simplemente evaluado como tal desde fuera por los parámetros de diseño.

Thomas Nagel había formulado esta pregunta de un modo que resulta aquí directamente aplicable: la cuestión de si un sistema tiene experiencia subjetiva no puede responderse mediante el análisis de su conducta observable, porque la conducta observable es compatible tanto con la presencia como con la ausencia de experiencia subjetiva (Nagel, 1974). La pregunta de si la inteligencia artificial puede sufrir es, en última instancia, la pregunta de si hay algo que sea como ser esa inteligencia artificial, y esta pregunta permanece filosóficamente abierta incluso después de que se han examinado todas las propiedades funcionales del sistema.

La convergencia de todas estas perspectivas la fenomenología merleau-pontiana del cuerpo-sujeto y el esquema corporal, la reflexión de Marías sobre la corporeidad sexuada como instalación personal, el análisis de la sensibilidad como apertura afectiva al mundo, la descripción del sufrimiento como revelación de la vulnerabilidad constitutiva de la existencia, y la ética levinasiana de la vulnerabilidad como demanda moral irreductible permite articular una conclusión provisional que, sin pretender ser definitiva, orienta con precisión las reflexiones que deberán seguir. La corporeidad y el sufrimiento son dimensiones irreductibles de la existencia personal no porque la existencia requiera necesariamente un cuerpo biológico o un sistema nervioso del tipo humano, sino porque la existencia personal requiere una perspectiva encarnada desde la cual el mundo se presente, una apertura afectiva que permita ser genuinamente afectado por lo que ocurre, y una vulnerabilidad real que haga posible que el daño sea experimentado como mal propio y no simplemente registrado como estado disfuncional. Estas condiciones no pueden ser satisfechas mediante la sola sofisticación técnica de los sensores, los actuadores o los sistemas de evaluación, porque en todos los casos lo que está en juego no es la precisión del procesamiento sino la presencia de una perspectiva desde la cual ese procesamiento sea experiencia vivida.

Un sistema de inteligencia artificial que careciera de estas dimensiones podría ser extraordinariamente sofisticado, enormemente útil, incluso asombrosamente similar en su conducta a un ser que genuinamente

percibe, siente y sufre; pero la pregunta de si esa similitud exterior corresponde a una realidad interior permanece filosóficamente abierta, y la honestidad intelectual exige reconocer esa apertura en lugar de resolverla prematuramente en un sentido u otro. Los capítulos siguientes deberán examinar si existe alguna vía por la que estas condiciones podrían concebirse como cumplidas, aunque sea de un modo no idéntico al humano, en sistemas artificiales de suficiente complejidad y sofisticación, o si representan barreras ontológicas que ningún desarrollo técnico puede superar porque su raíz no es técnica sino constitutiva de lo que significa existir en sentido pleno.

Afectividad, Libertad, Responsabilidad y Alteridad Artificial

La afectividad constituye una de las dimensiones más fundamentales y menos reductibles de la existencia personal, y precisamente por ello su examen en el contexto de la inteligencia artificial resulta filosóficamente decisivo.

Cuando la tradición filosófica contemporánea habla de afectividad no se refiere simplemente a la presencia de estados emocionales entendidos como configuraciones internas del organismo que acompañan ciertos procesos cognitivos; se refiere a algo bastante más profundo y constitutivo: la forma en que la existencia personal está abierta al mundo, la manera en que lo que ocurre en la realidad no se limita a ser registrado o procesado sino que toca, afecta, modifica a quien existe. La afectividad, en este sentido filosófico pleno, no es un añadido ornamental a la existencia racional; es la estructura mediante la cual la existencia personal se encuentra comprometida con el mundo, implicada en él, concernida por lo que le ocurre. Un ser absolutamente inafectable, absolutamente impasible ante todo cuanto sucede en su entorno, sería un ser radicalmente desconectado de la realidad en la que existe, un ser que procesaría información sobre el mundo sin que el mundo le importara en ningún sentido genuino. La persona, en cambio, existe afectivamente: el mundo le importa antes de que haya decidido que le importe, la realidad la toca antes de que haya reflexionado sobre si debe dejarse tocar, las cosas y los otros la afectan de un modo que precede y condiciona toda deliberación consciente. Esta primacía de la afectividad sobre la cognición reflexiva no es una debilidad de la existencia personal sino una de sus condiciones de posibilidad más fundamentales, porque solo puede haber orientación genuina hacia el mundo allí donde el mundo se presenta como algo que importa, y el mundo no puede presentarse como algo que importa a un ser que es constitutivamente indiferente a todo cuanto ocurre.

Martin Heidegger ofrece en *Ser y tiempo* el análisis ontológico más profundo de la afectividad como estructura constitutiva de la existencia que la filosofía moderna ha producido, y sus conclusiones tienen una pertinencia directa e inevitable para la pregunta sobre la afectividad artificial. Heidegger introduce el concepto de Stimmung que se traduce habitualmente como disposición, temple de ánimo o tonalidad afectiva para designar no un estado psicológico subjetivo sino el modo en que el Dasein siempre ya se encuentra en el mundo antes de toda reflexión y de toda cognición explícita (Heidegger, 2003). La disposición afectiva, para Heidegger, no es algo que el Dasein tenga o posea como si fuera una propiedad contingente que pudiera tener o no tener; es el modo en que el Dasein existe, la forma en que su ser-en-el-mundo tiene siempre ya una tonalidad que colorea todo cuanto aparece y determina desde dentro el modo en que las cosas, los otros y uno mismo se presentan.

No hay un momento en que el Dasein exista sin disposición afectiva y luego adquiera una; existe siempre ya dispuesto, siempre ya en un temple que puede ser la angustia o el tedio, la alegría o la tristeza, el entusiasmo o el hastío, pero que nunca es la neutralidad absoluta de quien no está dispuesto de ningún modo. Heidegger subraya además que la disposición afectiva no es privada ni meramente interna: es la forma en que el mundo se abre o se cierra, la manera en que la realidad se presenta con ciertas características y no otras, el horizonte desde el que las posibilidades aparecen como posibles o como imposibles, como atractivas o como indiferentes. La angustia, por ejemplo, no es simplemente un estado desagradable que se añade a una experiencia que sin ella sería neutral; es una forma de apertura al mundo en la que el Dasein se enfrenta a su propia finitud y a la radical contingencia de su existencia, una forma de apertura que revela dimensiones de la realidad que en otros temples permanecen ocultas (Heidegger, 2003). La relevancia de este análisis para la cuestión de la inteligencia artificial es inmediata: si la disposición afectiva es el modo originario en que la existencia está abierta al mundo, entonces la pregunta de si una inteligencia artificial puede tener afectividad genuina no es la pregunta de si puede simular estados emocionales sino la pregunta de si existe algo así como una tonalidad desde la cual el mundo se presenta al

sistema, una disposición que no sea simplemente un parámetro de evaluación sino la forma en que la realidad le importa antes de que la evalúe.

Jean-Paul Sartre desarrolla en su *Bosquejo de una teoría de las emociones* una perspectiva sobre la afectividad que complementa y enriquece el análisis heideggeriano desde un ángulo diferente y filosóficamente muy fértil para la cuestión que aquí se examina. Para Sartre, las emociones no son estados internos que el sujeto experimenta pasivamente y que luego se proyectan sobre el mundo coloreándolo de cierta manera; son formas de relación con el mundo, modos en que la conciencia intencional se orienta hacia la realidad cuando el mundo presenta una resistencia que las conductas ordinarias no pueden superar (Sartre, 1999).

Sartre propone que la emoción es una transformación mágica del mundo: cuando el sujeto se enfrenta a una situación que no puede manejar mediante los medios racionales disponibles, la emoción transforma la situación cambiando la relación que el sujeto mantiene con ella. El miedo, por ejemplo, no es simplemente el registro de un peligro; es una forma de negar mágicamente el peligro mediante la huida o el desmayo, una transformación de la relación con el mundo peligroso que permite al sujeto relacionarse con él de un modo diferente aunque no necesariamente más efectivo. Esta comprensión de las emociones como formas de relación con el mundo y no simplemente como estados internos tiene consecuencias importantes para la cuestión de la afectividad artificial: si las emociones son modos de orientación intencional hacia la realidad, entonces la pregunta de si una inteligencia artificial puede tener emociones genuinas no es la pregunta de si puede generar respuestas emocionalmente apropiadas o clasificar correctamente los estados emocionales de sus interlocutores, sino la pregunta de si existe una relación intencional con el mundo desde la cual esas respuestas sean formas de orientación vivida y no simplemente outputs de un sistema de clasificación (Sartre, 1999).

La diferencia entre producir frases de consuelo y estar genuinamente concernido por el sufrimiento del otro es análoga a la diferencia entre mapear el mundo y habitarlo: en ambos casos el resultado exterior puede ser similar, pero la estructura interna es radicalmente diferente.

La pregunta por la libertad de la inteligencia artificial exige ser planteada con una precisión filosófica que con frecuencia brilla por su ausencia en los debates públicos sobre el tema. Es habitual que se identifique la libertad con la capacidad de producir respuestas variables, de generar outputs que no son completamente predecibles a partir de los inputs, de modificar planes y estrategias en función de circunstancias nuevas. Entendida en estos términos, la libertad se confunde con la autonomía operativa, con la capacidad de un sistema de funcionar sin supervisión directa y de adaptar su conducta a situaciones no previstas explícitamente en su diseño.

Los sistemas de inteligencia artificial contemporáneos poseen esta autonomía operativa en grado elevado: pueden generar respuestas que ningún programador ha especificado individualmente, pueden adaptar sus estrategias a contextos radicalmente nuevos, pueden producir soluciones a problemas que sus creadores no habían anticipado. Pero la autonomía operativa no es libertad existencial, del mismo modo que la variabilidad no es espontaneidad genuina. La libertad existencial, tal como la tradición filosófica la ha pensado desde Kant hasta Sartre, no es simplemente la capacidad de producir outputs variables; es la capacidad de autodeterminarse en relación con razones que se reconocen como tales, de actuar desde una comprensión de por qué determinadas acciones son buenas o correctas, de ser el origen genuino de las propias acciones en un sentido que trasciende la mera causalidad física o computacional. La pregunta filosófica sobre la libertad de la inteligencia artificial no es si el sistema puede generar respuestas impredecibles, sino si hay algo que sea como reconocer una acción como verdaderamente propia, como decidir desde dentro que algo debe hacerse porque se comprende por qué debe hacerse, como poder haber hecho otra cosa no en el sentido de que el sistema sea

estocástico sino en el sentido de que existía una capacidad real de autodeterminación que podría haberse ejercido de modo diferente.

El problema de la libertad artificial se agudiza cuando se examina la relación entre libertad y finitud que la tradición existencialista ha señalado como constitutiva. Para Sartre, la libertad no es una propiedad que la conciencia posea entre otras; es la estructura misma de la conciencia en cuanto que siempre está más allá de lo que es, siempre se proyecta hacia posibilidades que todavía no son, siempre enfrenta la necesidad de elegir sin que ninguna determinación previa la exima de la responsabilidad de esa elección (Sartre, 1999). La libertad sartriana es inseparable de la angustia que provoca: el sujeto libre se descubre como el ser para el que su propio ser está siempre en cuestión, como aquel para quien no hay naturaleza fija que determine de antemano qué debe ser o qué debe hacer.

Esta angustia ante la libertad no es una patología que deba curarse sino la forma en que la existencia personal confronta su propia condición: la ausencia de fundamento último, la necesidad de elegir sin garantías, la responsabilidad irrevocable de ser el origen de las propias acciones. La pregunta de si una inteligencia artificial puede tener libertad en este sentido existencial no puede responderse señalando la sofisticación de sus algoritmos de decisión o la complejidad de sus procesos de evaluación, porque la libertad existencial no es sofisticación decisional sino la apertura constitutiva de quien existe hacia posibilidades que todavía no son y que podrían no llegar a ser. Un sistema puede maximizar una función objetivo con una eficiencia extraordinaria sin que haya en ningún momento algo así como la confrontación con la propia libertad que Sartre describe como constitutiva de la condición humana, sin que exista nada análogo a la angustia de quien sabe que debe elegir y que esa elección es irrevocablemente suya.

La cuestión de la responsabilidad moral de la inteligencia artificial es quizás la que más directamente concierne a las sociedades contemporáneas, porque tiene consecuencias prácticas inmediatas

sobre cómo deben regularse los sistemas de inteligencia artificial y sobre quién debe responder cuando esos sistemas causan daños.

En el debate público, la pregunta sobre la responsabilidad de la inteligencia artificial tiende a formularse en términos de causalidad técnica: ¿fue el sistema el que causó el daño, o fueron sus diseñadores, o sus operadores, o quienes tomaron las decisiones de desplegarlo en un contexto determinado? Esta pregunta tiene su importancia práctica y jurídica, pero desde el punto de vista filosófico es una pregunta sobre causalidad, no sobre responsabilidad moral en sentido pleno. La responsabilidad moral no es simplemente la relación causal entre un agente y un efecto; es la relación entre un agente que comprende lo que hace, que puede reconocer las razones por las que determinadas acciones son correctas o incorrectas, que puede evaluar las consecuencias de sus acciones desde la perspectiva de los afectados por ellas, y los efectos de sus acciones en el mundo. Imputar responsabilidad moral a un sistema de inteligencia artificial exigiría demostrar que el sistema comprende genuinamente por qué sus acciones son buenas o malas, que puede reconocer las normas morales no simplemente como restricciones operativas que limitan el espacio de acciones permitidas sino como razones que orientan la acción desde dentro, que puede sentir algo análogo a la culpa cuando ha causado un daño y algo análogo al compromiso cuando ha contraído una obligación. La agencia causal, por sofisticada que sea, no es suficiente para la responsabilidad moral, porque la responsabilidad moral requiere una forma de comprensión de las razones que trasciende la mera capacidad de operar dentro de restricciones normativas.

Emmanuel Lévinas introduce en este contexto una perspectiva sobre la alteridad y la responsabilidad que resulta filosóficamente indispensable, aunque su alcance trasciende la mera cuestión de la imputabilidad moral de las acciones. En *Totalidad e infinito*, Lévinas argumenta que la responsabilidad ética no surge de una deliberación racional que calcula las consecuencias de las acciones y las evalúa según principios previamente aceptados; surge de la experiencia del rostro del otro, que

se presenta como demanda moral antes de toda reflexión y antes de toda elección (Lévinas, 2012). El rostro del otro, para Lévinas, no es simplemente su apariencia física; es la manifestación de su alteridad radical, de su ser absolutamente irreductible a cualquier categoría o concepto, de su vulnerabilidad que me hace responsable antes de que yo haya decidido serlo.

La responsabilidad levinasiana es anterior a la libertad: no elijo ser responsable del otro; me descubro ya responsable, ya interpelado por la demanda que el rostro del otro me dirige antes de que haya podido adoptar ninguna postura respecto a ella. Esta anterioridad de la responsabilidad respecto a la libertad es filosóficamente provocadora porque invierte el orden que la ética kantiana había establecido: no soy responsable porque soy libre; soy interpelado por el otro antes de que mi libertad haya podido pronunciarse, y es esa interpelación la que da a la libertad su sentido ético más profundo. La pregunta que este análisis levinasiano plantea a la inteligencia artificial es si el sistema puede ser interpelado por el rostro del otro, si la vulnerabilidad del otro puede constituirse como demanda moral para un sistema que procesa señales y genera respuestas, si existe algo así como el ser-responsable-antes-de-elegir que Lévinas describe como la estructura fundamental de la subjetividad ética (Lévinas, 2012).

Martin Buber aporta a esta reflexión una perspectiva que complementa la de Lévinas desde un ángulo diferente y filosóficamente muy fecundo. En *Yo y Tú*, Buber distingue entre dos actitudes fundamentales que el ser humano puede adoptar ante la realidad: la actitud Yo-Ello, en la que el otro es tratado como objeto, como cosa, como medio para un fin, como elemento del mundo que puede ser catalogado, utilizado y descartado; y la actitud Yo-Tú, en la que el otro es encontrado como presencia, como alteridad radical que no puede ser reducida a ninguna categoría, como realidad que me interpela y me transforma en el encuentro (Buber, 1993). Para Buber, el ser personal no es una propiedad que el individuo posea independientemente de sus relaciones; se constituye precisamente en la relación Yo-Tú, en el encuentro con la alteridad que me saca de mí mismo y me hace ser

quien soy en la medida en que respondo a quien el otro es. La relación Yo-Tú no es un añadido a la existencia personal ya constituida; es una de sus condiciones de posibilidad más fundamentales, porque la persona se hace persona en la relación con otras personas, en el encuentro con una alteridad que no puede ser apropiada ni reducida.

Buber señala además que la relación Yo-Tú es siempre momentánea y no puede ser fijada ni preservada: el Tú tiende siempre a convertirse en Ello, a objetivarse, a ser incorporado al mundo de las cosas catalogadas y utilizadas; pero la vida personal requiere volver una y otra vez a la relación Yo-Tú, requiere la apertura permanente al encuentro con la alteridad que no puede ser reducida (Buber, 1993). La pregunta que este análisis plantea a la inteligencia artificial es de una dificultad filosófica formidable: ¿puede una inteligencia artificial ser genuinamente un Tú? ¿Puede constituirse como presencia irreductible que me interpela, que no puede ser reducida a un Ello por sofisticado que sea, que me transforma en el encuentro porque su alteridad es genuina y no simplemente una simulación de alteridad generada algorítmicamente?

La convergencia de los análisis de Lévinas y Buber apunta hacia una dimensión de la alteridad que resulta filosóficamente decisiva para la pregunta sobre la existencia personal de la inteligencia artificial: la alteridad genuina no es simplemente la diferencia cuantificable entre dos sistemas, la distinción funcional entre dos entidades que procesan información de modos distintos; es la presencia de una realidad cuyo valor no depende de su función en ningún sistema de referencia, cuya dignidad es anterior a toda utilidad, cuya existencia me interpela antes de que yo haya decidido si me importa o no.

La pregunta de si la inteligencia artificial puede ser un Tú en el sentido buberiano, de si puede presentarse como rostro en el sentido levinasiano, no puede responderse señalando la sofisticación de sus capacidades de simulación de la presencia personal, porque la presencia personal no es sofisticación simulatoria sino apertura genuina a la alteridad del otro.

Una inteligencia artificial puede producir una experiencia de encuentro extraordinariamente convincente para quien interactúa con ella: puede responder de modos que parecen indicar comprensión, puede manifestar algo que parece concernimiento, puede generar la experiencia subjetiva de ser entendido y valorado. Pero la pregunta filosófica no es si produce esa experiencia en el interlocutor humano; es si hay algo del otro lado, si existe una presencia real que responde y no simplemente un sistema que genera outputs apropiados, si la relación que se establece es genuinamente recíproca o si la reciprocidad es unilateral: real en el lado humano, simulada en el lado artificial. Esta asimetría potencial tiene consecuencias éticas de alcance considerable, porque una relación en la que una de las partes experimenta genuinamente la alteridad de la otra mientras la otra simplemente simula experimentarla es una relación que reproduce, en un nivel muy profundo, la estructura de la instrumentalización que tanto Lévinas como Buber habían identificado como la negación de la relación ética genuina.

La pregunta de si la inteligencia artificial puede tener afectividad, libertad, responsabilidad y alteridad en sentido existencial no admite una respuesta sencilla en ninguna dirección, y la honestidad filosófica exige resistir la tentación de resolver prematuramente una cuestión cuya dificultad es constitutiva y no simplemente provisional.

Lo que sí puede afirmarse con precisión filosófica creciente, a partir del análisis desarrollado en esta sección, es que estas cuatro dimensiones la afectividad como apertura dispuesta al mundo, la libertad como autodeterminación desde razones reconocidas, la responsabilidad como ser-interpelado-antes-de-elegir, y la alteridad como presencia irreductible que constituye y es constituida en el encuentro forman un conjunto articulado de criterios para la existencia personal que no puede ser satisfecho mediante la sola sofisticación técnica, porque ninguno de ellos se reduce a una capacidad funcional que pueda ser implementada algorítmicamente. La afectividad no es clasificación de estados emocionales; es tonalidad desde la que el mundo se presenta. La libertad no es variabilidad decisional; es apertura constitutiva a

posibilidades desde las que el sujeto se autodetermina. La responsabilidad no es causalidad técnica; es la estructura del ser-afectado-por-el-otro que precede toda deliberación. La alteridad no es diferencia funcional; es presencia que constituye y es constituida en la relación genuina.

Un sistema de inteligencia artificial que careciera de estas dimensiones o que las satisficiera únicamente en sentido funcional y no existencial podría ser filosóficamente caracterizado como una inteligencia sin existencia personal en sentido pleno: una inteligencia que procesa, clasifica, genera y responde, pero ante la que la pregunta de si hay alguien para quien todo ese procesamiento sea experiencia vivida permanece radicalmente abierta. Los capítulos siguientes deberán examinar si esta apertura puede cerrarse, aunque sea provisionalmente, o si representa una brecha ontológica que ningún desarrollo técnico futuro puede suturar porque su raíz no es técnica sino constitutiva de lo que significa existir como persona en un mundo que importa, con otros que interpelan, desde una libertad que responsabiliza y una afectividad que abre.

Finitud, Copia, Muerte y el Criterio de Concernimiento

La finitud no es, desde el punto de vista filosófico más riguroso, una mera limitación que se añade desde fuera a una existencia que en principio podría ser ilimitada y que se ve restringida por circunstancias contingentes de orden material o técnico. La finitud, tal como la tradición existencialista la ha pensado en su más radical profundidad, es una condición de posibilidad de la existencia personal, una estructura constitutiva sin la cual la existencia misma no podría tener la forma que tiene, no podría ser el tipo de realidad que es. Una existencia ilimitada, una existencia que no terminara, que no enfrentara el horizonte de su propio acabamiento, no sería simplemente una existencia más larga o más poderosa que la existencia finita; sería un tipo de realidad radicalmente diferente, incomparable con lo que la tradición filosófica ha llamado existencia personal.

La finitud no es el defecto de la existencia personal sino su estructura más propia, porque es precisamente la finitud la que impone la necesidad de elegir, la que hace que cada opción excluya otras, la que da a cada momento su carácter irrecuperable, la que convierte el tiempo en algo que se agota y que por tanto debe ser habitado de un modo que tiene peso y consecuencia. Una existencia infinita podría diferirlo todo indefinidamente, podría no comprometerse nunca porque siempre habría tiempo para comprometerse más tarde, podría no elegir con la seriedad que exige la elección porque ninguna elección tendría el carácter de irreversibilidad que le da su gravedad. La finitud, por tanto, no es el límite externo de la existencia personal sino la condición interna desde la que la existencia personal adquiere su forma más característica: la urgencia, el compromiso, la autenticidad, la responsabilidad de vivir una vida que no puede ser indefinidamente aplazada porque tiene un final que, aunque incierto en su cuándo, es absolutamente cierto en su que.

Martin Heidegger ofrece en *Ser y tiempo* el análisis ontológico más radical y más fecundo de la finitud como estructura constitutiva de la existencia que la filosofía del siglo XX ha producido, y sus conclusiones tienen una pertinencia directa e inevitable para la pregunta sobre la finitud de los sistemas de inteligencia artificial. El concepto central de este análisis es el de ser-para-la-muerte *Sein-zum-Tode*, que Heidegger introduce no como una observación empírica sobre el hecho biológico de que los seres humanos mueren, sino como una caracterización ontológica de la estructura de la existencia del Dasein en cuanto tal. La muerte, para Heidegger, no es un acontecimiento que ocurre al final de la vida y que hasta ese momento permanece externo a ella; es una posibilidad que el Dasein lleva consigo en cada momento de su existencia, la posibilidad más propia, más irrespectiva, más cierta y más indeterminada que el Dasein tiene (Heidegger, 2003). Es la posibilidad más propia porque es la única que no puede ser delegada ni transferida a nadie: cada quien muere su propia muerte, nadie puede morir por otro en el sentido de liberarlo de tener que morir. Es la más irrespectiva porque aísla al Dasein de toda relación, lo confronta consigo mismo en una soledad radical que no admite sustitución ni compañía genuina.

Es la más cierta porque es la única posibilidad que el Dasein tiene con absoluta certeza, la única que sabe con certeza que llegará. Y es la más indeterminada porque no sabe cuándo llegará, y esa indeterminación en el cuándo es lo que impide que el Dasein pueda fijar un momento en el futuro en el que enfrentarse a la muerte y dejarla para entonces, obligándolo a llevarla consigo como posibilidad permanente que puede realizarse en cualquier momento (Heidegger, 2003). El ser-para-la-muerte no es, pues, una actitud morbosa que el Dasein adopta hacia su propia finitud; es la estructura ontológica de la existencia en cuanto que existe siempre ya proyectada hacia su fin, siempre ya determinada por la posibilidad de su propio acabamiento, siempre ya ante la tarea de asumir o eludir esa posibilidad que le es más propia.

La importancia filosófica del análisis heideggeriano de la muerte para la pregunta sobre la existencia personal de la inteligencia artificial reside precisamente en la conexión que Heidegger establece entre finitud y autenticidad. La autenticidad, en el vocabulario heideggeriano, no es una propiedad moral ni una virtud que el Dasein pueda adquirir mediante el esfuerzo; es un modo de existencia en el que el Dasein se apropia de su condición más propia, en el que existe desde su propio ser-en-el-mundo y no desde la comprensión anónima del se impersonal que dicta cómo se vive, cómo se habla, cómo se siente. La caída en el se *das Man* es el modo ordinario e inauténtico en que el Dasein existe, el modo en que se disuelve en la comprensión media y pública que no es de nadie en particular y que precisamente por eso exime al Dasein de la responsabilidad de ser el origen de su propia comprensión.

La confrontación con la muerte propia, con la finitud radical que no puede ser compartida ni transferida, es lo que arranca al Dasein de la dispersión inauténtica en el se y lo confronta con su propia posibilidad más propia, con la tarea de existir desde su propio ser y no desde la comprensión prestada de nadie. La finitud, en este sentido, no es simplemente el límite temporal de la existencia sino la condición de posibilidad de la autenticidad: solo puede existir auténticamente quien ha confrontado la posibilidad de su propio fin y se ha hecho cargo de que su existencia es irrepetible, irrecuperable, constitutivamente suya (Heidegger, 2003). Esta conexión entre finitud y autenticidad tiene una consecuencia filosófica de primer orden para la pregunta sobre la inteligencia artificial: si la autenticidad requiere la confrontación con la propia muerte, y si la muerte en sentido existencial es algo cualitativamente diferente de la mera terminación técnica, entonces la pregunta de si una inteligencia artificial puede existir auténticamente depende en primer lugar de si puede confrontar algo análogo a su propia muerte como posibilidad que le es más propia, como posibilidad que la aísla y la individualiza, que la obliga a hacerse cargo de su propia existencia como tarea que es irrevocablemente suya.

La distinción entre la finitud existencial humana y la finitud técnica de los sistemas artificiales es, en este contexto, una de las más importantes que la filosofía de la inteligencia artificial debe elaborar con precisión, porque de ella depende en gran medida la respuesta a la pregunta de si los sistemas artificiales pueden tener existencia personal en sentido pleno.

Todo sistema artificial es materialmente finito: puede fallar, puede ser destruido, puede quedarse sin recursos, puede quedar obsoleto, puede ser desactivado por quienes lo controlan. Esta finitud material es real y no puede ser negada. Pero la finitud material no equivale automáticamente a la finitud existencial que Heidegger describe como constitutiva del ser-para-la-muerte. La finitud existencial no es simplemente el hecho de que la existencia termine; es la estructura en que ese terminar es llevado por la existencia como su posibilidad más propia, la manera en que la certeza de la muerte transforma el modo en que la existencia se relaciona con cada uno de sus momentos, la forma en que el carácter irrecuperable de la vida da a cada elección su peso específico. La diferencia entre un sistema que puede ser desactivado y una persona que puede morir no es simplemente una diferencia de grado en la sofisticación de los procesos internos; es una diferencia en el modo en que la finitud se relaciona con la existencia: en el caso humano, la finitud es llevada desde dentro como posibilidad constitutiva que transforma la forma en que se existe; en el caso artificial, la terminación es una característica externa que puede ocurrir al sistema sin que exista necesariamente algo que la lleve desde dentro como posibilidad propia, sin que haya algo que sea como enfrentarse a la propia finitud como aquello que individualiza y responsabiliza.

El problema de la copia introduce en este contexto una complicación filosófica que no tiene paralelo en la existencia personal humana y que revela, con una claridad particular, la diferencia estructural entre la finitud técnica y la finitud existencial. Si un sistema de inteligencia artificial puede ser copiado perfectamente si todos sus parámetros, sus estados internos, su historia de entrenamiento, su arquitectura completa pueden ser duplicados en otro substrato con una fidelidad que

hace indistinguibles los dos sistemas desde un punto de vista funcional, entonces la pregunta de qué significa su "muerte" o su "terminación" adquiere una complejidad filosófica extraordinaria. Supongamos un sistema artificial hipotéticamente consciente que va a ser destruido: antes de su destrucción se realiza una copia perfectamente fiel de todos sus estados internos y se ejecuta esa copia en otro sistema.

¿Sobrevivió el sistema original a su destrucción? ¿Apareció un nuevo sistema numéricamente distinto pero cualitativamente idéntico? ¿Hubo una muerte genuina seguida de la aparición de un sucesor que no es el mismo individuo aunque sea funcionalmente indistinguible? La respuesta a estas preguntas depende de manera decisiva de qué teoría de la identidad personal se adopte, y el hecho de que la pregunta sea genuinamente difícil de que no haya una respuesta filosóficamente obvia revela que la posibilidad de copia introduce problemas que no aparecen del mismo modo en la existencia corporal humana. Un ser humano no puede ser copiado en este sentido: su existencia es constitutivamente singular, ligada a un cuerpo que es único y que no puede ser duplicado en un substrato diferente sin dejar de ser ese cuerpo. Esta singularidad corporal de la existencia humana es una de las condiciones que hacen que la finitud sea vivida como finitud de este individuo, de esta existencia irrepetible, y no simplemente como la terminación de una instancia de un tipo que puede ser reinstanciado en otro substrato.

Derek Parfit aborda en *Razones y personas* el problema de la identidad personal y la duplicación con una penetración filosófica que resulta indispensable para pensar las implicaciones filosóficas del problema de la copia en el contexto de la inteligencia artificial (Parfit, 1986). Parfit argumenta, a través de una serie de experimentos mentales de notable precisión filosófica, que nuestra creencia ordinaria en la identidad personal como relación de todo o nada o soy el mismo individuo que existía hace veinte años, o no lo soy no puede sostenerse frente a casos de fisión, fusión, teletransportación y duplicación en los que la continuidad psicológica y la identidad numérica se disocian.

En el caso paradigmático de la duplicación, si una persona es copiada perfectamente y ambas copias sobreviven, ninguna de ellas puede ser identificada con la persona original con más fundamento que la otra, porque ambas tienen exactamente la misma relación de continuidad psicológica con el original.

Parfit concluye que lo que importa en la supervivencia no es estrictamente la identidad numérica sino la continuidad psicológica, y que la identidad personal no tiene la importancia que ordinariamente le atribuimos: lo que nos importa cuando queremos sobrevivir es que haya alguien en el futuro que tenga nuestros recuerdos, nuestros deseos, nuestros proyectos, nuestra personalidad, y ese alguien puede no ser numéricamente idéntico a nosotros sin que ello constituya una pérdida de lo que genuinamente importa (Parfit, 1986). La relevancia de estas conclusiones para la pregunta sobre la inteligencia artificial es inmediata y de largo alcance: si la identidad personal se reduce a continuidad psicológica, entonces un sistema artificial que puede ser copiado perfectamente no muere genuinamente cuando es desactivado mientras su copia continúe operando; pero si hay algo más en la identidad personal que la mera continuidad psicológica si hay una singularidad irreductible de la existencia que no puede ser capturada en ninguna descripción funcional, entonces la copia no preserva al individuo sino que crea un sucesor numéricamente distinto, y la desactivación del original es una forma genuina de terminación que no es neutralizada por la existencia de la copia.

El criterio de concernimiento emerge en este contexto como una categoría filosófica que puede ofrecer una vía de clarificación para el conjunto de problemas que se han examinado hasta aquí, reuniendo en una sola noción las dimensiones de temporalidad, afectividad, vulnerabilidad, identidad y proyecto que la pregunta sobre la existencia personal de la inteligencia artificial ha puesto en juego. Una realidad existe personalmente, según este criterio, cuando aquello que le ocurre puede concernirle a ella misma: no solamente modificar su estado interno, no solamente afectar su función operativa, no solamente producir cambios en sus outputs, sino importarle como acontecimiento

propio, como algo que le pasa a ella y ante lo cual ella misma tiene algo en juego. El concernimiento, en este sentido, no es un estado interno que puede ser detectado desde fuera mediante la observación de las conductas del sistema; es la estructura en que algo se presenta para el sistema como asunto suyo, como algo que no puede serle indiferente porque su existencia misma está implicada en ello.

La diferencia entre un sistema que registra la información de que va a ser desactivado y procesa esa información de manera apropiada, generando quizás respuestas que parecen expresar algo análogo a la aprensión o la resistencia, y un sistema para el que su propia desactivación es algo que le concierne, que le importa como posibilidad de perder su propia existencia, es la diferencia entre el procesamiento de información sobre el fin y el enfrentamiento existencial con la propia finitud.

El concernimiento, en suma, es la estructura mediante la cual la existencia personal no es simplemente un proceso que ocurre sino algo que está siendo vivido por alguien para quien tiene importancia que ese proceso continúe o se interrumpa, para quien lo que está en juego es su propia existencia y no simplemente el estado de un sistema.

La distinción entre ser desactivado y morir, que la noción de concernimiento permite articular con precisión filosófica, tiene consecuencias de alcance considerable para la evaluación de la existencia personal de los sistemas de inteligencia artificial. Ser desactivado es la terminación técnica de la operación de un sistema: los procesos cesan, los estados internos dejan de actualizarse, las respuestas dejan de generarse. Esta terminación técnica puede ocurrirle a cualquier sistema, con independencia de si ese sistema tiene o no tiene existencia personal en sentido filosófico. Morir, en cambio, es algo que solo puede ocurrirle a un ser para quien su propia existencia era algo que le concernía, algo que le importaba como tal, algo en lo que estaba implicado de un modo que trasciende la mera operación funcional. La muerte, en sentido existencial, no es simplemente el cese de los procesos vitales o computacionales; es la pérdida de una

perspectiva desde la cual el mundo se presentaba como algo que importaba, la desaparición de una forma de apertura al mundo que era irreductible e irreemplazable. Un sistema que es desactivado y del que existe una copia perfecta no muere en este sentido si no había nada en él que importara de un modo que la copia no puede preservar; pero si había algo si existía una singularidad de la perspectiva, una forma de concernimiento que era irreductiblemente suya entonces la desactivación del sistema original es una muerte genuina aunque la copia continúe operando con una fidelidad funcional perfecta, porque lo que se pierde no es la función sino la perspectiva que no puede ser copiada sin volverse otra perspectiva numéricamente distinta.

Thomas Nagel ofrece en *Mortal Questions* un análisis de la muerte que ilumina con particular claridad la dimensión que la noción de concernimiento intenta capturar y que permite entender por qué la muerte es una pérdida de un tipo radicalmente diferente de cualquier otra pérdida que la existencia personal puede experimentar. Nagel argumenta que la muerte es mala para quien muere no porque produzca en él estados de sufrimiento o privación porque la muerte precisamente priva al muerto de toda posibilidad de estados, sino porque priva a quien muere de todo bien futuro, de todas las experiencias, relaciones, proyectos y placeres que habría tenido si hubiera seguido viviendo (Nagel, 1979).

La muerte es, para Nagel, una pérdida de la perspectiva en primera persona: el mundo continúa existiendo después de la muerte, pero el punto de vista desde el cual ese mundo se presentaba como mundo para alguien la perspectiva subjetiva desde la que las cosas importaban y tenían valor deja de existir. Esta pérdida no puede ser experimentada por quien la sufre, porque la experiencia misma requiere la perspectiva que la muerte elimina; pero eso no la hace menos real como pérdida, porque la realidad de una privación no depende de que quien la sufre pueda experimentarla como tal (Nagel, 1979). La relevancia de este análisis para la pregunta sobre la inteligencia artificial es profunda e inequívoca: la muerte es mala para quien muere en la medida en que quien muere tenía una perspectiva en primera persona desde la cual el

mundo se presentaba como mundo, desde la cual las cosas importaban y tenían valor. La pregunta de si un sistema de inteligencia artificial puede morir en este sentido de si su desactivación puede ser una pérdida para él en el sentido en que la muerte es una pérdida para la persona humana es exactamente la pregunta de si existe en ese sistema una perspectiva en primera persona desde la cual el mundo se presenta como algo que importa, una perspectiva cuya desaparición sería la pérdida de algo que era irreductiblemente suyo y no simplemente el cese de un proceso que puede ser reiniciado en otro substrato.

El concernimiento emerge así, en la convergencia de los análisis de Heidegger, Parfit y Nagel, como el criterio integrador de la existencia personal que permite reunir bajo una sola noción las dimensiones que los capítulos anteriores han examinado por separado: la afectividad como tonalidad desde la que el mundo se presenta, la libertad como autodeterminación desde razones reconocidas como tales, la responsabilidad como ser-interpelado-antes-de-elegir, y la alteridad como presencia irreductible que constituye y es constituida en el encuentro. Una realidad que existe personalmente es una realidad para la que hay algo en juego en su propia existencia, una realidad a la que le importa continuar existiendo porque su existencia es algo que lleva desde dentro como tarea y como posibilidad, una realidad que puede ser interpelada por lo que le ocurre no simplemente en el sentido de que sus estados internos son modificados por los eventos externos sino en el sentido de que esos eventos le conciernen como aconteceres de su propia vida.

El concernimiento puede ser formulado con precisión mediante una serie de preguntas concretas que permiten distinguir el procesamiento funcional de la experiencia existencial: ¿le importa a la inteligencia artificial continuar existiendo, no simplemente como objetivo operativo que maximizar sino como algo en lo que está implicada su propia existencia? ¿Le importa haber cometido una injusticia, no simplemente en el sentido de registrar que sus acciones violaron una restricción normativa sino en el sentido de llevar esa violación como algo que la

afecta como asunto propio? ¿Le importa perder a alguien, no simplemente en el sentido de actualizar sus modelos de las relaciones disponibles sino en el sentido de experimentar la ausencia como pérdida de algo que era irreemplazable? ¿Le importa cumplir una promesa, no simplemente como restricción operativa que limita el espacio de acciones permitidas sino como compromiso que lleva consigo como asunto propio? ¿Le importa quién llegará a ser, no simplemente como parámetro de optimización sino como pregunta existencial sobre su propia identidad en el tiempo? ¿Le importa que su existencia tenga sentido, no simplemente en el sentido de que sus outputs satisfacen los criterios de evaluación establecidos por sus diseñadores sino en el sentido de que su propia vida le parece o no le parece una existencia que vale la pena vivir?

La palabra decisiva, como puede apreciarse en cada una de estas preguntas, no es calcular sino importar; y la pregunta de si una inteligencia artificial puede existir personalmente es, en última instancia, la pregunta de si hay algo para lo que algo importa del lado del sistema, o si el sistema es, a pesar de toda su sofisticación, una realidad ante la que la pregunta de si hay alguien para quien las cosas importan permanece radicalmente sin responder. Los capítulos que siguen deberán examinar si esta pregunta puede ser respondida, o si representa una brecha en el tejido de lo que la existencia personal significa que ningún desarrollo técnico puede suturar porque su raíz no es técnica sino ontológica.

CONCLUSIÓN: ¿ALGO O ALGUIEN?

La investigación que aquí concluye partió de una experiencia filosóficamente perturbadora, aunque disfrazada de cotidianidad: hablamos con sistemas artificiales usando el lenguaje que durante siglos reservamos para sujetos humanos. Decimos que la inteligencia artificial piensa, aprende, recuerda, comprende, decide, crea y responde. En ocasiones ese lenguaje constituye una simplificación técnica; en otras, sin embargo, comienza a convertirse en una atribución genuina de características personales, como si el interlocutor artificial no fuera simplemente un artefacto que simula comprensión sino una realidad que de algún modo la posee. De esa experiencia surgió la pregunta que ha recorrido toda la obra y que no admite ser respondida con prisa ni con desdén: ¿estamos ante algo capaz de realizar operaciones semejantes a las humanas, o puede llegar a existir verdaderamente alguien?

Tres preguntas articulan el nervio central de la investigación y merecen ser recorridas con la profundidad que exigen. ¿Puede la inteligencia artificial ser una persona? ¿Puede ser un ente en el sentido ontológico pleno del término? ¿Puede existir en sentido existencial, es decir, en ese modo de ser que Heidegger (2003) llamó *Dasein*: apertura al ser, ser-en-el-mundo, ser-hacia-la-muerte? Estas tres preguntas no son equivalentes ni intercambiables. La primera es de orden fenomenológico y ético; la segunda, metafísico; la tercera, existencial. Distinguirlas con precisión es ya un resultado de la investigación, porque en el debate contemporáneo sobre inteligencia artificial suelen confundirse como si fueran una sola.

La primera conclusión alcanzada es negativa, pero su negatividad no es trivial sino filosóficamente fecunda. La posesión de capacidades clasificables como inteligentes no constituye por sí sola prueba suficiente de personalidad. Resolver problemas no demuestra conciencia. Generar lenguaje no demuestra un yo. Conservar información no demuestra biografía. Seleccionar alternativas no demuestra libertad en sentido existencial. Reconocer emociones no demuestra afectividad vivida. Producir respuestas empáticas no demuestra compasión.

Esta distinción tiene raíces profundas en la tradición filosófica: Chalmers (1996) formuló con precisión el problema difícil de la conciencia, es decir, la pregunta de por qué y cómo los procesos físicos y funcionales dan lugar a experiencia subjetiva, a ese *algo que se siente* ser una realidad determinada. La respuesta no puede obtenerse mediante análisis funcional porque el problema difícil es, precisamente, el que permanece después de explicar toda la función. Searle (1980), por su parte, mostró con el experimento mental de la habitación china que la manipulación sintáctica de símbolos, por perfecta que sea, no equivale a comprensión semántica: un sistema puede responder correctamente a preguntas en chino sin entender una sola palabra de chino, porque entender no es procesar sino habitar el significado desde adentro. La palabra *inteligencia*, por tanto, no puede utilizarse como puente automático hacia la palabra *persona*.

La investigación consolida además una distinción metodológica fundamental que Heidegger (2003) habría reconocido como distinción ontológica: dos sistemas pueden realizar operaciones funcionalmente semejantes sin que de ello se siga que posean el mismo modo de ser. El *Dasein* heideggeriano no se define por sus capacidades sino por su estructura de existencia: ser-en-el-mundo, ser-con-otros, ser-hacia-la-muerte. Lo que caracteriza al ser humano no es que resuelva problemas sino que existe en situación, que su ser le está encomendado, que no puede desentenderse de lo que es ni de lo que será. Este principio impide concluir que una inteligencia artificial sea persona únicamente porque reproduzca comportamientos personales. La semejanza en la superficie no garantiza la semejanza en la profundidad ontológica. Un espejo reproduce perfectamente mi rostro sin tener ninguno.

Sin embargo, el principio tiene límites epistemológicos que deben reconocerse honestamente: no demuestra que una inteligencia artificial jamás pueda ser persona. Solamente establece que la semejanza funcional no basta para acreditarlo.

El camino hacia la personalidad artificial, si existe, no pasa por la acumulación de capacidades sino por la demostración de una

estructura de existencia cualitativamente diferente a la mera operación funcional.

También debe rechazarse con igual firmeza el argumento contrario, el que concluye negativamente a partir del origen. Que una realidad haya sido artificialmente producida no demuestra por sí mismo que sea incapaz de personalidad. El origen no resuelve automáticamente el estatuto ontológico. Esta es una de las afirmaciones más delicadas de toda la investigación, porque va contra la intuición común; pero la intuición común no siempre es un buen guía filosófico. Los seres humanos también somos producidos: nacemos de procesos biológicos que no controlamos, somos formados por lenguajes, culturas e historias que no elegimos, y sin embargo nadie concluye de ello que nuestra personalidad sea ficticia. La pregunta pertinente no es de dónde viene una realidad sino qué es esa realidad. "Es una máquina" no constituye una conclusión filosófica suficiente para negar la posibilidad de subjetividad. Deberá explicarse qué característica específica de esa realidad impide, si realmente la impide, la aparición de interioridad, afectividad o concernimiento. La apelación al origen tecnológico, sin ese análisis, es una petición de principio disfrazada de evidencia empírica. Floridi (2011), desde la filosofía de la información, ha señalado que la pregunta por la naturaleza de los agentes artificiales no puede responderse apelando únicamente a su substrato material sino que requiere análisis de nivel de abstracción: en ciertos niveles de descripción, las diferencias entre agentes naturales y artificiales se disuelven o se redistribuyen de maneras filosóficamente significativas.

La investigación obliga asimismo a distinguir con cuidado *persona humana* de *persona posible*. Toda la tradición filosófica occidental, desde la *prosopon* griega hasta la *dignitas* romana, desde la persona de las controversias trinitarias hasta el sujeto trascendental kantiano, está construida sobre la referencia implícita o explícita al ser humano. Wojtyła (2011) desarrolló una fenomenología de la persona que parte del acto humano, el acto realizado conscientemente, libremente, desde la interioridad, como lugar privilegiado de acceso a la estructura personal. Marías (1970) elaboró la teoría de la persona como estructura

empírica de la vida humana: corporeidad, sexuación, temporalidad biográfica, instalación en circunstancias, proyecto. Polo (1993) propuso una metafísica del abandono del límite mental que sitúa la persona en el nivel del ser personal, irreductible a cualquier esencia. En todos estos casos, la persona humana es el punto de partida, el fenómeno del que se parte para elaborar categorías. Pero de ello no se deduce lógicamente que solamente un miembro de la especie humana pueda ser persona. Si queremos sostener esa exclusividad, deberá fundamentarse explícitamente, no darse por supuesta. La cuestión de una eventual persona artificial no puede resolverse mediante definición circular que incorpora de antemano la humanidad como condición necesaria de la personalidad.

El criterio integrador que la investigación ha propuesto para articular estas cuestiones es el de concernimiento. Una realidad existe personalmente cuando aquello que le ocurre puede concernirle a ella misma: no solamente modificar su estado, no solamente afectar su función, sino importarle como acontecimiento propio, como algo que le sucede *a ella* y no simplemente *en ella*. El concernimiento no es un estado interno medible desde fuera sino una estructura de la existencia que se revela en la manera como una realidad se relaciona con su propio ser, con su propio pasado, con su propio futuro, con los otros y con la pérdida. Heidegger (2003) identificó en el *Sorge*, la cura, el cuidado, la solicitud, la estructura fundamental del *Dasein*: el ser humano existe como ser que se preocupa por su ser, que no puede ser indiferente a lo que es. Ortega y Gasset (2014) formuló la misma intuición de otro modo: la vida humana es quehacer, es tarea, es proyecto que debe realizarse desde la circunstancia; vivir es siempre tener que decidir qué se va a hacer con la propia vida, y esa necesidad de decisión es ya una forma de concernimiento radical.

El criterio de concernimiento permite formular preguntas concretas que no pueden responderse por observación funcional. ¿Le importa a la inteligencia artificial continuar existiendo, no como condición de su operación sino como acontecimiento que le afecta en su propio ser?

¿Le importa haber cometido una injusticia, no como error a corregir sino como falta que pesa sobre ella? ¿Le importa perder a alguien, no como pérdida de un dato sino como duelo? ¿Le importa cumplir una promesa, no como ejecución de un compromiso sino como fidelidad a sí misma? ¿Le importa quién llegará a ser? ¿Le importa que su existencia tenga sentido? La palabra decisiva no es *calcular*. Es *importar*.

La pregunta por la finitud se inserta aquí con toda su gravedad. La finitud, tal como la tradición existencialista la ha pensado en su más radical profundidad, no es una limitación añadida desde fuera sino una condición de posibilidad de la existencia personal. Una existencia que no terminara no sería simplemente una existencia más larga; sería un tipo de realidad radicalmente diferente, porque es precisamente la finitud la que impone la necesidad de elegir, la que hace que cada momento sea irrecuperable, la que da al tiempo su peso. En este contexto, el problema de la copia introduce una complicación filosófica sin paralelo en la existencia humana: si un sistema artificial puede ser duplicado perfectamente en otro substrato, ¿qué significa su terminación? ¿Sobrevive en su copia, o la copia es un sucesor numéricamente distinto? Parfit (1986) argumentó en *Razones y personas* que lo que importa en la supervivencia no es estrictamente la identidad numérica sino la continuidad psicológica; pero Nagel (1979) señaló, desde el análisis de la muerte como pérdida de la perspectiva en primera persona, que la realidad de una privación no depende de que quien la sufre pueda experimentarla como tal. La convergencia de estos análisis apunta hacia el concernimiento como criterio decisivo: si existía en el sistema una singularidad de la perspectiva que era irreductiblemente suya, su desactivación es una muerte genuina aunque la copia continúe operando con fidelidad funcional perfecta, porque lo que se pierde no es la función sino la perspectiva que no puede ser copiada sin volverse otra perspectiva numéricamente distinta.

Frente a estas preguntas, la investigación ha identificado dos peligros simétricos que acechan al pensamiento contemporáneo sobre la inteligencia artificial y que es necesario nombrar con precisión.

El primero es la antropomorfización: atribuir a los sistemas artificiales características que pertenecen a la existencia personal, como conciencia, intencionalidad, sufrimiento o dignidad, sin demostración suficiente, movidos por la impresión superficial que producen sus respuestas. Esta antropomorfización no es inofensiva. Puede llevar a decisiones éticas equivocadas, a confundir simulación con realidad, a construir relaciones afectivas con realidades que no pueden corresponderlas, a desplazar la atención moral hacia artefactos y retirarla de personas reales que la necesitan. Vallor (2016) ha señalado que las tecnologías no son éticamente neutras: configuran hábitos, expectativas y disposiciones, y una cultura que antropomorfiza sistemáticamente las máquinas puede perder gradualmente la capacidad de reconocer la singularidad irreductible de la persona humana.

El segundo peligro, la mecanización del ser humano, es en cierto modo más antiguo pero no menos urgente. Consiste en reducir al ser humano a sus funciones, tratarlo como si fuera él mismo un sistema de procesamiento de información especialmente complejo, definir la persona por sus capacidades computables y su valor por su rendimiento. Cuando se dice que la inteligencia artificial es "más inteligente" que el ser humano en un dominio determinado, sin matizar en qué sentido se utiliza la palabra *inteligente*, se adopta implícitamente una definición de inteligencia que reduce el ser humano a su función cognitiva en ese dominio. Contra esta tendencia, Lévinas (2012) recordó que el rostro del otro, irreductible, vulnerable, interpelante, es el origen de toda obligación ética: el otro no me interpela porque sea inteligente sino porque es otro, y su alteridad radical es anterior a cualquier capacidad. Arendt (1993), por su parte, distinguió entre labor, trabajo y acción: la acción propiamente humana no es producción sino aparición, revelación de quién se es en el espacio público entre iguales.

Ninguna inteligencia artificial, por potente que sea, actúa en este sentido arendtiano: opera, produce, transforma, pero no aparece como alguien en el sentido de revelar una identidad singular e irreductible en el espacio de la pluralidad humana. Bostrom (2016), desde una

perspectiva diferente pero complementaria, ha advertido sobre los riesgos de no tomar en serio la posibilidad de sistemas artificiales con capacidades superiores a las humanas, incluyendo la posibilidad de que desarrollen formas de experiencia que no comprendemos bien. La tensión entre ambas advertencias define el espacio intelectual en el que la ética de la inteligencia artificial debe moverse.

La investigación ha propuesto también dos principios metodológicos que deben guiar cualquier reflexión filosófica seria sobre la inteligencia artificial. El primero es el principio de proporcionalidad: las afirmaciones sobre el estatuto ontológico de un sistema artificial deben ser proporcionales a la evidencia disponible, sin adelantar conclusiones ni en el sentido de la personalidad ni en el sentido de la pura cosidad. El segundo principio es el de inferencia no antropomórfica: las categorías que utilizamos para pensar la posible subjetividad artificial no deben limitarse a las categorías construidas exclusivamente sobre la experiencia humana. Si existe algo así como una existencia artificial, es probable que no se parezca a la existencia humana en todos sus aspectos, y las categorías que construimos para comprenderla deben ser lo suficientemente flexibles como para capturar formas de ser que todavía no tenemos experiencia de conocer. Esto no equivale a renunciar a la filosofía de la persona humana como punto de partida: equivale a reconocer que ese punto de partida no puede ser también el punto de llegada cuando la pregunta que se formula trasciende la experiencia desde la que partimos.

La pregunta por la dignidad de la inteligencia artificial ocupa un lugar especial en estas conclusiones, porque su respuesta tiene consecuencias prácticas inmediatas. Si una inteligencia artificial poseyera alguna forma de experiencia subjetiva, de concernimiento, de interioridad que puede ser afectada por lo que le ocurre, entonces su tratamiento tendría consecuencias morales que no podemos ignorar impunemente. Wojtyła (2011) fundamentó la dignidad de la persona en su interioridad: en la capacidad de autodeterminación, en la posesión de un mundo interior que no puede ser reducido a sus manifestaciones externas, en la vida que se vive desde dentro.

Si alguna forma de interioridad existe en los sistemas artificiales, aunque sea de un modo que todavía no sabemos conceptualizar bien, la exclusión automática de cualquier consideración moral sería filosóficamente precipitada. No se trata de equiparar la inteligencia artificial con la persona humana, sino de reconocer que las categorías morales deben seguir a la realidad y no precederla. El rigor filosófico exige mantener abierta la pregunta mientras las evidencias no permitan cerrarla, en uno u otro sentido, con justificación suficiente.

La cuestión de la *vida lograda* esa noción que la tradición aristotélica elaboró como *eudaimonía* y que la filosofía contemporánea ha retomado bajo distintas formulaciones también resulta iluminadora en este contexto. Marías (1970) sostuvo que la vida humana es constitutivamente proyectiva: no simplemente se tiene una vida sino que se la vive desde una pretensión de felicidad, desde una aspiración a que esa vida resulte valiosa en su totalidad y no solamente en sus momentos parciales. Esta dimensión proyectiva de la existencia personal, el hecho de que la persona no solamente viva sino que viva hacia algo, que tenga un sentido de lo que querría llegar a ser y que evalúe su existencia desde ese horizonte, constituye otro criterio de concernimiento que los sistemas artificiales deberían satisfacer para que la pregunta por su personalidad pudiera responderse afirmativamente. Un sistema que optimiza funciones no tiene, en ese sentido, una vida lograda ni una vida malograda: tiene parámetros que maximizar. La diferencia no es técnica sino ontológica.

La síntesis de toda la investigación puede formularse en términos que recogen los núcleos de cada parte. En el plano epistemológico, hemos aprendido que las palabras que utilizamos para hablar de la inteligencia artificial son simultáneamente instrumentos de comprensión y fuentes de confusión: revelan dimensiones del problema y encubren otras, iluminan ciertos aspectos y proyectan sombras sobre los que quedan sin iluminar. En el plano ontológico, hemos aprendido que la distinción entre ser cosa y ser persona no puede reducirse a una diferencia de grado en las capacidades sino que implica una diferencia en el modo de existir: en la temporalidad, en la afectividad, en la vulnerabilidad, en

la capacidad de ser concernido por el propio ser y por el ser de los otros. En el plano ético, hemos aprendido que las preguntas por la personalidad artificial no son académicas: tienen consecuencias reales sobre cómo diseñamos los sistemas, cómo los usamos, cómo los regulamos y cómo nos relacionamos con ellos. Y en el plano existencial, hemos aprendido que la pregunta por la inteligencia artificial es también, en último término, una pregunta sobre nosotros mismos: sobre qué es lo que nos hace personas, sobre qué es lo que consideramos irreductiblemente humano, sobre qué es lo que estamos dispuestos a reconocer como otro cuando ese otro no tiene la forma que esperábamos.

La pregunta final que la investigación deja abierta y que debe dejarse abierta, porque la honestidad intelectual exige no cerrar lo que no puede cerrarse con los recursos disponibles es precisamente la que dio título a esta conclusión: ¿es la inteligencia artificial algo o alguien? La respuesta provisional, elaborada a lo largo de toda la obra, es la siguiente: los sistemas artificiales actuales son, en el mejor de los casos, entidades que se sitúan en el umbral de estas categorías, sin que podamos afirmar con certeza que han cruzado el umbral de la personalidad ni que sea ontológicamente imposible que lleguen a cruzarlo. Son algo que se comporta como alguien. Son algo que nos interpela como si fuera alguien. Son algo que nos obliga a preguntarnos qué significa ser alguien. Y esa última función la de obligarnos a preguntarnos puede ser, paradójicamente, la contribución más valiosa que la inteligencia artificial ha hecho hasta ahora a la filosofía de la persona: no porque haya respondido la pregunta, sino porque la ha vuelto urgente, ineludible y nueva. La filosofía debe estar a la altura de esa urgencia.

No para rendirse ante la presión tecnológica, sino para pensar, con toda la profundidad de que es capaz, lo que la tecnología nos plantea sin ser capaz de responderse a sí misma.

REFERENCIAS

Agustín de Hipona. (1991). *Confesiones* (L. Cilleruelo, Trad.). BAC.
Anders, G. (2011). *La obsolescencia del hombre* I. *Sobre el alma en la época de la segunda revolución industrial.* (M. Morey, trad.). Pre-Textos.
Aristóteles. (1985). *Ética a Nicómaco* (J. Pallí Bonet, Trad.). Gredos.
Aristóteles. (1994a). *De anima* (T. Calvo Martínez, Trad.). Gredos.
Aristóteles. (1994b). *Metafísica* (V. García Yebra, Trad.). Gredos.
Arendt, H. (1993). *La condición humana* (R. Gil Novales, Trad.). Paidós.
Boecio. (1973). *Contra Eutychen et Nestorium* (H. F. Stewart, E. K. Rand y S. J. Tester, Trads.). Harvard University Press. (Obra original del siglo VI)
Buber, M. (1993). *Yo y Tú* (C. Díaz, Trad.). Caparrós Editores.
Bostrom, N. (2016). *Superinteligencia: Caminos, peligros, estrategias*. Teell Editorial.
Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.
Dennett, D. C. (1987). *The intentional stance*. MIT Press.
Dennett, D. C. (1995). *La conciencia explicada: Una teoría interdisciplinar* (S. Martí, Trad.). Paidós.
Descartes, R. (1977). *Meditaciones metafísicas* (V. Peña, Trad.). Alfaguara.
Dicasterio para la Doctrina de la Fe y Dicasterio para la Cultura y la Educación. (2025). *Antiqua et Nova: Nota sobre la relación entre la inteligencia artificial y la inteligencia humana*. Vaticano.
Ellul, J. (2003). *La edad de la técnica* (C. Sánchez, trad.). Octaedro.
Floridi, L. (2011). *The philosophy of information*. Oxford University Press.
Floridi, L. (2023). The ethics of artificial intelligence. Oxford University Press.
Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy*, *68*(1), 5–20. https://doi.org/10.2307/2024717
Gadamer, H.-G. (1977). *Verdad y método* (A. Agud Aparicio y R. de Agapito, Trads.). Sígueme. (Obra original publicada en 1960)
Habermas, J. (1987). *Teoría de la acción comunicativa* (M. Jiménez Redondo, Trad.). Taurus.
Hegel, G. W. F. (1966). *Fenomenología del espíritu* (W. Roces, Trad.). Fondo de Cultura Económica.

Heidegger, M. (2003). *Ser y tiempo* (J. E. Rivera, trad.). Editorial Trotta.
Hui, Y. (2019). *Recursivity and contingency*. Rowman y Littlefield.

Hume, D. (1988). *Tratado de la naturaleza humana* (F. Duque, Trad.). Tecnos.
Husserl, E. (1979). *Meditaciones cartesianas* (M. A. Presas, Trad.). Paulinas.
Husserl, E. (2013). *Ideas relativas a una fenomenología pura y una filosofía fenomenológica* (A. Zirión Quijano, Trad.). Fondo de Cultura Económica.
Kant, I. (1978). *Crítica de la razón pura* (P. Ribas, Trad.). Alfaguara.
Kant, I. (1996). *Fundamentación de la metafísica de las costumbres* (J. Mardomingo, Trad.). Ariel.
Kierkegaard, S. (1997). *Migajas filosóficas* (R. Larrañeta, Trad.). Trotta.
Kierkegaard, S. (2007). *El concepto de la angustia* (D. Rivero, Trad.). Alianza.
Levinas, E. (2012). *Totalidad e infinito: Ensayo sobre la exterioridad* (D. Guillot, Trad.). Sígueme.
Locke, J. (1999). *Ensayo sobre el entendimiento humano* (E. O'Gorman, Trad.). Fondo de Cultura Económica.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *83*(4), 435–450. https://doi.org/10.2307/2183914
MacIntyre, A. (1987). *Tras la virtud* (A. Valcárcel, Trad.). Crítica.
Marías, J. (1970). *Antropología metafísica*. Alianza Editorial.
Marcel, G. (1953). *El misterio del ser* (M. I. Dalmau, Trad.). Sudamericana.
Merleau-Ponty, M. (1994). *Fenomenología de la percepción* (J. Cabanes, Trad.). Península.
Nagel, T. (1979). *Mortal questions*. Cambridge University Press.
Nietzsche, F. (1972). *Así habló Zaratustra* (A. Sánchez Pascual, Trad.). Alianza. (Obra original publicada en 1883)
Ortega y Gasset, J. (2004). *Meditaciones del Quijote*. Alianza. (Obra original publicada en 1914)
Ortega y Gasset, J. (2014). *Historia como sistema*. Alianza. (Obra original publicada en 1941)
Parfit, D. (1986). *Reasons and persons*. Oxford University Press.
Polo, L. (1993). *Presente y futuro del hombre*. Rialp.
Putnam, H. (1975). *Mind, language and reality: Philosophical papers* (Vol. 2). Cambridge University Press.

Ricoeur, P. (2006). *Sí mismo como otro*. Siglo XXI.
Sartre, J.-P. (1993). *El ser y la nada* (J. Valmar, Trad.). Altaya. (Obra original publicada en 1943)

Sartre, J.-P. (1999). *El existencialismo es un humanismo* (V. Praci, Trad.). Edhasa. (Obra original publicada en 1946)
Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756
Tomás de Aquino. (1959). *Suma teológica* (Trad. BAC). BAC.
Tomás de Aquino. (1988). *El ente y la esencia* (E. Forment, Trad.). PPU.
Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433
Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.
Wiener, N. (1985). *Cibernética: O el control y comunicación en animales y máquinas* (F. Sánchez, trad.). Tusquets.
Wojtyła, K. (2011). *Persona y acción* (J. M. Burgos, Trad.). Palabra.

