“*Errare machinarum est*: error e intencionalidad más allá de lo humano”

Trabajo de grado presentado para optar al título de Profesional en Filosofía

Programa: Filosofía

Autora:
Dana Isabella Acosta Castillo

Director:
Carlos G. Patarroyo G.

Universidad del Rosario
Escuela de Estudios Sociales, Políticos e Internacionales
Junio de 2026

A mis profesores de filosofía,
a mis amigos,
a Arturo,
a Jorge, Jacquie, Nathy y Cris,
y a quienes me enseñaron que:
las veces que no me equivoco no me vuelven un mejor humano.[1]

[1] Adaptado de *Debí hacerte caso*, de Méne (2023).

Contenido

1. Introducción

En 1986 Honda inició una investigación respecto a la locomoción bípeda, investigación que poco a poco llevó a la creación del robot *Asimo*[2] en el año 2000. Por aquella época Asimo no era un robot autónomo; debía programarse para realizar tareas específicas y muchas veces necesitaba tener una programación muy detallada y puntual con respecto al nivel, tamaño, pasos y más información sobre el plano a recorrer o las escaleras particulares que debía de subir. Era frecuente verlo caerse cuando la superficie sobre la cual caminaba tenía algún desnivel o las escaleras que debía subir no eran de la altura previamente programada en su sistema. En ese entonces, al estar *Asimo* en el suelo, era común escuchar a la gente que presenciaba el evento decir que el robot "había fallado" o que "había un fallo" en su programación. *Asimo* no era diferente, entonces, de una licuadora que se descompone por desgaste, o de un televisor que deja de funcionar después de años de uso. Sin embargo, hoy la "descendencia" de *Asimo* y las evoluciones que se han hecho por otras empresas -como *Figure-03*[3] de 2025- son robots autónomos, que cuentan con reconocimiento externo, inteligencia artificial y que no dependen de la programación exacta de parte de sus creadores para enfrentar situaciones nuevas e inesperadas. Se adaptan el entorno y, de alguna manera, corrigen sus aproximaciones cuando algo ha salido mal en el pasado. Estos robots están diseñados para aprender y autocorregirse, lo cual lleva a la pregunta acerca de qué tanto se puede seguir diciendo que, cuando algo sale mal en su desempeño, hay un "fallo" en el robot o su código, o que el robot simplemente ha "fallado". ¿Sigue siendo válido decir que el robot en este caso no es diferente de la licuadora que falla o del televisor que deja de funcionar? Cuando se mira la literatura en robótica e ingeniería, se puede ver que se utiliza con frecuencia la palabra "error" para referirse a este desempeño que no ha salido como se esperaba. Este uso puede encontrarse en trabajos sobre interfaces cerebro-máquina (Chavarriaga et al., 2014), robots sociales (Mirnig et al., 2017), interacción humano-robot (Honig & Oron-Gilad, 2018)[4], entre otros. Por supuesto, esto puede deberse a una imprecisión lingüística de parte de los investigadores en estas disciplinas, pero no deja de ser llamativo, desde el punto de vista filosófico, si no hay, detrás de este uso, una pregunta más profunda e inquietante: ¿se puede atribuir a un robot, con precisión y solidez argumentativas, la capacidad de cometer errores?

El primer obstáculo al que se enfrenta esta pegunta es el antiquísimo adagio -atribuido a Cicerón- de acuerdo con el cual *errare humanum est* (errar es humano) que, si bien no atribuye exclusividad al humano para la atribución de la categoría de "error", sí se ha utilizado como si dijera "errar es *sólo* de humanos" o "errar es *únicamente* de humanos". ¿Es esto cierto? ¿Qué características tiene el error que puedan llevar a pensar que otros seres (entre ellos seres artificiales como los robots) no serían susceptibles de incurrir en él?

---

[2] Ver: https://www.youtube.com/watch?v=VTlV0Y5yAww

[3] Ver: https://www.youtube.com/watch?v=4ZP943-gARQ

[4] En este último se presenta una taxonomía de los distintos fallos que se pueden presentar entre humanos y robots. Allí el error "humano" recae dentro de los fallos de interacción, a diferencia de los fallos de diseño como dentro del software o hardware. Ver Figura 1.

En su libro *On Error*, Nicholas Rescher defiende que el error es, ante todo, un fenómeno teleológico, es decir, solo existe en relación con un propósito. Como él mismo afirma:

> Los errores suelen surgir en relación con los objetivos y propósitos. **Requieren intención**—al menos implícitamente respecto a los tipos de propósitos que la gente debería tener. El error es un concepto fundamental e intencional que tiene en cuenta la realización de ciertos objetivos. (2007, p.3, mi traducción y negrilla).

Si Rescher tiene razón, y la intencionalidad es una característica *sine qua non* del error, defender que los robots pueden errar implicaría defender que pueden exhibir intencionalidad, algo cuando menos polémico. Esta es una posibilidad que el mismo Rescher rechaza tajantemente

> Decimos: 'John cometió un error', y no: 'Un error le sucedió a John'. Un ser incapaz de actuar voluntariamente puede, en efecto, hacer cosas que son contraproducentes para sus intereses […]; pero no comete errores […]. Los dispositivos inertes – máquinas e instrumentos - no cometen errores; funcionan mal [*they malfunction*] (Rescher, 2007, p. 3, mi traducción).

Que Rescher lo afirme con tanta seguridad, no necesariamente implica que tenga razón. Pero antes de explorar esa vía, es necesario ver si la caracterización ofrecida por Rescher es o no correcta. Para ello, se emprenderá un breve recorrido histórico para comprender cómo se ha tratado el concepto de error en la historia de la filosofía y extraer de ahí una caracterización que permita ver qué tanto la posición de Rescher le hace justicia.

2. Recorrido histórico

2.1. Aristóteles

Sin duda un referente ineludible en la filosofía, y en este caso también en la filosofía de la acción, es Aristóteles[5]. Se iniciará el recorrido histórico por su posición frente a las acciones, la voluntariedad de estas y el error. Aristóteles distinguía entre 3 tipos de acciones: 1) *Voluntarias*: "Llamo voluntario, […], a lo que hace uno estando en su poder hacerlo y sabiendo, y no ignorando, a quién, con qué y para qué lo hace" (Aristóteles, Ética a Nicómaco, 1135a). Se trata de aquellas cuyo principio está en el agente y que se realiza con conocimiento de las circunstancias relevantes. Implican deliberación, elección y control. El agente sabe lo que hace, considera alternativas y actúa sin coacción externa. 2) *Involuntarias*: "[…] aquello cuyo principio es extrínseco, siendo tal que en él no pone nada de su parte el agente." (1110a 1-3). Son aquellas acciones que se realizan por fuerza o por ignorancia de circunstancias particulares, y que generan pesar o arrepentimiento cuando se reconoce lo ocurrido. Esto implica que, a diferencia de lo voluntario, el agente no es propiamente el origen de la acción, sino más bien aquello sobre lo que esta recae, en la medida en que su comportamiento se encuentra coaccionado por factores externos. El ejemplo que ofrece Aristóteles es del marinero que, en medio de la tormenta y ante la posibilidad de naufragio, se ve obligado a tirar su mercancía por la borda para aligerar el peso de la embarcación. En este caso, si bien tirar la mercancía por la borda ha sido una acción del marinero, el origen de esta ha estado en circunstancias externas que lo han forzado a ella. En el caso de la ignorancia, el carácter involuntario radica en ignorar aspectos decisivos de la acción, como a quién afecta, con qué medios se realiza o cuáles son sus consecuencias; el ejemplo de Aristóteles es el de la persona que revela lo que alguien más le ha dicho, sin saber que le fue dicho en confidencia o secreto. En suma, se trata de acciones porque las realiza el agente (en oposición a que el agente sea empujado por otro objeto o algo así), pero lo que el agente realiza va en contra de lo que más desearía -el marinero es un comerciante y no desea perder su mercancía, sin embargo, la tormenta, como causa externa, lo fuerza a tirarla por la borda-. 3) *No voluntarias*: "el que por ignorancia hace algo, cualquier cosa que ello sea, sin sentir el menor desagrado por su acción, no ha obrado voluntariamente, puesto que no sabía lo que hacía, pero tampoco involuntariamente, ya que no sentía pesar" (Aristóteles, Ética a Nicómaco, 1110b). Estas son acciones que se parecen más a las voluntarias que a las involuntarias. El agente las hace sin aparente coacción o presión externa. Sin embargo, el hecho de que no generen arrepentimiento es para Aristóteles indicador de que hay algo mal en el agente. Y en este caso se trata de una ignorancia de lo bueno o lo correcto. Por eso dice que son acciones que, de todas maneras, se realizan por ignorancia[6]. Aquí el problema no es solo no saber un detalle,

[5] Principalmente porque para Platón el error era un asunto de conocimiento, específicamente de ignorancia. Aristóteles entra a ampliar este debate y enriquecerlo con perspectivas más profundas.

[6] Aristóteles distingue entre acciones realizadas *por* ignorancia, de las realizadas *con* ignorancia. Para él alguien ebrio o llevado a la violencia por una pasión desenfrenada actúa *con* ignorancia. Es decir, no ve en él un fallo fundamental sino simplemente una ausencia temporal del buen juicio debido a alguna circunstancia específica (el

sino no estar en condiciones de deliberar adecuadamente. El agente actúa sin una comprensión suficiente de lo que está en juego y esto no le genera ningún remordimiento.

Hay, si se quiere, error en las acciones involuntarias y no en las no-voluntarias. En las primeras el agente no logra hacer lo que quería hacer (en el caso del marinero, llegar con su carga salva al puerto, o en caso del amigo, respetar los vínculos de amistad) pero no lo ha logrado porque ha sido forzado por causas externas a desviarse de su propósito (marinero) o porque ignoraba algo que era esencial para su acción (amigo al ignorar que la información era un secreto). Los errores aquí no son una muestra de una mala voluntad o un mal carácter de parte del agente. Son circunstancias ajenas a él que le han desviado de su propósito. En cambio, en las acciones no-voluntarias lo que se produce es la exitosa consecución de algo que se desea, pero que no es lo debido. Se escoge mal, por una falla en el carácter. Y esta elección revela un carácter desviado, maligno[7]. Apelando, *por mor* de la claridad, a un anacronismo, se podría decir que en las acciones no-voluntarias no se ve un error, sino mala fe.

Pero Aristóteles no habla del error únicamente en el caso de las acciones. En *De anima* trata también el error perceptual. Para él el intelecto puede aprehender objetos simples, así como también construir composiciones con ellos. La relación del intelecto con los objetos simples es directa, pura y simple. No hay lugar allí para el error. Este aparece únicamente en las composiciones:

> La intelección de los indivisibles tiene lugar en aquellos objetos acerca de los cuales no cabe el error. En cuanto a los objetos en que cabe tanto el error como la verdad, tiene lugar ya una composición de conceptos que viene a constituir como una unidad. […] En cuanto a los acontecimientos pasados o futuros, el tiempo forma parte también de la intelección y la composición. El error, en efecto, tiene lugar siempre en la composición. Y es que al afirmar que lo blanco es no-blanco, se ha hecho entrar a lo no-blanco en composición. (*De anima*, III, 6, 430ª)

El error no hace parte de los objetos, dice Aristóteles, esto lo tiene claro también en la *Metafísica*: "La falsedad y la verdad no se dan, pues, en las cosas (como si lo bueno fuera verdadero y lo malo, inmediatamente falso) sino en el pensamiento. Y tratándose de las cosas simples y del qué-es, ni siquiera en el pensamiento." (*Metafísica*, VI, 1027b-25) esta cita muestra dos cosas. La primera de ellas es que recalca lo dicho anteriormente, a saber, que el error se da acerca de las composiciones y no acerca de lo simple. Y lo segundo es que muestra que lo verdadero y lo falso, y, por ende, el error, no se dan en las cosas mismas sino en el pensamiento al hacer las composiciones. El individuo, al querer capturar en una descripción la forma de algún objeto, realiza una composición con varios conceptos y *erra* cuando introduce allí alguno inadecuado. En esto Aristóteles no se distancia mucho de Epicuro, y este último quien ahonda más en este asunto, pues Aristóteles poco dice acerca de este tipo de error y de sus causas[8].

---

alcohol o la ira). En cambio, actúa *por* ignorancia quien, sin condiciones externas apremiantes, escoge lo inconveniente y no siente pesar por ello: "Pues todo lo malvado desconoce lo que debe hacer y de lo que debe apartarse, y por tal falta son injustos y, en general, malos." (110b 25)

[7] Las acciones involuntarias, realizadas por ignorancia y acompañadas de arrepentimiento, muestran una ruptura entre lo que el agente quería hacer y lo que efectivamente hizo. En cambio, en las acciones no voluntarias, aunque también haya ignorancia, no se da esta distancia en términos de la propia intención, lo que modifica la manera en que atribuimos responsabilidad.

[8] Keeler, en su artículo de 1932 "Aristotle and the Problem of Error" describe esta situación con las siguientes palabras: "Al pasar de Platón a Aristóteles pareciera que el problema del error ha sido deliberadamente archivado.

2.2. Epicuro

Epicuro afirma que: “El error surge siempre de la <<opinión>> (dóxa)” (Carta a Heródoto, [37])” - . Para él, el error no está en la percepción, pues:

> Todas las sensaciones son verdaderas (...). No hay nada que pueda refutar a las sensaciones: ni una sensación similar puede refutar a otra similar por tener el mismo valor, ni una disimilar a una disimilar, ya que no juzgan el mismo objeto. (Epicuro, Carta a Heródoto, 50-52)

El problema no está en sentir, pues esta percepción es siempre verdadera, sino en lo que se añade después, el juicio que se basa en esta percepción. El error aparece cuando se forma una opinión que va más allá de lo dado en la sensación. Es allí donde puede aparecer el error. Por ejemplo, alguien que ve una torre a lo lejos puede pensar que la torre es cilíndrica, cuando en realidad, en una inspección más cercana, resulta que es un ortoedro. (Lucrecio, *De la naturaleza de las cosas*, 323-468, p.50). La sensación, en cuanto tal, es verdadera: así se presenta. El error surge al afirmar que la torre es realmente cilíndrica, sin examinarla más de cerca. El lugar del error no es la percepción misma, sino el exceso en el juicio. Es decir, el error ocurre cuando la mente añade, sustrae o modifica información no proporcionada por los sentidos sin esperar su confirmación o refutación. Para Epicuro el error no es un defecto del alma por ser material, ni una falla inevitable de la constitución corporal humana. Surge en el paso que se da desde la sensación hacia la opinión. Es precisamente porque alma y cuerpo están unidos que es posible sentir y percibir (cuerpo) y luego formar juicios sobre lo experimentado (alma). Sin embargo, este proceso debe hacerse con cuidado, y la premura de pasar de lo uno a lo otro, sin verificación o evidencia sensorial suficiente, es lo que induce al error.

> La mentira y el error se encuentran entre los supuestos pendientes siempre de ser confirmados o no ser confirmados por un testimonio, surgiendo la mentira y el error precisamente cuando esos supuestos no son confirmados luego por el testimonio de la imagen inamovible […] conectada con la aprehensión imaginativa del cuerpo sólido, pero que guarda respecto a este una separación. (Epicuro, *Carta a Heródoto*, 50)

Como muestra la cita, el error para Epicuro es esta premura de formar una opinión o juicio en ausencia del *testimonio* que ofrece esa imagen que es inamovible y que está ligada al objeto externo, es decir, la percepción más cercana y fidedigna de él. Se evita, entonces, con la mesura y la verificación, con el llamado al testimonio faltante antes de apresurarse a la formación del juicio o la opinión. El error, en suma, es también (como lo era en Aristóteles) una desviación de lo que se quiere hacer. Se desea describir acertada o correctamente un cierto objeto, pero por premura y afán, no se espera la verificación de la experiencia más cercana y fidedigna, lo que lleva a afirmar cosas sobre él que no le corresponden verdaderamente.

2.3. Agustín de Hipona

Por su parte, San Agustín de Hipona presenta una aproximación diferente. En él el error hace parte del enfoque sobre el vivir bien, y no afecta solo la serenidad o el juicio correcto, sino la orientación

---

Aristóteles no muestra signo alguno, de encontrarse perplejo por él. En ningún lugar lo encontramos luchando con la pregunta, tan familiar para la generación previa, acerca de cómo es posible decir o pensar lo que no es.” (p. 242)

misma del alma. Conocer y vivir bien no son asuntos separados; pero ahora esa unión se inscribe en una relación con la verdad divina[9] y con la responsabilidad ante ella. según Charles Bolyard en "Augustine on Error and Knowing That One Does Not Know", dentro de la literatura de Agustin se pueden distinguir dos categorías del error – errores cognitivos y errores volitivos –. Cada una de ellas tiene, a su vez, distintos tipos de error. Por un lado, en los errores cognitivos se encuentran los siguientes: i) Error estándar – "El error, pienso yo, consiste en afirmar lo falso por verdadero" (Contra Acad, I, 4, 11), ii) Duda – o afirmar sin certeza. No es simplemente ignorar, es dar el paso del asentimiento sin fundamento suficiente. Por ejemplo, alguien escucha un rumor y, aunque no tiene pruebas, lo da, por cierto. Y, iii) "El poder del diablo para engañarnos" – En *De Trinitate*, San Agustín sugiere un tipo de error diferente a los dos mencionados anteriormente; sugiere que algo en el "hombre interno" está siendo persuadido para vivir en un estado de engaño y sin una búsqueda que pueda conducir a la verdad. El error no sería solo una conclusión mal construida, sino una especie de seducción interior. Algo persuade, inclina, arrastra. Ahora bien, esto nos lleva a que el error puede tener una dimensión interior que excede el cálculo racional. No hay error sin algún tipo de asentimiento. Y no hay asentimiento sin voluntad. Y la voluntad puede ser inclinada, tentada o seducida por otras fuerzas. Combatirlas es parte de la responsabilidad y deber de quien busca tanto la verdad como llevar una vida buena.

Por otro lado, hay errores que no dependen del contenido del juicio, sino que tienen un enfoque hacia la actitud frente a la verdad, así se definen: iv) No dar asentimiento- o de buscar y no encontrar, pero aun así asentir o instalarse en la confusión, por ejemplo, en alguien que investiga un tema, encuentra información contradictoria o insuficiente y, en lugar de suspender el juicio o seguir indagando, decide aceptar apresuradamente una conclusión sin fundamento claro. Y por la misma línea, v) Por pereza – El que no trata de buscar (la verdad) en absoluto y lo cataloga el Santo como un error "más grave". Por ejemplo, alguien que tiene la posibilidad de examinar, de preguntar, de aprender, pero elige no hacerlo por comodidad o desinterés. Aquí el error no es meramente una proposición falsa concreta, sino que refleja una disposición interior.

Estas distinciones permiten ver que algunos errores surgen de cómo se conoce; otros surgen de cómo se actúa para conocer. Así, mientras en Epicuro el error surge en el juicio que excede la percepción, y en Aristóteles se vincula con la acción y las condiciones en que esta se realiza, en Agustin de Hipona el error se desplaza hacia la interioridad del sujeto: no es solo una falla en lo que se conoce o se hace, sino en la disposición misma frente a la verdad. Por eso, en Agustín la voluntad ocupa un lugar central. El error no es solo un tropiezo intelectual, sino también una forma de orientar (o no) la propia vida frente a la verdad.

2.4. Descartes

Ahora bien, para entender la aproximación de Descartes al error, primero se ha de recordar que en las *Meditaciones* aplica su método de la duda hiperbólica o duda metódica a fin de llegar a un conocimiento indubitable que le permita sentar sobre él la reconstrucción del edificio del

[9] La verdad divina es la verdad eterna e inmutable que no depende de nuestra mente y que fundamenta todas las demás verdades; se obtiene por revelación a quien ha trabajado para merecerla, es por eso por lo que en Agustín llevar una vida buena y buscar la verdad son inseparables. La verdad divina (i.e. la verdad última o definitiva) es presentada en *De Trinitate* como idéntica a la relación con Dios como *summa veritas*. En *Confesiones* (VII, 10, 16) lo expresa así: "Dios es "interior a lo más íntimo mío y superior a lo más alto mío", es decir, la verdad que ilumina desde dentro y nos trasciende.

conocimiento. Es en este proceso de reconstrucción, una vez que ha llegado a su famoso *cogito ergo sum* que, en la cuarta meditación, trata el tema el error. Si bien ha llegado a un fundamento indubitable para su edificio, y ha demostrado (o cree haber demostrado) la existencia de un dios que es la suprema bondad y, con ello, la no existencia de un genio maligno engañador, el error sigue existiendo en los asuntos humanos y merece de una explicación, para, en primer lugar, entenderlo, y con ello, saber cómo evitarlo.

Descartes presenta el siguiente escenario: si el ser humano es capaz de alcanzar el conocimiento de las otras cosas a través de la contemplación de Dios, pues en él se encuentran contenidos los tesoros de las ciencias y de la sabiduría, y el ser humano es a la vez parte de su creación, entonces, ¿cómo es posible que se cometan errores? (Descartes, *Meditaciones de filosofía primera*, Meditación IV, AT VII, 53). Para Descartes esta cuestión no tiene una respuesta sencilla, ya que errar, como pecar, son cosas que no parecen encajar en la perfección divina. Él sugiere tres pasos clave para entender cómo se llega al error. Primero, Dios no puede ser engañador, pues eso implicaría imperfección y demostraría malicia o debilidad, cualidades que no son propias de Dios. Segundo, la facultad de juzgar del ser humano ha debido ser dada por Él. Y, tercero, como por el primer punto no puede ni quiere engañar, entonces no ha debido dar al ser humano la facultad de equivocarse cuando usa su capacidad de juzgar correctamente. Esto indica que, en teoría, el ser humano debería poder no equivocarse (*Meditaciones de filosofía primera*, Meditación IV, AT VII, 53-54). Sin embargo, es evidente que esto no es así. La clave, entonces radica en el uso de la facultad de juzgar. Para Descartes, el error no proviene de Dios ni de la facultad en sí misma, sino de la manera en que es utilizada. Al juzgar precipitadamente o sin suficiente claridad, se incurre en el error. Dios ha dado la capacidad de conocer la verdad, pero es responsabilidad de quien la tiene emplearla correctamente, reconociendo siempre sus límites como ser finito, débil y dependiente.

En otras palabras, los errores se producen porque la voluntad es más amplia que el entendimiento. El afán o el deseo de llegar a una opinión o emitir un juicio superan a la información que el entendimiento proporciona y allí aparecen la imprecisión y el error. Descartes muestra que el error no surge de una imperfección en las facultades que Dios ha dado, sino del mal uso de libre albedrío humano: “Es en ese uso incorrecto del libre albedrío donde se encuentra la privación que constituye el error” (*Meditaciones de filosofía primera*, Meditación IV, AT VII, 60). Esta privación o ausencia no se halla en la facultad de juzgar ni tampoco en la libertad que se ha concedido al ser humano. En lugar de ello, la operación que implica saber hasta dónde debe llegar el entendimiento, dada la amplitud de la voluntad, no depende de Dios, sino de cada persona que tiene esas facultades. Esta falta se encuentra, específicamente, en el mal uso de esa libertad, ya que "la percepción del entendimiento debe preceder siempre a la determinación de la voluntad" (*Meditaciones de filosofía primera*, Meditación IV, AT VII, 60). Es decir, primero se ha comprender claramente antes de emitir un juicio o tomar una decisión.

Pero, si esto es así, surge otra pregunta: ¿por qué Dios no ha dado el ser humano una voluntad más ajustada a los límites de su entendimiento, de modo que no se equivoque? Para Descartes la voluntad, al ser ilimitada, es parte de lo que hace semejante al ser humano con Dios; es la libertad donde está la semejanza y la diferencia yace en la lejanía con su capacidad infinita de comprensión. En realidad, la existencia del error y la posibilidad de equivocarse no disminuyen la perfección del mundo, sino que la incrementan en la medida en que permiten la diversidad, la libertad y la individualidad de las criaturas. Un universo donde todo fuera perfecto en el sentido de que nadie pudiera errar sería también un universo en el que no habría verdadera libertad, ni autonomía, ni

responsabilidad. Para Descartes, esa no sería una realidad más perfecta, sino más limitada, porque prescindiría de la riqueza que supone que cada criatura actúe según sus propias facultades. Así pues, Descartes afirma: "No hay imperfección en Dios porque me haya dado libertad para asentir o no a cosas que no ha puesto en mi entendimiento con una percepción clara y distinta; la imperfección está en mí, porque no uso bien esa libertad y juzgo sobre lo que no entiendo correctamente" (*Meditaciones de filosofía primera*, Meditación IV, AT VII, 60-61). Esto parece resonar con ideas que ya se han visto anteriormente. Epicuro también sostenía que el error se encuentra en la opinión; no en el mundo o en las percepciones, sino dentro de cada uno. Según él, el error radica en la opinión que no espera confirmación, en los juicios precipitados que se emiten sin conocer todos los hechos o pruebas, y que no se toman el tiempo de esperar, juzgando lo que no se comprende correctamente.

2.5. Locke

Ahora giremos hacia otra postulación: la empirista. Si los autores anteriores tendían a explicar el error desde el papel de la razón, la voluntad o el grado de perfección del entendimiento, el empirismo introduce un cambio importante de perspectiva: el análisis del conocimiento debe comenzar por examinar el origen de nuestras ideas y la forma en que la experiencia configura nuestro entendimiento. En este contexto, el problema del error ya no se explica únicamente por un uso indebido de la voluntad o por la imperfección del intelecto, sino también por la manera en que formamos, combinamos y evaluamos nuestras ideas. Por esta razón, resulta especialmente interesante analizar la propuesta de John Locke, particularmente en el *Ensayo sobre el entendimiento humano*, Libro Cuarto: Del conocimiento, capítulo XX. Del falso asentimiento, o del error.

Locke abre este capítulo con una afirmación que delimita con claridad el lugar del error: "Puesto que el conocimiento no se obtiene sino de la verdad cierta y visible, el error no es una falla de nuestro conocimiento, sino un equívoco de nuestro juicio que presta su asentimiento a lo que no es verdadero" (Locke, 1689/2005, p. 713; Lib. IV, Cap. XVI, § 1). Desde el inicio, Locke distingue cuidadosamente entre conocimiento y juicio. El conocimiento, en sentido estricto, implica verdad; por eso, cuando se erra, no se lo hace en el conocimiento mismo, sino en el momento en que se asiente a algo sin que exista evidencia suficiente para sostenerlo. En este sentido, Locke parece entender el conocimiento más como un estado o resultado correcto del entendimiento que como una facultad susceptible de fallar por sí misma. A partir de esta distinción, Locke se propone explicar por qué los seres humanos caen en el error. Para ello identifica cuatro causas principales[10]:

1) **La falta de pruebas:** Para desarrollarla, distingue entre dos situaciones distintas. Por un lado, existen casos en los que las pruebas simplemente no están al alcance. Por otro, hay situaciones en las que las pruebas sí podrían obtenerse, pero no se las busca, muchas veces por lo que se suele llamar "falta de tiempo"[11]. Esta explicación deja abiertas dos formas distintas de enfrentar el error. En los casos en que las pruebas existen, pero no se las busca,

[10] Aunque estas causas recuerdan en parte a explicaciones que se han visto anteriormente, la manera en que Locke las articula introduce un giro: el error no surge tanto de una imperfección metafísica del entendimiento, sino de problemas en la forma en que se evalúa, interpreta o utiliza la evidencia disponible.

[11] La expresión está entre comillas porque el propio Locke matiza inmediatamente esta excusa. Al responder a la posible objeción de que muchas personas están demasiado ocupadas para investigar ciertos temas, afirma que nadie carece realmente de tiempo para reflexionar sobre cuestiones fundamentales, como el alma o la moral.

el problema es claramente una forma de negligencia intelectual: se prefiere dedicar el tiempo a otras cosas y se termina aceptando afirmaciones sin examinarlas adecuadamente. En cambio, cuando las pruebas son realmente imposibles de obtener, la actitud correcta no es emitir un juicio definitivo, sino suspender el asentimiento o mantener la afirmación en el ámbito de la probabilidad.

2) **La falta de habilidad para emplearlas:** Las personas que yerran por esta razón no necesariamente carecen de evidencia, sino que no saben cómo utilizarla correctamente. Según Locke, esto ocurre cuando alguien es incapaz de seguir mentalmente una cadena de consecuencias o de ponderar adecuadamente los distintos grados de probabilidad. Como resultado, puede terminar dando su asentimiento a proposiciones que en realidad no son suficientemente probables. Locke ilustra esta deficiencia al señalar que "Hay hombres de un silogismo, otros de dos silogismos, y eso es todo, y otros que pueden dar un paso más. Gente como esa no siempre es capaz de discernir de qué lado están las pruebas más convincentes" (Locke, 1689/2005, p. 716; Lib. IV, Cap. XVI, § 5).

3) **La falta de voluntad para utilizarlas:** A diferencia de la primera y segunda causa, aquí las pruebas pueden estar al alcance del sujeto, pero este decide no examinarlas con seriedad. El error surge entonces no por falta de información ni por incapacidad para entenderla, sino por una disposición voluntaria a evitar el esfuerzo intelectual que implicaría evaluarla. Esta falta de voluntad puede manifestarse de distintas maneras. A veces se debe a la comodidad o al autoengaño: alguien prefiere aceptar una creencia que le resulta conveniente antes que revisar críticamente la evidencia que podría contradecirla. Piénsese, por ejemplo, en quien intenta justificar un hábito como fumar, diciendo que “todo el mundo lo hace” o que “no hay evidencia clara de que sea tan dañino”, ignorando deliberadamente los datos disponibles. En otros casos aparece como simple negligencia intelectual. Locke observa que muchas personas cuentan con el tiempo y las capacidades necesarias para perfeccionar su entendimiento, pero aun así prefieren permanecer en una especie de ignorancia voluntaria. Nótese que, lo que comenzó como una omisión voluntaria de la evidencia (tercera causa) deriva con el tiempo en una incapacidad funcional para procesarla (segunda causa). Como él mismo afirma: “No sabría decir cuántos hombres hay cuya amplitud de medios les permite el tiempo necesario para perfeccionar sus entendimientos, pero que se conforman con permanecer en una ociosa ignorancia” (Locke, 1689/2005, p. 717; Lib. IV, Cap. XVI, § 6).

4) **El uso de falsas medidas de probabilidad:** Este tipo de error aparece cuando, al no tener certeza completa, las personas deben juzgar basándose en lo que consideran más probable. El problema surge cuando los criterios que utilizan para medir esa probabilidad son defectuosos. Locke identifica cuatro fuentes principales de este tipo de error:

    a. **Las proposiciones dudosas aceptadas como principios**, se trata de afirmaciones que se aceptan como verdaderas, aunque carezcan de pruebas sólidas. Muchas veces esto ocurre porque la idea parece intuitivamente plausible o porque se ha repetido tantas veces que termina adquiriendo apariencia de verdad. Locke menciona, a modo de ejemplo, cómo a algunos niños se le inculca el miedo a la oscuridad mediante cuentos o fábulas sobre criaturas imaginarias que, por costumbre y por haberlo escuchado desde que eran muy pequeños, se consolida como una creencia acerca de la existencia de criaturas fantásticas en la oscuridad.

b. **Las hipótesis recibidas**, son teorías o suposiciones adoptadas sin suficiente examen crítico. El problema aparece cuando estas hipótesis comienzan a funcionar como un trasfondo o fundamento desde el cual se interpreta nueva información. Por ejemplo, en ciertos momentos de la historia algunas teorías médicas fueron aceptadas durante siglos simplemente porque formaban parte de la tradición académica; es el caso de la antigua teoría de los humores, que explicaba las enfermedades como desequilibrios entre sangre, flema, bilis amarilla y bilis negra, y que fue repetida y enseñada durante mucho tiempo, incluso cuando muchas observaciones empíricas comenzaban a ponerla en duda[12].

c. **Las pasiones predominantes,** aquí el error surge cuando emociones fuertes influyen en el juicio y nos llevan a aceptar como probable aquello que en realidad solo coincide con deseos o temores. El orgullo, el miedo, la ambición o el deseo de reconocimiento pueden inclinar el entendimiento hacia ciertas conclusiones. El ejemplo más claro lo da el mismo Locke cuando dice: "Dígasele a un hombre apasionadamente enamorado que se le engaña; preséntesele un ejército de testigos acerca de la infidelidad de su amante; es diez contra uno, que bastarán tres palabras amorosas de ella para invalidar todos estos testimonios. *Quod volumus, facile credimus* "Fácilmente creemos lo que deseamos"." (Locke, 1689/2005, p. 722; Lib. IV, Cap. XVI, § 12). La pasión es una fuerza enorme contra la cual la evidencia, el silogismo y el testimonio contrario parecen inefectivos y apenas débiles oposiciones sin chance alguno.

d. **La autoridad,** esta es, quizá, una de las fuentes más comunes de error, pues en estos casos las personas aceptan una afirmación simplemente porque proviene de alguien con prestigio, poder o fama, sin examinar por sí mismas la evidencia que la respalda. El respeto o la admiración hacia ciertas figuras pueden llevar a que sus opiniones se tomen como pruebas suficientes, incluso cuando no lo son; "¡Como si los hombres de buena fe o los hombres muy leídos no pudieran incurrir en errores; como si la verdad pudiera establecerse por medio del sufragio de las multitudes! Sin embargo, a eso se acomodan la mayoría de los hombres". (Locke, 1689/2005, p. 725; Lib. IV, Cap. XVI, § 17).

En general, Locke introduce una diferencia respecto al error dado por el mal uso de la voluntad. Para Locke, el asentimiento no depende enteramente de la libertad, cuando la mente percibe claramente la relación entre dos ideas o cuando, después de examinar una cuestión, una opción aparece como más probable que otra, el entendimiento tiende necesariamente a aceptarla. En sus propias palabras, "[…] así como el conocimiento no es en nada más arbitrario que la percepción, así también yo creo que el asentimiento no está más en nuestro poder que el conocimiento" (Locke, 1689/2005, p. 724; Lib. IV, Cap. XVI, § 16). De este modo, el sujeto no puede simplemente elegir creer lo que quiera cuando la evidencia se presenta con suficiente claridad o probabilidad. Sin

[12] ¿No es esto muy similar a las proposiciones dudosas aceptadas como principios? Locke es consciente de esta posible similitud y ofrece una aclaración. Para él aquellos que han recibido y cimentado en su mente las proposiciones dudosas y las han establecido como principios, son diferentes (o actúan de manera diferente) de aquellos a quienes aquejan las hipótesis recibidas: "Junto a aquellos hombres [refiriéndose a los de las proposiciones dudosas aceptadas como principios], hay otros cuyos entendimientos están vaciados en un molde y cincelados para solo ajustarse a una hipótesis recibida. La diferencia entre unos y otros consiste en que estos últimos admiten las cuestiones de hecho, y se ponen de acuerdo con ese respecto con sus adversarios, pero difieren solo en las razones asignadas y en la explicación de las maneras de operación." (Locke, 1689/2005, p. 720; Lib. IV, Cap. XVI, § 11).

embargo, esto no elimina por completo la responsabilidad por el error. Locke sostiene que sí está en poder del ser humano suspender la investigación o dejar de emplear sus facultades en la búsqueda de la verdad. En consecuencia, el límite de la agencia no se encuentra en el momento final del juicio, sino en la decisión previa de examinar o no las pruebas disponibles. Entonces, el error no surge tanto de elegir mal entre dos creencias evidentes, sino de no haber investigado lo suficiente para medir correctamente la probabilidad de aquello que se afirma, como ya se mencionó anteriormente.

2.6. Leibniz

A continuación, se contrastará la visión de Locke con el capítulo XX de los *Nuevos ensayos sobre el entendimiento humano* de Leibniz, una obra concebida precisamente como respuesta al Ensayo lockeano. Por ello, conviene aclarar un aspecto estructural del texto: en los *Nuevos ensayos* no se encuentra una exposición lineal, sino un diálogo donde Filaletes representa la posición empirista cercana a Locke, mientras que Teófilo encarna la perspectiva de Leibniz. Si bien en los pasajes que siguen es Filaletes quien lleva la voz al enumerar los tipos de error, Teófilo interviene para señalar en qué puntos complementa esta visión o en cuáles se distancia de ella. Esta estructura dialógica permite que la exposición de la postura contraria sirva como el marco necesario para el desarrollo de la tesis leibniziana.

Así pues, Filaletes enumera las cuatro causas del error tomadas de Locke: la carencia de pruebas, la poca habilidad para usarlas, la falta de voluntad para utilizarlas y, finalmente, las reglas falsas de probabilidad, las cuales a su vez se derivan de cuatro fuentes distintas: las proposiciones dudosas, las hipótesis admitidas, las pasiones dominantes y la autoridad.

Como se puede notar, estas causas y fuentes ya han sido planteadas desde Locke, aquí Leibniz en el papel de Teófilo contraargumenta a Filaletes en: *La poca habilidad de usar las pruebas*, *las proposiciones dudosas y, las hipótesis admitidas*[13]. Así pues, queda en evidencia cuál es la cuestión diferencial: la distinción entre percepción y juicio. Este aspecto es resaltado por autores como Stefano Di Bella en *Leibniz on Error: Between Descartes and Spinoza. Will, Judgement and the Concept of Reality, 2016*; Keya Maitra en *Leibniz's Account of Error,* 2002 y; Matteo Favaretti Camposampiero en *Before Judging. Leibniz on the Ultimate Origin of Error*, 2016. Particularmente, Camposampiero examina por qué se cometen errores si, según Leibniz, toda percepción es verdadera por fundarse en la visión divina del universo. El núcleo de este análisis radica en que el error no pertenece a la percepción, sino al juicio: las percepciones 'siempre son verdaderas', pero son los juicios los que engañan. Aunque esta postura guarda cierta afinidad con Epicuro al rescatar la verdad de la percepción, Leibniz se distancia de su materialismo. Mientras que para Epicuro el alma recibe los átomos de manera directa y sin mediaciones, Leibniz propone

[13] *Poca habilidad:* Leibniz discrepa con respecto a las limitaciones mentales o sobre la falta de educación que postula Locke. El entendimiento humano es perfectamente capaz de seguir la lógica elemental. El problema no es una incapacidad o "defecto físico" de la mente, sino la falta de atención y de memoria al momento de mantener el hilo argumentativo.
Las proposiciones dudosas y las hipótesis admitidas (como reglas falsas de probabilidad): Leibniz Sostiene que la mente nunca elige el error de manera libre y consciente. El juicio no decide equivocarse. Lo que ocurre es que es engañado a nivel perceptivo. Las "pequeñas percepciones" insensibles y las pasiones confunden al alma, presentándole lo falso bajo la apariencia de verdadero.
En general, Leibniz apoya a Locke en cuáles son los detonantes externos del error (carencia de pruebas, falta de voluntad, pasiones dominantes y la autoridad), pero discrepa en las causas internas.

una dimensión interna y activa del alma racional. Aquí, la percepción pasa por filtros y niveles organizativos que pueden provocar el error, no debido a una desviación física, sino a una interpretación apresurada o incompleta. De este modo, el error leibniziano no es un accidente fortuito, sino el reflejo del grado de perfección del alma. Por ello, cuando Leibniz afirma que esto depende del "grado presente de perfección del agente", introduce algo que complica la apelación directa a la libertad. Ese grado de perfección no es algo que se elige instantáneamente; es el resultado de la constitución humana, de nuestra historia, nuestra formación y el lugar que ocupamos en el orden del mundo. Entonces, ¿desaparece la libertad? No exactamente. Pero tampoco ocupa el mismo lugar central que en Descartes. Pues, para Descartes el error ocurre porque la voluntad se adelanta. Hay un acto libre de asentimiento indebido. En contraste, para Leibniz, el juicio mismo está determinado por el estado actual de la mente. Se juzga según el grado de claridad que se posee en ese momento. Si la percepción es confusa, el juicio tenderá a ser confuso. La atención misma no surge como un acto absolutamente espontáneo, sino que está condicionada por el nivel de desarrollo intelectual.

Sin embargo, esto no elimina toda responsabilidad. Lo que cambia es su profundidad temporal. No se trata tanto de que en el instante en el cual se cometió un error, se pudo haber hecho otra cosa, con absoluta independencia, sino de que la mente puede perfeccionarse progresivamente. El entrenamiento, el hábito de examinar y, el cultivo del entendimiento, forman parte del desarrollo del agente. Se es responsable en la medida en que se participa en ese proceso de perfeccionamiento, aunque no se controle absolutamente cada acto singular. Como lo expresa Leibniz, la libertad y, por extensión, la responsabilidad, consisten en "el uso de la razón, que nos pone en estado de elegir lo mejor" (1999, p. 204) Así, el error en Leibniz está vinculado a la libertad en un sentido más amplio, como desarrollo gradual de la racionalidad, no como decisión aislada en el momento del juicio. Esto no es un asunto menor porque, como será más adelante, es esta capacidad de corregir, de mejorar, de aprender del error para no cometerlo, será esencial en la comprensión del error que esta investigación busca lograr.

Dado que el error en Leibniz no posee una realidad positiva, resulta preciso examinar dos nociones fundamentales que interpelan al juicio y a la percepción: la privación y la ignorancia. Al no gozar de entidad metafísica propia, el error no opera como una fuerza disruptiva en el entendimiento ni como una manifestación activa del mal, sino que se define como una carencia. Podría decirse, entonces, que el error reviste un carácter privativo, no representa una presencia, sino una omisión de claridad, distinción y atención. De manera análoga a la oscuridad, que no existe por sí misma, sino como ausencia de luz, el error constituye la falta de un conocimiento adecuado.

Lo anterior permite esclarecer su estrecha relación con la ignorancia. Lejos de ser siempre culpable, existe una ignorancia de carácter natural que resulta indisociable de la finitud humana. Al no disponer de un acceso simultáneo a toda la información ni de facultades plenamente desenvueltas desde el origen, el error emerge como el reflejo de nuestro estadio evolutivo. De ahí que Leibniz lo considere un indicador del grado presente de perfección del alma racional. Evidencia un estado incompleto de claridad que no exige una relación con respecto a la voluntad, sino al cultivo y perfeccionamiento gradual de las potencias cognitivas.

Aquí hay algo distintivo de Leibniz. El error no rompe la armonía del universo o la estructura perfecta del universo. Forma parte de una gradación de perfecciones. Las sustancias finitas reflejan el mundo con mayor o menor claridad, y esa claridad admite grados. Por eso, cuando se afirma que Dios no puede errar, no se está simplemente repitiendo una tesis teológica común. En Leibniz, esto significa que en Dios no hay grados, ni oscuridad, ni perspectiva parcial. Su conocimiento no

es progresivo ni limitado. Mientras que el ser humano conoce por grados de claridad, Dios conoce de manera absolutamente distinta. Así, el error no se explica como una culpa original ni como un mal introducido en el sistema, sino como la marca inevitable de una sustancia finita que participa en la armonía universal sin agotarla.

En resumen, a primera vista, la clasificación lockeana parece muy cercana a la que se encuentra en la respuesta de Leibniz. Sin embargo, la coincidencia es más aparente que real. Para Locke, el error surge fundamentalmente cuando se otorga asentimiento a proposiciones que no cuentan con evidencia suficiente. Se trata, ante todo, de una falla del juicio. Leibniz, por el contrario, interpreta esas mismas situaciones desde una perspectiva más amplia: el error no consiste únicamente en juzgar mal, sino en la inevitable limitación cognitiva de las sustancias finitas. Mientras Locke analiza el error como un problema epistemológico ligado al uso de las pruebas y la probabilidad, Leibniz lo entiende como una privación de claridad y perfección en el conocimiento.

2.6. Spinoza

Pasando a la última teoría del error a tratar en esta sección, se hablará de Spinoza, quien no solo ofrece aportes interesantes sobre este problema, sino que también permite esclarecer, en cierta medida, la evolución de las propuestas de Descartes y Leibniz. Su enfoque introduce un cambio importante en la forma de comprender la relación entre conocimiento, mente y naturaleza. Para comenzar, conviene situar a Spinoza en un plano general, pues su naturalismo resulta distintivo al sostener que los seres humanos no ocupan un estatus ontológico especial dentro de la realidad. Spinoza afirma que todo lo que ocurre en los seres humanos debe poder explicarse mediante las mismas leyes que rigen el resto de la naturaleza. En palabras de Bennett (1984), para Spinoza el ser humano no constituye "un dominio dentro de un dominio" (p. 41), sino una parte más del orden natural. En consecuencia, cualquier fenómeno humano, incluyendo las creencias, errores o pasiones, debe comprenderse del mismo modo que se explican otros procesos naturales, como el movimiento de los planetas o el crecimiento de las plantas.

En este marco, Spinoza desarrolla una teoría del conocimiento (y, por consiguiente, del error) que busca explicar estos fenómenos sin recurrir a entidades sobrenaturales ni a facultades excepcionales. Distingue tres formas de conocimiento: el conocimiento por experiencia vaga o imaginación, el conocimiento racional y el conocimiento intuitivo, siendo este último el más elevado por su certeza y claridad. Conocer adecuadamente consiste en captar las causas necesarias de las cosas; una idea verdadera no solo representa su objeto, sino que surge de una comprensión adecuada de su lugar dentro de la cadena causal. Desde esta perspectiva, el error no se concibe simplemente como discrepancia con la verdad, sino como el resultado de una comprensión parcial o confusa. Una idea inadecuada no refleja de manera clara ni necesaria las causas de aquello que representa. Por ello, la verdad no depende únicamente de una correspondencia externa entre la idea y la cosa, sino de la adecuación interna de la propia idea, es decir, de que esta exprese de manera completa y necesaria la causa de aquello que comprende.

Para entenderlo mejor, se puede señalar que Spinoza afirma que cada modificación del cuerpo tiene una idea correspondiente en la mente. Por ejemplo, cuando se observa el sol en el cielo, se tiende a imaginarlo como un objeto relativamente pequeño y cercano. Sin embargo, esta idea es inadecuada, pues no refleja las verdaderas relaciones causales que explican su tamaño y su distancia respecto de la Tierra. La mente posee una idea del Sol, pero esa idea surge únicamente de los efectos que produce en nuestro cuerpo y no de un conocimiento adecuado de sus causas.

Otro ejemplo aparece cuando Spinoza analiza la creencia humana en el libre albedrío. Los seres humanos suelen pensar que son libres porque son conscientes de sus deseos y decisiones, pero ignoran las causas que los determinan: "Los hombres se engañan porque creen ser libres; y esta opinión sólo consiste en que son conscientes de sus acciones, pero ignoran las causas que los determinan. Por lo tanto, la idea de su libertad consiste en no reconocer ninguna causa de sus acciones" (Spinoza, 1677/2020, 2p35s). En este caso, la mente posee una idea de su propia acción, pero no comprende el entramado causal que la produce. El error, entonces, no consiste en que la mente tenga una idea completamente falsa, sino en que posee una idea parcial que no logra situar su objeto dentro del orden necesario de la naturaleza.

Así pues, para Spinoza, "[…] ignorar y errar son cosas diversas" (1677/2020, 2p35d), aunque a primera vista puedan parecer equivalentes. El error no constituye una entidad positiva ni un estado especial del alma producido por una facultad particular. A diferencia de autores como Descartes o Locke, que explican el error a partir de un uso inadecuado del juicio o de la voluntad, Spinoza sostiene que este surge cuando la mente posee ideas inadecuadas. Se trata de ideas fragmentarias, confusas o incompletas porque no captan las causas que determinan aquello que representan. En este sentido, Bennett señala que el error puede entenderse como una forma particular de ignorancia. No se trata de una ignorancia absoluta, es decir, de la ausencia total de conocimiento, sino de poseer una idea parcial sin advertir su carácter incompleto. Se erra, no porque se carezca por completo de ideas, sino porque las ideas que se tiene no expresan adecuadamente el orden causal de las cosas. El error es, por tanto, una consecuencia natural de la posición limitada que ocupa la mente humana dentro del orden de la realidad. Esta concepción plantea una dificultad importante. Si, como afirma Spinoza en la demostración de 2p32, "Todas las ideas, en efecto, que son en Dios, concuerdan totalmente con lo ideado por ellas; por tanto, todas son verdaderas", entonces ¿cómo puede hablarse de ideas falsas? La respuesta de Spinoza, según Wilson, consiste en reconocer que las mentes humanas son limitadas y constituyen expresiones parciales de la infinitud divina: "Lo que Spinoza parece querer decir es que las ideas dan lugar a la falsedad cuando ocurren en mentes limitadas, en separación del orden causal total en el cual están con respecto de la mente divina" (Wilson, 1996, p. 109).

Regresemos al ejemplo de la persona que ve el sol como algo pequeño y cercano. Ciertamente posee una idea del Sol y, en tanto que esta idea es algo positivo, también está en Dios, tal como afirma la proposición 32 de la segunda parte (Spinoza, 1677/2020, 2p32). Si además tomamos en cuenta lo que Spinoza sostiene: "Nada hay de positivo en las ideas en virtud de lo cual se digan falsas" (2p33), entonces la falsedad no puede consistir en algo positivo presente en la idea misma. Debe consistir, más bien, en una ausencia. Y esa ausencia corresponde precisamente a las conexiones causales que permitirían comprender que el Sol está muy lejos y que, por ello, se percibe como pequeño. La mente finita toma únicamente una porción limitada del entramado total de relaciones y la presenta de forma aislada. La idea es entonces incompleta y, por ello, inadecuada.

El error puede entenderse, así, como una carencia. Nace de una deficiencia estructural de la idea: le faltan conexiones causales, claridad y comprensión de su lugar dentro del orden del pensamiento. En este punto puede apreciarse cómo la propuesta de Spinoza permite releer las teorías anteriores. Frente a Descartes, el error ya no se explica por una desproporción entre voluntad y entendimiento. Frente a Leibniz, tampoco se presenta simplemente como una limitación gradual del alma racional; pues en Leibniz la finitud de la mente implica grados de claridad en la percepción, de modo que el error aparece como el resultado de una aprehensión confusa del mundo. Percibimos, pero lo hacemos mal porque nuestra perspectiva es limitada. Spinoza, en

cambio, desplaza el problema hacia la estructura misma de las ideas: una idea es falsa no porque haya sido mal juzgada, sino porque expresa la realidad de manera parcial e incompleta. El punto no es solo que percibamos de manera parcial, sino que la propia estructura de las ideas no garantiza por sí misma su adecuación. No se trata únicamente de una diferencia de grado en la perfección del alma, sino de una diferencia en el modo de producción de las ideas dentro del orden de la naturaleza. El error deja de ser un acto del sujeto para convertirse en una consecuencia natural de la forma en que una mente finita participa del orden infinito de la naturaleza. Esto nos plantea un problema. Pues, si el error es estructural y no voluntario, ¿cómo es posible progresar en los grados de conocimiento? La respuesta de Spinoza es que no se supera el error mediante actos voluntarios de corrección, ni mediante un cambio externo en la voluntad, sino a través del aumento de la potencia de obrar y de pensar. En la medida en que se adquieren ideas más adecuadas y se comprenden con mayor claridad las causas que determinan las cosas, disminuye el ámbito de las ideas inadecuadas. El punto no es eliminar el error como si se tratara de una falla puntual del sujeto, sino reorganizar progresivamente el sistema de ideas de modo que cada vez dependa menos de relaciones parciales y más de conexiones causales completas.

Esto, sin embargo, abre una tensión que será importante más adelante: si el error es estructural y no simplemente episódico, entonces el aprendizaje no puede entenderse como una mera corrección puntual de fallos, sino como una transformación global del modo en que un sistema produce y conecta sus ideas. En ese sentido, la noción misma de "corrección" deja de ser un acto aislado y pasa a depender de procesos más amplios de reorganización del conocimiento.

2.7. Recapitulación a la luz de Rescher

Para cerrar esta sección, debemos volver a la propuesta de Nicholas Rescher. En general, distingue entre tres tipos de error definidos en función de aquello que el agente no logra alcanzar:

(1) El error cognitivo: Ocurre cuando no alcanzamos la verdad. Es el error más familiar y se presenta cuando nuestra intención es informarnos o responder una pregunta sobre la realidad, pero terminamos aceptando como verdadero algo que es falso (o viceversa). Es aquí donde resuenan las preocupaciones de Epicuro, San Agustín, Descartes, Leibniz y Locke sobre el juicio apresurado.

(2) El error práctico: Acontece cuando no logramos satisfacer un deseo o cumplir un propósito operativo. Aquí no fallamos necesariamente en lo que sabemos, sino en lo que hacemos; es el terreno de la ineficacia, donde la acción frustra el objetivo. Esta dimensión se liga directamente con Aristóteles y su análisis de la acción orientada a fines, donde el error puede darse tanto en la deliberación como en la elección de los medios para alcanzar el bien o la eficacia.

(3) El error axiológico: Sucede cuando no reflejamos ni apreciamos adecuadamente el valor de algo. Este es quizás el aporte más sutil de Rescher, pues ocurre cuando somos incapaces de asignar la prioridad o el valor correcto a las cosas en nuestras apreciaciones. En el plano ético, el ejemplo claro es quien valora más la gratificación momentánea de un impulso que la lealtad a un compromiso de años. El error reside en una «percepción errónea del valor» (Rescher, 2007, p. 14). Esta categoría dialoga estrechamente con San Agustín y todo el inconveniente del buscar sin encontrar la verdad. O con Spinoza y la diferencia que hay en el modo de producción de las ideas dentro del orden de la naturaleza.

Esta clasificación resulta útil porque demuestra que el error no se limita a un único dominio, sino que atraviesa distintas dimensiones de nuestra relación con el mundo. Como señala el propio Rescher, los errores requieren siempre de una intención, al menos de manera implícita, respecto a los objetivos que los agentes desean alcanzar. Para el autor, el error posee un componente estrictamente normativo: no erramos simplemente porque las cosas salieron mal, sino porque no cumplimos con un estándar que nosotros mismos, como seres racionales, reconocemos como válido.

De este modo, comprender el error implica reconocer tres elementos fundamentales: (i) una vulnerabilidad necesaria, pues el error solo es posible para un agente que asume riesgos; (ii) una intencionalidad, pero no de cualquier tipo, sino reflexiva, ya que no basta con tener una meta, sino entender que se debería alcanzar; y (iii) un mecanismo de retroalimentación (feedback), donde el error cognitivo advierte que nuestro mapa del mundo es incorrecto, el práctico que nuestras herramientas son inadecuadas, y el axiológico que nuestra escala de valores necesita calibración; mecanismo que, a su vez, ha de permitir la autorregulación (o aprendizaje o mejoramiento) para que el error no se repita o al menos se vaya disminuyendo periódicamente su ocurrencia. Si un sistema cualquiera (digamos, un humano) cumple con ser falible, con tener una intención adecuada y con tener un mecanismo de retroalimentación que le muestra dónde está el problema que le impide lograr su objetivo, pero sigue fallando una y otra vez, después de un tiempo empezaremos a pensar que hay algo mal en este sistema, y en vez de hablar de error, pensaremos que algo está "fallando" en sus capacidades cognitivas, de razonamiento o agenciales. Lo que esto muestra es que la capacidad de aprender del error y tomar pasos o medidas para evitarlo parece esencial también para que se diga que un cierto sistema "erra" en vez de que "falla". Y es la ausencia de esta capacidad la que invita a hablar de fallo en lugar de error.

Al mirar atrás, vemos que, aunque los diversos filósofos revisados difieren en sus explicaciones específicas, comparten una intuición común: errar no es simplemente producir un resultado numérico o fáctico incorrecto. se vincula a la acción, al juicio, a la finitud de la mente y, a la estimación del valor.

La propuesta de Rescher permite formular esta intuición con especial claridad en el horizonte actual. Si el error supone la frustración de una intención orientada a fines, entonces no todo fallo constituye un error. Una licuadora falla por desgaste, un televisor deja de funcionar después de años de uso y, hasta una aspiradora robot puede trabarse por absorber una media. Sin embargo, desde la perspectiva aquí trazada, ninguno de estos casos constituye todavía un error en sentido estricto, sino un mal funcionamiento. El error exige una implicación del agente con aquello que hace: la capacidad de asumir fines como propios y de reconocer normativamente la distancia entre lo que pretendía alcanzar y lo que efectivamente alcanzó. Y si aceptamos esta caracterización, se sigue una consecuencia crítica para el problema que nos ocupa: si las máquinas carecen de intencionalidad (reflexiva), entonces no pueden errar; únicamente pueden fallar. No obstante, esta conclusión se vuelve problemática si queremos sostener que ciertos sistemas artificiales contemporáneos realizan algo más que operaciones mecánicas ciegas. En este punto, la discusión ya no gira únicamente en torno al error, sino en torno a la naturaleza de la intencionalidad misma. Dicho de otro modo: si queremos defender la posibilidad de que una máquina yerre, primero debemos preguntarnos si puede ser considerada un agente intencional.

Por ello, la siguiente sección no se centrará directamente en el error, sino en aquello que parece constituir su condición de posibilidad. Si las teorías estudiadas ofrecen buenas razones para pensar que el error requiere alguna forma de agencia intencional, resulta obligatorio examinar si esta

característica puede atribuirse también a los sistemas artificiales. Solo una vez respondida esta cuestión podremos determinar si tiene sentido hablar, en sentido estricto, de si errar es una condición exclusivamente humana.

3. Hablar de máquinas como agentes: Turing y la objeción de Searle

Una postura interesante que parece estar en tensión con la postulación de Rescher es la de Alan Turing (1950) en su artículo *Computing Machinery and Intelligence*. Allí, además de afirmar que "podemos esperar que, con el tiempo, las máquinas compitan con los hombres en todos los ámbitos puramente intelectuales", sostiene la tesis de que la inteligencia no es una propiedad interna misteriosa, sino una capacidad observable a través del comportamiento y el lenguaje. Así, el artículo no busca responder de manera directa a la pregunta de si las máquinas pueden pensar, sino reformularla en términos operativos a través del llamado juego de la imitación[14]. De este modo, la cuestión deja de depender de definiciones problemáticas como pensamiento o inteligencia y se traslada al terreno de la ejecución. En este marco, Turing introduce una serie de objeciones clásicas[15] a la posibilidad de la cognición artificial para luego refutarlas metódicamente.

Con este trasfondo, resulta relevante la objeción que Alan Turing denomina precisamente "argumentos desde diversas incapacidades". Allí recoge una serie de afirmaciones recurrentes según las cuales las máquinas no podrían, por ejemplo:

> ser amables, ingeniosas, bellas, amables, tener iniciativa, sentido del humor, distinguir el bien del mal, **cometer errores**, enamorarse, disfrutar de fresas con crema, hacer que alguien se enamore de ella, aprender de la experiencia, usar las palabras correctamente, ser sujeto de su propio pensamiento, tener tanta diversidad de comportamientos como un hombre, hacer algo realmente nuevo. […][16] (Turing, 1950, P. 447, mi traducción y mi negrilla.)

Es en este punto donde Turing introduce una distinción que resulta especialmente útil para nuestro análisis. Parte de la idea de que la afirmación según la cual las máquinas no pueden equivocarse descansa en un malentendido. En el contexto del juego de la imitación, se asume que una máquina sería fácilmente distinguible de un humano porque resolvería problemas, por ejemplo, aritméticos, con una precisión absoluta y una rapidez sorprendente. Su exactitud la delataría. Por ello, Turing plantea, de manera deliberadamente simple e incluso algo irónica, que una máquina

[14] En él participan tres personas: un hombre (A), una mujer (B) y un interrogador (C), que puede ser de cualquier sexo. El interrogador permanece en una habitación separada de los otros dos. El objetivo del juego para el interrogador es determinar cuál de los otros dos es el hombre y cuál es la mujer. Los conoce por las etiquetas X e Y, y al final del juego dice «X es A y Y es B» o «X es B y Y es A». […] El objetivo de A en el juego es intentar que C cometa un error en la identificación. […] El objetivo del juego para el jugador (B) es ayudar al interrogador. […Al respecto, Turing nos cuestiona] "¿Qué pasará cuando una máquina asuma el papel de A en este juego? ¿Tomará el interrogador decisiones erróneas con la misma frecuencia cuando el juego se desarrolle de esta manera que cuando se juega entre un hombre y una mujer? Estas preguntas sustituyen a nuestra pregunta original: «¿Pueden pensar las máquinas?»" (1950, p. 433-434, traducción propia).

[15] Entre ellas se encuentran: la problemática de la reformulación del problema; la objeción teológica, que reserva el pensamiento al alma humana; la llamada "cabeza en la arena", que rechaza la idea por sus consecuencias indeseables; la objeción matemática, que apela a límites formales del cálculo; y el argumento de la conciencia, que exige experiencia subjetiva como condición del pensamiento.

[16] Turing no niega completamente estas dificultades, pero las desplaza: muestra que son distinciones de grado o dependen de supuestos discutibles.

diseñada para participar en el juego de la imitación no tendría por qué ofrecer siempre la respuesta correcta. Podría, en cambio, introducir errores de manera estratégica, ajustando su comportamiento con el fin de no ser distinguida de un agente humano. De igual manera, podría dilatar o demorar su respuesta para no quedar en evidencia por la rapidez de sus cálculos. En este sentido, la ausencia de error no sería una condición necesaria de la máquina, sino una consecuencia del modo en que ha sido diseñada. Sin embargo, su respuesta no debe entenderse como la simple idea de que el programador añade errores para hacerla parecer más humana. Más bien, lo que sugiere es que la relación entre máquina y error no es tan rígida como suele asumirse. Si una máquina puede ser programada para equivocarse de manera plausible, entonces la producción de resultados incorrectos no queda completamente excluida de su funcionamiento. En otras palabras, evitar el error puede ser, en ciertos casos, una exigencia más estricta que permitirlo.

Sin embargo, cuando Turing habla del error en relación con máquinas, no se limita a repetir estas nociones, sino que obliga a reconsiderarlas. Frente a quienes sostienen que las máquinas no pueden errar, su propuesta introduce un desplazamiento, pues no todo error es del mismo tipo. Por un lado, están los errores de funcionamiento, asociados a fallas mecánicas o eléctricas. Se ha de notar aquí que Turing no está siendo impreciso con el lenguaje aquí, todo lo contrario, busca aclarar la diferencia entre "error" y "fallo" que tanto interesa a esta investigación. Para ello, busca recoger las expresiones populares y aclararlas. Por ello, dice primero que se habla de "dos tipos de errores" ("it seems to me that this criticism depends on a confusion between two kinds of mistake") y procede a distinguirlos para mostrar, luego, que uno es simplemente un fallo y el otro propiamente un error ("We may call them "errors of functioning" and "errors of conclusion". Errors of functioning are due to some mechanical or electrical fault which causes the machine to behave otherwise than it was designed to do. […] Errors of conclusion can only arise when some meaning is attached to the output signals of the machine.") El primer tipo de "error", que se debe a un fallo (fault) mecánico o eléctrico parece entenderse sin demasiada dificultad. Tanto en máquinas como en humanos hablamos de fallar cuando una acción no produce el resultado esperado. Por ejemplo, cuando alguien ha estado sentado mucho tiempo en el escritorio trabajando y, cuando se va a levantar, una de sus piernas no responde y se caes al piso. La gramática misma con la que expresamos esto muestra ya que generamos una "distancia" con el suceso y hasta con su causa. Decimos que *la pierna* "no responde", hablando como si la pierna fuera un tercero, fuera diferente y aislada de nosotros. En pocas palabras, lo tratamos como algo que *nos pasa*, ante lo que somos meros pacientes, y no algo *que hacemos*, ente lo que seríamos agentes. Lo mismo ocurre cuando un computador deja de responder por un problema eléctrico o cuando un sistema arroja un resultado absurdo debido a un daño técnico. En ambos casos, el problema parece relativamente claro: algo en el funcionamiento falló.

Ahora bien, la dificultad aparece con el segundo tipo de error. Los que Turing llama, *errores de conclusión*, y que surgen en el propio proceso de cálculo o inferencia, aun cuando el sistema funciona correctamente desde el punto de vista técnico. Si lo pensamos en términos humanos, llegar a una conclusión errada no siempre implica un fallo en el razonamiento mismo. Puede ocurrir que, a partir de información incompleta o incorrecta, sigamos un proceso lógico adecuado y, aun así, lleguemos a un resultado equivocado. Por ejemplo, como Carl Sagan señala en su libro *El mundo y sus demonios* que los científicos de la NASA observaron nubes en Venus. Así que su razonamiento era que, si había nubes, había agua, porque las nubes son agua evaporada. Si había agua, podía haber vida. En realidad, años después se descubrió que las nubes eran de vapor de ácido sulfúrico y que toda vida en esas condiciones era del todo imposible. El procedimiento que llevó a cabo es correcto, pero el resultado no lo es. En este tipo de casos no dudamos en decir que

hubo un error, aunque no necesariamente en el razonamiento mismo, sino en las condiciones bajo las cuales este se realizó (una de las premisas iniciales era falsa).

En el caso de una máquina, es relativamente sencillo pensar la cuestión en los términos en que Turing la planteaba. Cuando habla de errores de conclusión, propone imaginar una máquina que devuelve ecuaciones matemáticas o frases en determinado idioma. Si una proposición falsa aparece como respuesta, podemos decir, siguiendo a Turing, que la máquina cometió un error de conclusión. Del mismo modo, podría extraer conclusiones a partir de un proceso de inducción científica y, mediante ese método, llegar a un resultado incorrecto. Hasta aquí resulta fácil comprender e intuir la línea de pensamiento que Turing estaba siguiendo. Sin embargo, la dificultad aumenta si consideramos un caso similar al de los humanos mencionado anteriormente. Si el sistema es alimentado con información incorrecta o sesgada, puede procesarla adecuadamente y, aun así, producir resultados sistemáticamente erróneos. ¿Seguiría siendo esto un error de conclusión o debería considerarse un error de funcionamiento? Parecería que no encaja del todo en ninguna de las dos categorías, aun cuando el resultado obtenido sea incorrecto.

El caso es aún más complejo cuando dejamos de pensar en licuadoras, televisores y maquinas diseñadas para engañar y pasamos a sistemas que aprenden de la "experiencia". Máquinas como aspiradoras inteligentes[17], robots de asistencia o colaboración social[18] o, inteligencias artificiales generativas[19], las cuales pueden cometer errores basados ya sea en fallos o en este escenario de ser alimentados por información incorrecta. Este tipo de resultados no se limitan a casos aislados, sino que puede reproducirse y consolidarse en el tiempo, en la medida en que esa información pasa a formar parte de sus criterios de decisión. Si esto ocurre, entonces parece insuficiente decir que la máquina simplemente "falló". Hay algo más complejo ocurriendo allí. Visto así, parece que Turing se queda corto al explicar los errores en máquinas que aprenden, pues existen ciertos errores que no dependen únicamente de daños físicos o fallas técnicas, sino también del modo en que un sistema interpreta y responde a la información que recibe[20]. En consecuencia, necesitamos un marco que vaya más allá del mero funcionamiento mecánico, aunque no necesariamente en un sentido cognitivo ni apelando a elementos exclusivamente humanos. Después de todo, como el propio Turing plantea, se está tratando con "varias discapacidades", y no es necesario incorporar la dimensión sensible o cognitiva humana para abordar este tipo de problemas. Al discutir la inteligencia y otros rasgos que suelen emplearse para afirmar o negar que una máquina los posea, basta con concentrarnos en aquello que efectivamente se manifiesta en su comportamiento.

En este punto, el eco de la intencionalidad reflexiva de Rescher parece volver a resonar. Aunque para Rescher "los dispositivos inertes -máquinas e instrumentos- no cometen errores; funcionan mal [they malfunction]" (Rescher, 2007, p. 3, mi traducción), los casos presentados anteriormente parecen situarse en una zona gris. Da la impresión de que van más allá del simple mal funcionamiento, pero tampoco podemos atribuirles intencionalidad solo porque su conducta

[17] Las cuales aprenden el plano de la casa, esquivan objetos y, mejoran su calidad de limpieza dependiendo de la superficie.

[18] Los cuales están constantemente aprendiendo del entorno y de las reacciones que el humano pueda tener. Ya sean emocionales o físicas.

[19] Que también aprenden en cada iteración, tienen "memoria" y si se les señala en que pueden estar equivocadas, lo implementan en su aprendizaje para futuras respuestas.

[20] En toda justicia, mal puede acusarse a Turing de negligencia o algo similar, si se tiene en cuenta que su artículo fue publicado en *Mind* en 1950 y que el avance de la computación, en ese entonces, era mínimo. Turing fue, a todas luces, un visionario que se adelantó a su época y concibió un potencial inmenso a la computación que, en su época, era apenas exigua.

parezca orientada hacia determinados resultados. Para abordar esta dificultad, será necesario ir más allá de la comprensión del sistema y de sus procesos de aprendizaje, y dirigir la atención hacia la manera en que se entiende la intencionalidad misma.

Siguiendo a Jorge Burgos (2026), la intencionalidad se utiliza para referirse a aquello que solemos llamar significado o contenido mental. En términos generales, se trata de la capacidad que tienen ciertos estados, eventos o símbolos de representar algo o de estar dirigidos hacia algo. Esta concepción se encuentra fuertemente influenciada por la tradición inaugurada por Brentano, para quien la intencionalidad, el hecho de que un estado mental sea siempre acerca de algo, constituye una de las características distintivas de lo mental. Desde esta perspectiva, los pensamientos humanos poseen una direccionalidad o referencia intrínseca que los sistemas artificiales no tendrían en sentido propio.

Gran parte de las posiciones tradicionales sobre la mente parten precisamente de esta idea para sostener que las máquinas no pueden poseer intencionalidad en sentido estricto. Una de las críticas más influyentes en esta dirección es la formulada por John Searle mediante el argumento de la Habitación China, concebido como una respuesta a las pretensiones más fuertes derivadas del test de Turing. En este experimento mental, Searle imagina que un humano encerrado en una habitación siguiendo un libro de instrucciones para responder a caracteres extraños (él no sabe que son, en realidad, caracteres chinos) introducidos por debajo de una puerta. Aunque no comprende una sola palabra de chino, el libro le indica que ante X carácter recibido en el papel, ha de escribir Y carácter, y así sucesivamente. Sin saberlo, y gracias al libro, esos dibujos que para él no tienen sentido, corresponden a respuestas adecuadas a preguntas que le están siendo formuladas en chino en los papeles que se introducen en la habitación. Este ser humano está manipulando símbolos de acuerdo con reglas preestablecidas, exactamente del mismo modo en que lo haría una computadora. Para quienes observan desde afuera, sus respuestas serían indistinguibles de las de un hablante competente del idioma chino. Sin embargo, sostiene Searle, es evidente que el hombre dentro de la habitación no tiene comprensión alguna ni del idioma ni de lo que se le pregunta o de lo que él está respondiendo.

La conclusión del argumento es que ejecutar correctamente un programa puede producir la apariencia de comprensión, pero no comprensión real. Las computadoras operan mediante reglas sintácticas que les permiten manipular símbolos, aunque carecen de acceso al significado de esos símbolos. En consecuencia, el test de Turing resultaría insuficiente como criterio para atribuir inteligencia o comprensión, ya que evalúa únicamente el comportamiento observable. Más aún, Searle considera que este argumento cuestiona la idea de que la mente humana pueda entenderse simplemente como un sistema computacional de procesamiento de información. Para él, la intencionalidad no surge de la mera manipulación formal de símbolos, sino que depende de las capacidades causales propias de los sistemas biológicos. Como afirma el propio Searle:

> ¿Podría pensar una máquina?" Mi opinión es que solo una máquina podía pensar, y de hecho solo tipos muy especiales de máquinas, es decir, cerebros y máquinas que tenían los mismos poderes causales que los cerebros. Y esa es la razón principal por la que una IA fuerte nos ha contado poco sobre el pensamiento, ya que no tiene nada que decirnos sobre las máquinas. Por su propia definición, se trata de programas, y los programas no son máquinas. Sea lo que sea la intencionalidad, es un fenómeno biológico, y es tan probable que dependa causalmente de la bioquímica específica de sus orígenes como la lactancia, la fotosíntesis o cualquier otro fenómeno biológico. (Searle, *Minds, Brains and Programs*. p. 424)

Con este ejemplo, Searle intenta mostrar que una máquina puede producir respuestas indistinguibles de las de un hablante competente sin comprender realmente lo que hace. El sistema puede seguir reglas, reconocer patrones y generar respuestas adecuadas, pero ello no implica que los símbolos con los que opera posean significado para él. La crítica apunta, en última instancia, a la diferencia entre sintaxis y semántica: una máquina puede procesar correctamente estructuras formales sin comprender su contenido. Y si no existe comprensión genuina, entonces tampoco existiría intencionalidad en el sentido clásico del término.

Así, la posición estándar termina sosteniendo que las máquinas solo simulan comprensión. Parecen responder, interpretar o incluso equivocarse, pero no habría en ellas estados intencionales genuinos. Desde esta perspectiva, hablar de error en sentido fuerte sigue siendo problemático, porque el error estaría ligado a creencias, comprensión y contenido mental.

Sin embargo, si aceptamos esta conclusión, parece que muchos de los problemas planteados por los sistemas contemporáneos quedan sin explicación. Por ello, vale la pena preguntarse: ¿debe la intencionalidad entenderse necesariamente en términos biológicos e internos?

Una respuesta que puede ofrecérsele a Searle yace en decir que el hombre dentro de la habitación no sabe chino ni comprende las preguntas y respuestas en las que está involucrado, pero que este enfoque es inadecuado, pues es el *sistema entero* el que se ha de evaluar como competente y comprensivo del chino o no. El sistema incluye el libro de instrucciones, al ser humano, los papeles, etc.; tal vez el humano no sepa chino, pero el conjunto entero sí parece saberlo. De manera análoga, es como si se dijera que la corteza occipital de mi cerebro participa del proceso de visión, pero que ella misma no "ve" nada. Para ella esos impulsos eléctricos y activaciones neuronales no son, en realidad, representaciones de árboles, casas o animales. Son solo impulsos eléctricos. Esto es inadecuado porque toma sólo una parte ínfima del cerebro y le exige ser una suerte de homúnculo con comprensión total del proceso, algo que ciertamente no está diseñada para hacer. Pero si se habla de la cabeza entera, o incluso del agente entero, la historia es muy diferente. Es justamente por ello que, hablando de sistemas funcionalistas de la mente, Hilary Putnam propone una serie de condiciones:

> (1) Todos los organismos capaces de sentir dolor son Autómatas Probabilistas.
>
> (2) Todo organismo capaz de sentir dolor posee por lo menos una Descripción de un cierto tipo (esto es, ser capaz de sentir dolor es poseer una clase especial de Organización Funcional).
>
> (3) Ningún organismo capaz de sentir dolor puede descomponerse en partes que posean separadamente Descripciones del tipo de las mencionadas en (2).
>
> (4) Para cada Descripción de la clase mencionada en (2), existe un subconjunto de entradas sensoriales tal que un organismo con dicha Descripción siente dolor cuando, y sólo cuando, algunas de sus entradas sensoriales están en ese subconjunto. (Putnam, *La naturaleza de los estados mentales*, p.5)

Nos aleja de nuestro objetivo entrar en las discusiones acerca de qué es un autómata funcional y otros detalles acerca de la propuesta funcionalista de Putnam. Lo esencial en este momento es ver que la condición 3 está estableciendo, justamente, que es el sistema completo el que ha de tener un cierto estado mental (sea este sentir dolor, o saber chino), y no una parte de él. De esta manera, se hace evidente que el planteamiento, por parte de Searle de la objeción de la habitación china

está poniendo el foco en una parte del sistema, y le está exigiendo -inadecuadamente- comportarse como el todo.

Sin embargo, tener alguna manera de responder a Searle esta objeción no implica, por sí solo, que se ha solucionado ya el tema de la posibilidad de la intencionalidad en sistemas artificiales. Esta sigue siendo una tarea aun por enfrentar y para eso Daniel Dennett parece tener la llave perfecta. La Actitud Intencional.

4. La actitud intencional

4.1. El problema:

En resumen, siguiendo la línea que se venía trabajando, errar conlleva los tres rasgos mencionados por Rescher y sustentados por los filósofos revisados anteriormente: (1) una vulnerabilidad necesaria; (2) una intencionalidad reflexiva; y (3) un mecanismo de retroalimentación. Respecto de la primera característica, no parece haber mayor dificultad, pues no solo se trata de un rasgo básico y común que comparte con el fallo, sino que además se sitúa en el nivel del accionar. En cambio, la característica (2) presenta un problema distinto, ya que remite a actitudes proposicionales.

Según Fodor (1978), las actitudes proposicionales son estados mentales que pueden comprenderse como relaciones entre un organismo y determinadas representaciones internas. Esto significa que no se trata simplemente de estados aislados, sino de contenidos que adoptan la forma de una oración subordinada del tipo "S cree que P", "S desea que P" o "S espera que P", donde "P" funciona como una proposición completa. En este sentido, uno de sus rasgos centrales es su carácter opaco: el valor de verdad de la actitud no depende únicamente del referente de "P", sino de la manera en que dicho contenido es representado por el sujeto, lo que impide sustituir términos dentro de la proposición sin alterar potencialmente el estado mental atribuido.

Por su parte, Churchland (1999) sostiene que las actitudes proposicionales reciben este nombre porque expresan una actitud frente a una proposición. Ejemplos de ello son "el *pensamiento* de que (los niños son maravillosos), la *creencia* de que (los seres humanos tienen grandes potencialidades) o, el *temor* de que (la civilización pueda caer en otra Edad Oscura)[21]" (Materia y conciencia, p. 101). En cuanto a su carácter intencional, este no debe entenderse como una acción deliberada del sujeto, sino como la propiedad de estar dirigido hacia algo distinto de sí mismo. Así, dichas actitudes remiten a objetos, estados de cosas o situaciones del mundo, como los niños, los seres humanos o la civilización.

En términos generales, puede observarse que las actitudes proposicionales son intencionales en el sentido de Brentano[22] y se caracterizan por dos rasgos fundamentales. En primer lugar, (i) son

[21] Lo que se encuentra dentro de los paréntesis se puede intercambiar por P, donde P, es una proposición.

[22] Brentano defiende que todos los fenómenos psicológicos, y sólo los fenómenos psicológicos, son intencionados. Sostiene que creer, es creer algo. Así afirma que:

> Todo fenómeno mental incluye algo como objeto dentro de sí mismo, aunque no lo hagan de la misma manera. En la presentación, algo se presenta, en el juicio algo se afirma o se niega, en el amor se ama, en el odio se odia, en el deseo es deseado, y así sucesivamente.
> Esta inexistencia [se ha de tener en cuenta que Brentano toma el término de los escolásticos del medioevo, así que "inexistencia" no significa "no-existencia" sino "existencia-en"] intencionada es característica exclusivamente de los fenómenos mentales. Ningún fenómeno físico muestra algo parecido. Por tanto, podemos definir los fenómenos mentales diciendo que son aquellos fenómenos que contienen un objeto intencionadamente dentro de sí mismos. (*Psicología desde un punto de vista empírico,* 1874, 88-89)

estados mentales que involucran una oración subordinada, como en expresiones del tipo "María cree que P", "María sabe que P" o "María espera que P", donde "P" corresponde a una proposición completa, por ejemplo: "Colombia clasifique a la segunda ronda del mundial". En segundo lugar, (ii) son *opacas* (lo que suele asociarse con la denominada condición de Frege). Esto significa que, aun cuando dos expresiones refieran al mismo objeto, reemplazar una por otra dentro de una actitud proposicional puede alterar el valor de verdad de la oración (es decir, no son sustituibles *salva veritate*). Por ejemplo, aunque "el Lucero de la mañana" y "el Lucero de la tarde" designen a Venus, del hecho de que María crea que el Lucero de la mañana es visible al amanecer no se sigue necesariamente que crea que el Lucero de la tarde es visible al amanecer, pues puede ser el caso que María no sepa que el Lucero de la mañana y el lucero de la tarde sean efectivamente el mismo objeto.

Esto hace que se vea como problemático hacer compatibles a las actitudes proposicionales con una visión científica y objetiva del mundo. El mismo Churchland lo dice:

> A veces se ha mencionado la intencionalidad de estas actitudes proposicionales como el rasgo decisivo que distingue lo mental de lo meramente físico, como algo que no puede manifestar ningún estado puramente físico. En parte esta afirmación podría ser correcta en el sentido de que la manipulación racional de las actitudes proposicionales efectivamente puede ser el rasgo distintivo de la inteligencia consciente. Pero, aunque la intencionalidad haya sido mencionada con frecuencia como "la marca de lo mental", esto no constituye necesariamente una presunción en favor de alguna forma de dualismo. (Churchland, 1999, 102)

Para Erasmus Mayr el problema es muy similar. Toda agencia, para ser verdadera agencia y no mero mecanismo ciego, ha de involucrar actitudes proposicionales y el gran problema está en cómo hacer compatible este rasgo de la agencia, con la explicación científica del mundo:

> "Comencemos con las siguientes tres afirmaciones acerca de la agencia:
>
> Tesis uno: Las acciones son instancias de actividad. El agente es activo respecto de lo que está haciendo y no solamente un ente pasivo.
>
> Tesis dos: Las acciones son fenómenos naturales, es decir, parte del orden natural.
>
> Tesis tres: Las acciones, en tanto que son intencionales, pueden ser explicadas citando las razones por las cuales fueron realizadas. […]
>
> Abandonar cualquiera de las tesis uno, dos o tres parece absurdo. Pues ello implicaría, o bien que las acciones no son en absoluto diferentes de las cosas que sufrimos -tales como dolores de muela-, o que nosotros estamos por fuera del orden natural, o que nunca actuamos por razones; y ninguna de estas posibilidades es algo que podamos considerar seriamente en tanto que nos entendamos como agentes en absoluto." (Mayr, 2011, 6)

Por supuesto, si Mayr habla de "abandonar" alguna de las tesis, lo hace porque sabe que las tres están en fuerte tensión entre sí. En particular, para el caso que aquí interesa, hay una fuerte tensión entre la tesis dos y la tesis tres. La tensión se debe a que el orden natural del que habla la tesis dos se explica por medio de las leyes naturales. Estas leyes son leyes causales, que postulan el intercambio, injerencia o transferencia de fuerza o energía entre cuerpos de acuerdo con ciertas fórmulas ya establecidas. La tesis tres, en cambio, habla de una causalidad muy diferente, a saber, la causalidad por razones. Las razones funcionan como causas que guían las acciones de un individuo. Si las razones no son un mero epifenómeno, han de tener poder causal en la conducta

del individuo *qua razones*. Pero, ¿cómo ofrecer una explicación causal compatible con el modelo científico que a la vez atienda a la opacidad de las razones que han de fungir como causa? Fodor establece como quinta condición de una teoría de las actitudes proposicionales el que "Una teoría de las actitudes proposicionales debe interrelacionarse con las explicaciones empíricas de los procesos mentales" (1978, p.508) Así, el problema está en cómo hacer compatible la intencionalidad, con las explicaciones científicas del mundo, cuando la intencionalidad parece ser el rasgo distintivo de lo mental e irreductible a lo físico. El mismo Fodor no lo ve como una tarea imposible, pero sí de entrada mucho más complicada de lo que podría imaginarse o se ha imaginado hasta el momento: "Creo, de hecho, que el requisito de que una teoría de las actitudes proposicionales deba ser empíricamente plausible puede significar una cantidad abrumadora de trabajo; mucho más trabajo del que los filósofos usualmente han pensado." (1978, p.508)

4.2. La propuesta de Dennett

La dificultad no es menor, pero no por ello es infranqueable. Al igual que ocurre con muchas de las estrategias adoptadas a lo largo de la historia de la filosofía, el énfasis puede desplazarse para comprender mejor aquello que se está estudiando. En este sentido, Daniel Dennett propone un giro radical en la estrategia tradicional: invierte la forma en que el pensamiento y la mente han sido abordados históricamente, una tradición que puede rastrearse, al menos, hasta Descartes. De hecho, el propio Dennett señala que: "La filosofía usualmente no produce 'resultados' estables y confiables, así como la ciencia lo hace. Puede, sin embargo, producir nuevas formas de ver las cosas, formas de enmarcar las preguntas, de ver qué es importante y por qué" (IS, p. 2).

Así, desde la tradición filosófica dominante, el estudio de la mente suele comenzar por la introspección, por la conciencia. Pero de inmediato aparece una barrera. La conciencia es subjetiva, privada y de acceso a la primera persona únicamente. Cualquier intento de entender la mente de una manera compatible con la ciencia (que explica las cosas desde el punto objetivo de la tercera persona) está condenado al fracaso si el punto de partida es subjetivo, privado y de primera persona. Por eso Dennett parte de un estudio de tercera persona de lo mental, separando el asunto de la mente del asunto de la conciencia. No hay que entender la conciencia primero para poder entender la mente, sino que hay que entender la mente para luego poder dar una explicación de la conciencia. Este es el giro que propone Dennett, como bien lo explica Tadeusz Zawidzki:

> Todo el proyecto de Dennett comienza invirtiendo la prioridad tradicional de la conciencia sobre el pensamiento. Intenta entender qué es el pensamiento y qué significa tener un pensamiento particular, independientemente de entender la conciencia. Luego intenta explicar la conciencia y la autoconciencia como tipos especiales de pensamiento. Según Dennett, esto es crucial para el proyecto de reconciliar la imagen manifiesta y la científica de las personas. Dado que la ciencia se centra en información objetiva y en tercera persona, cualquier estrategia para entender el pensamiento y la mente que comience con la conciencia, que supuestamente es un ámbito de información subjetiva en primera persona, pone el pensamiento y la mente fuera del alcance de la ciencia. Dennett comienza intentando comprender el pensamiento en el proyecto de Dennett en el contexto de términos objetivos en tercera persona que sean susceptibles de estudio científico. Luego intenta comprender la conciencia, la subjetividad y la perspectiva en primera persona como una especie de pensamiento, que en última instancia también es susceptible de estudio

> científico. Esta estrategia es fundamental para comprender todo el proyecto filosófico de Dennett. (Zawidzki, 2007, 9, mi traducción)

Así, Dennett propone que los distintos sistemas (sean humanos, eventos naturales, máquinas artificiales, órganos biológicos, animales no humanos, etc.) pueden entenderse utilizando una estrategia predictiva y explicativa que él denomina "la actitud intencional", esta constituye la versión más compleja y, en muchos sentidos, la más contraintuitiva de la propuesta de Dennett. Para llegar a ella, primero es necesario comprender las tres estrategias explicativas que, según él, pueden adoptarse para predecir y explicar el comportamiento de un sistema.

La primera es la actitud física. Desde esta perspectiva, explicamos el comportamiento de algo atendiendo a sus componentes materiales y a las leyes físicas que los gobiernan. Por ejemplo, un cocinero puede anticipar que, si deja una olla sobre el fogón durante demasiado tiempo, el agua se evaporará y la comida terminará quemándose. Sin embargo, esta estrategia parece quedarse corta en ciertos casos. Pensemos en una máquina que juega ajedrez: aunque una descripción física puede explicar qué ocurre en sus componentes electrónicos, no parece suficiente para explicar por qué el sistema movió el caballo y no la torre en una jugada determinada. Esto implicaría que hay que asumir una actitud diferente de la física. Como señala Dennett, que toda explicación debe reducirse, en última instancia, a la actitud física constituye una suerte de dogma cientificista más que una exigencia metodológica efectiva.

Es así como aparece la segunda estrategia: la actitud de diseño. En este caso, ya no es necesario examinar cada uno de los componentes materiales del sistema. Basta con comprender para qué fue diseñado y cómo debería funcionar bajo condiciones normales. Si una licuadora deja de funcionar, por ejemplo, solemos preguntarnos qué parte del mecanismo falló respecto de la función para la que fue diseñada, sin necesidad de reconstruir toda la física involucrada en su operación. Esta estrategia resulta mucho más eficiente que la anterior, y mucho más útil para sistemas más complejos que el de una olla con alimento dentro hirviendo, o mecanismos sencillos. Pero tampoco parece suficiente en casos de aun mayor complejidad; si volvemos al ejemplo de la máquina de ajedrez, saber que fue diseñada para jugar y ganar partidas no explica adecuadamente por qué realizó una jugada específica en una situación concreta del tablero. Supongamos que la máquina puede, en una cierta jugada, capturar a la reina de su oponente, pero no lo hace, sino que, con el caballo, elimina un alfil. Sabe que la máquina fue diseñada para jugar y ganar en el juego de ajedrez no permite entender esta movida. Se requiere aun de un tipo de actitud diferente, más robusta y compleja que permita entender esto. El diseño general del sistema permite comprender su función, pero no necesariamente sus decisiones particulares.

De este modo llegamos a la tercera estrategia explicativa: la actitud[23] intencional. Es en este nivel donde Dennett propone que, para comprender el comportamiento de ciertos sistemas, resulta útil tratarlos como si fueran agentes dotados de racionalidad, quienes poseen creencias, deseos y objetivos. El propio Dennett la define de la siguiente manera:

> Primero usted decide tratar el objeto cuyo comportamiento debe predecirse como un agente racional; luego descubre qué creencias debería tener ese agente, dado su lugar en el mundo y su propósito. Luego determina qué deseos debería tener, basándose en las mismas

[23] Dennett utiliza el término "Intentional Stance", y he decidido traducir "stance" por "actitud", siguiendo la definición de la RAE de este término: "Actitud: Disposición de ánimo manifestada de algún modo". (RAE). También puede traducirse como instancia o postura. Lo esencial es comprender que se asume una cierta actitud/postura frente a la comprensión de un sistema.

> consideraciones, y finalmente predice que ese agente racional actuará para avanzar en sus objetivos a la luz de sus creencias. Un poco de razonamiento práctico a partir del conjunto elegido de creencias y deseos en muchos — pero no todos — casos, dará lugar a una decisión sobre lo que el agente debe hacer; eso es lo que predice que hará el agente. (Dennett, IS, p.17, mi traducción).

Así, Dennett defiende que la actitud intencional puede aplicarse no solo a los seres humanos, sino también a otros animales e incluso a ciertos artefactos. Uno de sus ejemplos favoritos es el de una máquina que juega ajedrez. Tomemos, como ejemplo, a la máquina real DeepBlue.

Desde la actitud intencional, no intentamos explicar su comportamiento recurriendo a sus componentes físicos ni a los detalles de su programación. En cambio, la tratamos como un agente racional. y nos preguntamos ¿Qué desea? ganar la partida, evitar la derrota, conservar piezas valiosas o alcanzar posiciones favorables. ¿En qué cree? En que dar jaque mate conduce a la victoria, que perder piezas importantes disminuye sus posibilidades de ganar o que determinadas posiciones del tablero son más ventajosas que otras. Una vez realizadas estas atribuciones, el comportamiento de la máquina se vuelve comprensible y predecible. Regresemos al ejemplo de la jugada en la que se podría haber capturado a la reina, pero se optó por eliminar un alfil. Podemos ahora comprender este movimiento diciendo que la jugada de capturar a la reina había revelado la estrategia que tenía la máquina y, entre sus creencias, la máquina tiene aquella según la cual aumentan las posibilidades de ganar si el oponente no logra descifrar la estrategia contra la cual se enfrenta. Podría tener también una creencia acerca del valor o priorización de ciertas cosas, como, por ejemplo, que mantener secreta su estrategia es más importante que capturar a la reina, y otras similares. De esta manera, asumir la actitud intencional frente a DeepBlue permite comprender sus movimientos (¿por qué sacrificó una pieza si se supone que debe mantener la mayoría de las piezas?) a la luz de pensarlo como un agente racional que tiene un objetivo, unos deseos y una serie de creencias y que actúa con base en ellos.

Hasta este punto puede resultar evidente que existen importantes similitudes entre la actitud de diseño y la actitud intencional. De hecho, ambas comparten al menos tres características fundamentales. En primer lugar, son estrategias explicativa y predictivamente más eficientes que la actitud física. En lugar de requerir una descripción de los componentes materiales de un sistema y de las leyes que los gobiernan, permiten generar predicciones acertadas a partir de un conjunto mucho más reducido de supuestos. Esto las convierte en herramientas especialmente útiles cuando una descripción física completa es impracticable o innecesaria. En segundo lugar, incorporan un elemento normativo. Ambas parten de la idea de que el sistema *debe* comportarse de determinada manera: en el caso de la actitud de diseño, de acuerdo con la función para la cual fue diseñado; en el caso de la actitud intencional, de acuerdo con lo que un agente racional debería creer o hacer dadas sus metas y la información disponible. Las explicaciones y predicciones se construyen precisamente a partir de estas expectativas normativas. Y finalmente, ambas son falibles. Dado que sus predicciones dependen del supuesto de que el sistema funciona correctamente, siempre existe la posibilidad de error. Un artefacto puede estar averiado, del mismo modo que un agente puede actuar de manera irracional o disponer de información equivocada. Por esta razón, aunque estas estrategias suelen ser extraordinariamente útiles, también implican un margen de riesgo que no puede eliminarse por completo.

Las diferencias entre la actitud de diseño y la actitud intencional vienen dadas en que la intencional tiene una suposición normativa aún más fuerte y, por ende, es aún más riesgosa o falible:

> La actitud intencional plantea una suposición normativa aún más estricta que la postura de diseño: se asume que los sistemas intencionales se comportan no solo como están diseñados, sino de la manera más racional posible. Como consecuencia, la postura intencionada es aún más arriesgada que la postura de diseño." (Zawidzki, 2007, 37, traducción mía)

Esto, lejos de ser un problema de la teoría, puede ser tal vez uno de sus puntos fuertes. Como se verá más adelante, la actitud intencional puede permanecer neutral frente a cuál sea el mecanismo físico exacto que ha dado lugar a un cierto comportamiento. No se compromete, por ejemplo, en el caso del ser humano, con una teoría neurológica o científica específica acerca del funcionamiento del cerebro. Si hay comportamiento que se ven afectados, al menos en parte, por este tipo de funcionamiento físico, es de esperar que las predicciones hechas desde la actitud intencional fallen en algunas ocasiones. Pero este es un precio pequeño por pagar para evitar comprometerse con una teoría reduccionista y asumir, con ella, sus enormes problemas.[24]

4.3. ¿Solución o vía sin salida?

Recordando que nuestro principal problema es que errar conlleva intencionalidad y aprendizaje, y que ambas características parecen corresponder a actitudes proposicionales, surge una dificultad importante. Si esto es así, entonces dependemos de una explicación de la conciencia que la filosofía, hasta ahora, no ha logrado desarrollar siguiendo un "camino seguro de la ciencia". Es precisamente por eso que el giro propuesto por Dennett nos ayuda a superar este problema. La aplicación de esta estrategia permite abordar el problema desde una perspectiva distinta, y el propio Dennett lo deja ver en *Brainstorms* cuando dice:

> El concepto de sistema intencional es una noción relativamente sencilla y poco metafísica, abstraída como está de cuestiones sobre la composición, constitución, conciencia, moralidad o divinidad de las entidades que lo engloban. Así, por ejemplo, es mucho más fácil decidir si una máquina puede ser un sistema intencional que decidir si una máquina puede realmente pensar, o ser consciente, o moralmente responsable. Esta simplicidad la hace ideal como fuente de orden y organización en los análisis filosóficos de conceptos "mentales". Sea lo que sea una persona—mente o alma encarnada, agente moral autoconsciente, forma "emergente" de inteligencia—es un sistema intencional, y todo lo que se deriva solo de ser un sistema intencional es así cierto para una persona. (Dennett, BS, 16, mi traducción)

Esto nos recuerda al giro que realiza Turing frente a la pregunta por la inteligencia de las máquinas. Responderla de manera directa resulta complejo, pues parece que, como señala Searle, se necesita algo más que sintaxis: cerebros y estados mentales humanos. Sin embargo, cuando la discusión se

[24] Zawidzki lo explica haciendo referencia a una distinción entre una aproximación causal y una definicional: "Hay una diferencia, por ejemplo, entre preguntar qué *hace* [makes] a alguien ser un marido, y preguntar *cómo llega* [comes to be] alguien a convertirse en un marido. La primera es una pregunta de *definición*. Lo que hace a alguien un marido es, al menos en parte, que es reconocido como tal por las autoridades apropiadas. […] La segunda pregunta es una de *causación*: ¿cómo llega una persona particular a convertirse en un marido, esto es, ¿qué causó que se casara con alguien y, por lo tanto, se convirtiera en marido? Esta pregunta tiene tantas respuestas como hay maridos. Algunos son causados a ser maridos al enamorarse, otros por su deseo de tener el estatus de residente en un cierto país y otros por padres iracundos empuñando una escopeta." (p. 45) Dennett está del lado de las definiciones y no de las causaciones.

traslada al terreno práctico, la demostración parece darse por sí misma. En lugar de preguntarnos qué es exactamente la inteligencia, conciencia, y demás, más bien observamos qué hace el sistema. Si una máquina se comporta de maneras que normalmente asociamos con agentes inteligentes, entonces la cuestión deja de ser qué ocurre en su interior y pasa a ser por qué no deberíamos considerarla inteligente, buena, consciente, etc.[25]

Además, adoptar esta posición no entra en conflicto con lo que ya sabemos gracias a otras ciencias. La actitud intencional no niega las explicaciones ofrecidas por disciplinas como la neurociencia, la biología o la física, opera en un nivel distinto. Así, por ejemplo, la neurociencia puede explicar cómo determinados procesos neuronales intervienen en la toma de decisiones, mientras que la actitud intencional permite comprender esas mismas decisiones atribuyendo al agente ciertas creencias, deseos y objetivos. Ambas explicaciones pueden coexistir sin excluirse mutuamente. Del mismo modo, tampoco afecta la forma en que nos relacionamos con otros seres. De hecho, la utilizamos constantemente para comprender, predecir y responder al comportamiento de quienes nos rodean, ya sean amigos, vecinos, familiares o incluso animales. Rara vez explicamos las acciones de los demás apelando a procesos físicos o de diseño, normalmente interpretamos lo que hacen atribuyéndoles razones, expectativas, creencias o deseos.

Dennett ejemplifica esto con una rana en el bosque. Pues cuando nos acercamos a ella y esta salta para alejarse, pensamos "está huyendo". No analizamos los procesos físicos que produjeron la contracción de sus músculos ni intentamos explicar su comportamiento a partir de una supuesta finalidad para la que fue diseñada (sea por un creador, una naturaleza sabia o como se quiera). En cambio, recurrimos de manera casi automática a la actitud intencional. Tratamos a la rana como un agente racional dentro de sus limitaciones y le atribuimos ciertas creencias y deseos que permiten explicar su conducta. Desde esta perspectiva, suponemos que la rana percibe un peligro, que "cree" que puede ser capturada o devorada, que desea preservar su vida y que, por tanto, debe escapar. Gracias a estas atribuciones, su comportamiento resulta comprensible y predecible: la rana salta porque, dadas las circunstancias y los objetivos que le atribuimos, esa parece ser la acción más razonable. La utilidad de la explicación no depende de demostrar que la rana posea exactamente esos estados mentales, sino de que tales atribuciones permiten entender y anticipar exitosamente su comportamiento.

Zawidzki lo dice aún más claro cuando afirma:

> Si algo es o no un sistema intencional, según Dennett, es un asunto perfectamente objetivo y de tercera persona. Además, el concepto de sistema intencional de Dennett no está

[25] Este giro es una estrategia utilizada en filosofía en varios frentes. Es famoso, por ejemplo, el caso de Peter F. Strawson quien en *Freedom and Resentment* y enfrentado a la pregunta acerca de si la eventual verdad del determinismo elimina o no la responsabilidad moral, decide hacer un giro en toda la discusión. Hasta entonces, se empezaba por el *being* (es decir, las condiciones que alguien ha de cumplir para ser moralmente responsable) y luego se miraba si, dadas esas condiciones se podía *considerar* a alguien como moralmente responsable (lo que se conoce como el *holding*). El debate estaba estancado por siglos, porque entre las condiciones se encontraba el libre albedrío, el control autónomo y otras tantas que son sumamente difíciles de dilucidar. Strawson decide invertir la ecuación del being al holding, y la considera mejor del holding al being. Así, Strawson comienza por preguntarse cómo es que tradicionalmente consideramos (hold) a las personas como moralmente responsables y en qué circunstancias dejamos de considerarlas así. Luego se pregunta si la eventual verdad del determinismo universalizaría o no esas circunstancias en las que no consideramos a alguien moralmente responsable (la universalización de la enfermedad mental, del accidente, de la ignorancia, etc.) y ve que no, no se universalizan, de manera que se ubica del bando de los que creen que la eventual verdad del determinismo no atenta contra la responsabilidad moral.

> relacionado con otros conceptos científicamente problemáticos de la misma manera que lo está el concepto manifiesto de creencia. Para determinar si un sistema cuenta como un sistema intencional, en el sentido de Dennett, no necesitamos determinar si es consciente ni de qué es consciente. No necesitamos determinar si es una persona o no. Lo único que importa es que, de hecho, la mejor manera de entenderlo implica adoptar una postura intencional hacia él. Los seres humanos, los ordenadores y al menos algunos animales, según esta visión, cuentan como sistemas intencionales porque no se puede negar que la mejor a menudo la única forma de entender su comportamiento es adoptando la postura intencional. (Zawidzki, 2007, 37, mi traducción)

En síntesis, lo importante de la propuesta de Dennett, es que nos ayuda a salir del problema inicialmente descrito, pues desplaza la discusión de la pregunta por la naturaleza interna de los estados mentales y la conciencia, hacia la pregunta por su utilidad en la explicación. Desde esta perspectiva, no es necesario resolver primero o del todo los problemas asociados a la experiencia en primera persona para poder atribuir creencias, deseos e intenciones. Basta con que tales atribuciones constituyan la mejor estrategia disponible para comprender y predecir el comportamiento de un sistema.

4.4. Explicación tipo cascada

Es así como sabemos que aplicar la actitud intencional no resuelve automáticamente todos los problemas. Lejos de ello, Dennett sostiene que este tipo de explicaciones constituyen lo que denomina un "préstamo de inteligencia" ("loan of intelligence"). Cuando describimos el comportamiento de un sistema atribuyéndole creencias, deseos, intenciones o representaciones, estamos optando por una estrategia metodológica que dista de ser una explicación exhaustiva Lo explica así en *Brainstorms:*

> Cada vez que un constructor de teorías propone llamar a cualquier evento, estado, estructura, etc., en cualquier sistema (por ejemplo, el cerebro de un organismo) señal, mensaje o comando o le otorga contenido, toma un préstamo de inteligencia. Implícitamente postula, junto con sus señales, mensajes o órdenes, algo que puede servir como lector de señales, comprensor de mensajes o comandante, de lo contrario sus "señales" serán en vano, se desintegrarán sin recibir o sin comprender. Este préstamo debe pagarse eventualmente encontrando y analizando a estos lectores o comprensores; porque, si esto no funciona, la teoría tendrá entre sus elementos análogos humanos no analizados dotados de suficiente inteligencia para leer las señales, etc.… y por tanto la teoría pospondrá responder a la pregunta principal: ¿qué a la inteligencia? La intencionalidad de todo este tipo de discusiones sobre señales y órdenes nos recuerda que la racionalidad se está dando por sentada, y de este modo nos muestra dónde una teoría está incompleta. Es esta característica la que, en mi opinión, otorga un valor a la tarea aún inconclusa de idear una definición rigurosa de intencionalidad, pues si podemos reclamar un criterio puramente formal de discurso intencional, tendremos lo que equivale a un medio de intercambio para evaluar teorías del comportamiento. La intencionalidad abstrae de los detalles no-esenciales de las diversas formas que pueden adoptar los préstamos de inteligencia (por ejemplo, lectores de señales, emisores de voluntad, bibliotecarios en los pasillos de la memoria, egos y superyós) y sirve como un medio fiable para detectar exactamente dónde una teoría está en números rojos respecto a la tarea de explicar la inteligencia; dondequiera

> que una teoría se apoya en una formulación que lleva las marcas lógicas de la intencionalidad, allí se oculta un hombrecito." (Dennett, BS, 12, mi traducción)

Cada vez que una teoría habla de señales, mensajes, órdenes, representaciones o información, parece requerir también la existencia de algo capaz de interpretar esas señales, comprender esos mensajes o ejecutar esas órdenes. Pensemos en el teatro cartesiano como explicación del funcionamiento de la mente. Ese gran teatro supone, a su vez, a un cierto homúnculo sentado en una de las sillas viendo lo que se proyecta en esa gran pantalla. Este homúnculo, se asume, tiene la capacidad de ver y entender lo proyectado, y surge la pregunta de cómo es que lo puede hacer. No se gana nada -en términos explicativos- si a su vez se piensa la mente del homúnculo como un gran teatro en el cual a su vez hay un homúnculo más pequeño… y así *ad infinitum*. Si la teoría no puede explicar cómo surge ese intérprete y simplemente lo presupone, entonces la deuda permanece impaga. La explicación funciona, pero solo porque ha introducido de manera implícita una instancia de inteligencia cuya existencia aún no ha sido justificada. Por ello, la cuestión no consiste en determinar si la intencionalidad atribuida a un agente es original o "pura"[26]. Lo importante es reconocer que toda atribución intencional implica un préstamo pendiente.

Por esto, la actitud intencional permite comprender y predecir exitosamente el comportamiento de un sistema, pero por sí sola no explica cómo es posible que dicho sistema llegue a comportarse de esa manera. La explicación completa exige mostrar de dónde surge la racionalidad que inicialmente hemos supuesto. Es por esta razón que Dennett presenta las tres actitudes en una estructura jerárquica. En la base se encuentra la actitud física, sobre ella, la actitud de diseño y, en la punta, la actitud intencional. Estas no son explicaciones incompatibles, sino niveles complementarios, con una no es suficiente explicar en su totalidad a ciertos sistemas complejos, como lo son humanos, computadoras y ranas. Así, cuando adoptamos la actitud intencional estamos tomando un préstamo de inteligencia, llamado racionalidad (que se les atribuye a los agentes) y, que eventualmente debe ser pagado mostrando cómo las regularidades atribuidas al agente pueden entenderse desde la perspectiva del diseño y, en última instancia, desde la perspectiva física[27]. En consecuencia, una explicación intencional adecuada no elimina los niveles inferiores, sino que descansa sobre ellos.

O en palabras de Dennett, cuando ofrecemos una explicación intencional, ya que esto es una actitud proposicional, tenemos un estado que tiene una frase subordinada. Asumimos que hay un *intérprete* de esa frase, y ese intérprete es como un pequeño *homúnculo* en el sistema. Este *homúnculo* debe ser explicado, o de lo contrario la explicación es circular e inútil: "La psicología sin homúnculos es imposible. Pero la psicología con homúnculos está condenada a la circularidad o a la regresión infinita." (Dennett, BS, 122)

La manera de explicar estos *homúnculos* o *intérpretes* es analizándoles en términos de cosas cada vez más sencillas y más "tontas", hasta quedar con una descripción física que ya no encierra misterios:

> Los homúnculos solo son monstruos si duplican todos los talentos que se les llama para explicar. Si uno puede conseguir que un equipo o comité de homúnculos relativamente ignorantes, de mente cerrada y ciegos, produzca el comportamiento inteligente de todos, esto es progreso. Un diagrama de flujo suele ser el organigrama de un comité de homúnculos (investigadores, bibliotecarios, contables, ejecutivos); cada caja especifica un

[26] Volveremos a este problema sobre intencionalidad original y derivada, en la sección 4.6.
[27] Que es lejos de ser un reduccionismo con pasos adicionales, pero también volveremos a ello más adelante.

> homúnculo prescribiendo una función sin decir cómo debe realizarse (se dice, en efecto: pon un hombrecito ahí para hacer el trabajo). Si luego miramos más de cerca las cajas individuales, vemos que la función de cada una se cumple subdividiéndola mediante otro diagrama de flujo en homúnculos aún más pequeños y tontos. Al final, este anidamiento de cajas dentro de cajas lleva a homúnculos tan tontos (que solo tienen que recordar si decir sí o no cuando se les pregunta) que pueden ser, como se dice, "reemplazados por una máquina". Uno libera homúnculos sofisticados de su plan organizando ejércitos de esos idiotas para hacer el trabajo. (Dennett, BS, 123-124, mi traducción)

La propuesta de Dennett consiste en convertir lo que tradicionalmente se consideraba un problema en una estrategia explicativa. El error no está en postular un intérprete o un homúnculo, pues toda explicación intencional parece requerir alguno. El verdadero problema aparece cuando ese intérprete es tan complejo como aquello que se pretende explicar. Si para explicar una mente inteligente postulamos otra mente igualmente inteligente en su interior, entonces no hemos explicado nada; simplemente hemos desplazado el problema un nivel más abajo. Por esta razón, Dennett propone analizar cada homúnculo en términos de funciones cada vez más simples. Cada uno de estos subsistemas puede seguir siendo descrito de manera intencional, pero con atribuciones cada vez más limitadas y especializadas.

La clave es que la complejidad del sistema no debe encontrarse en ninguna de sus partes consideradas aisladamente. Lo que llamamos inteligencia emerge de la organización y coordinación de múltiples procesos relativamente simples. Así, en lugar de explicar una mente inteligente apelando a otra mente inteligente, la explicación procede mostrando cómo una gran cantidad de mecanismos limitados y "tontos" pueden, al interactuar, producir comportamientos que desde un nivel superior parecen inteligentes.

Este proceder no es un invento de Dennett ni es nada extravagante, es, justamente, el proceder de la ciencia computacional y la ciencia cognitiva. En áreas computacionales usualmente se empieza por una descripción de alto nivel de aquello que el sistema debe hacer. Parten de una caracterización que guarda importantes semejanzas con la actitud intencional.

Así, por ejemplo, en ciencia cognitiva es común comenzar describiendo el sistema en términos de representaciones, reglas o decisiones, aunque luego se intente explicar cómo esas propiedades emergen de mecanismos de nivel inferior. Los modelos conexionistas y emergentistas muestran precisamente este movimiento: partir de una descripción funcional de alto nivel y después reconstruirla como el resultado de interacciones distribuidas entre unidades simples, sin asumir necesariamente que existan símbolos explícitos en el nivel físico del sistema.

Lo explica de manera más sencilla y claramente Zawidzki:

> La visión de Dennett sobre cómo la postura intencional se relaciona con otras formas de describir sistemas en la explicación del comportamiento inteligente está inspirada en metodologías estándar de informática y ciencia cognitiva. Cuando los programadores desarrollan software, comienzan con una descripción de postura intencional y de alto nivel de la tarea que quieren que el ordenador realice. Por ejemplo, conciben un programa de ajedrez como objetivo de ganar partidas de ajedrez. Luego descienden a lo que algunos llaman el 'nivel algorítmico': elaboran instrucciones, ejecutables por componentes de ordenadores que, en conjunto, pueden aproximarse a la competencia en el ajedrez. Finalmente, desarrollan formas de implementar estas instrucciones en el hardware físico

> de ordenadores reales. La ciencia cognitiva suele emplear una metodología similar. (Zawidzki, 2007, 43)

La importancia de este punto radica en que las explicaciones intencionales no son reemplazadas por las explicaciones de diseño o por las explicaciones físicas. Más bien, cada nivel cumple una función distinta dentro de una misma cadena explicativa. Los niveles superiores permiten comprender qué hace un sistema y por qué lo hace y, los niveles inferiores muestran cómo es posible que lo haga.

Por ello, el objetivo de Dennett es mostrar que las atribuciones intencionales pueden ser científicamente respetables siempre que formen parte de una estructura explicativa más amplia, capaz de conectarlas con mecanismos de diseño y, en última instancia, con procesos físicos. En este sentido, la actitud intencional no constituye una alternativa a la ciencia, sino una herramienta indispensable para describir fenómenos complejos antes de proceder a su análisis en niveles más fundamentales.

4.5. Naturalista, sí; Reduccionista, no.

Con todo lo anterior en mente, podría parecer que Dennett está defendiendo una forma de reduccionismo fuerte. Después de todo, si toda atribución intencional constituye un préstamo de inteligencia que eventualmente debe ser pagado, entonces parecería que la deuda solo puede saldarse descendiendo hasta una explicación completamente física. Sin embargo, esta interpretación resulta equívoca. La propuesta de Dennett no consiste en reducir las explicaciones intencionales a una teoría física particular, ni en afirmar que actualmente carecemos de una explicación adecuada únicamente por limitaciones tecnológicas. Por el contrario, una de las principales virtudes de la actitud intencional es precisamente su neutralidad respecto de cuál será la explicación definitiva en los niveles inferiores de la cascada explicativa. Lo dice claramente Zawidzki:

> La fortaleza de la comprensión de Dennett sobre la creencia y otras actitudes proposicionales en términos de la postura intencional es que ofrece una neutralidad bienvenida respecto de los desarrollos empíricos en las ciencias cognitivas. La postura intencional explica lo que significa ser creyente [*what it is to be a believer*]; no toma una postura sobre cómo el cerebro humano alcanza este estatus. Este tipo de neutralidad es bienvenida porque, como resultado, nuestro estatus como creyentes no depende de las tendencias empíricas de la ciencia. Sea lo que sea que la ciencia descubra sobre el cerebro, según Dennett, seguiremos siendo sistemas intencionales y, por tanto, creyentes. Esto se debe a que, pase lo que pase, seguirá siendo cierto que el comportamiento humano es fiable y en gran medida desde la actitud intencional. (Zawidzki, 2007, 44)

Es así como el estatus de un agente como sistema intencional no se ve amenazado cada vez que cambian nuestras teorías científicas. Podemos descubrir nuevos mecanismos neuronales, nuevas arquitecturas computacionales o incluso principios explicativos completamente distintos de los actualmente conocidos, y aun así seguiría siendo correcto describir a los sistemas como agentes que *creen, desean, esperan* o *temen*. Esto se debe a que la validez de la actitud intencional no depende de una hipótesis concreta sobre la constitución física de los sistemas, sino de su capacidad para explicar y predecir su comportamiento.

Por ello Dennett no es ingenuo en las complicaciones que hay sobre lo mental. Sabe muy bien de los riesgos que se corren si se adopta una visión, por ejemplo, materialista reductiva de lo mental. Una de las críticas más influyentes a este tipo de propuestas fue formulada por Hilary Putnam cuando criticó el materialismo en *The Nature of Mental States*. Allí discute con quienes creen que se puede reducir el *dolor* (o el estado mental de dolor) a un cierto estado cerebral, a lo cual responde:

> Consideremos lo que el teórico de los estados cerebrales tiene que hacer para dar validez a sus afirmaciones. Tiene que especificar un estado fisicoquímico tal que *cualquier* organismo (no solamente un mamífero) siente dolor si y sólo si (a) posee un cerebro con una estructura fisicoquímica adecuada y (b) su cerebro está en ese estado fisicoquímico. Esto significa que el estado fisicoquímico en cuestión tiene que ser un estado posible de un cerebro de mamífero, de un cerebro de reptil, de un cerebro de molusco (los pulpos son moluscos y ciertamente sienten dolor), etcétera. Al mismo tiempo, tiene que ser un estado que *no* sea posible (físicamente posible) para el cerebro de ninguna creatura físicamente posible que no pueda sentir dolor. Aun cuando pueda encontrarse un estado semejante, tiene que ser nomológicamente cierto que será también un estado del cerebro de cualquier vida extraterrestre que pudiera descubrirse y que fuese capaz de sentir dolor, antes de que podamos siquiera mantener la suposición de que este estado *sea* el dolor. (Putnam, 1967 [reimpreso en Lycan, 1990, 31])

Dennett es consciente de esta enorme dificultad y no se va a casar con una teoría que sea tan demandante metafísicamente. Su propuesta no consiste en identificar las creencias, los deseos o las intenciones con estados físicos específicos, sino en caracterizarlos a partir del papel que desempeñan. La pregunta relevante deja de ser qué estructura material realiza una creencia y pasa a ser: bajo qué condiciones resulta apropiado atribuir creencias a un sistema. Así, su interpretación desde la actitud intencional deja la puerta abierta a que haya intencionalidad realizada de distintas maneras, en algunos organismos puede ser biológica y en otros de orden no-biológico o artificial. El mismo lo expresa de la siguiente manera:

> Las capacidades cognitivas de los animales que no utilizan el lenguaje […] también deben tenerse en cuenta, y no solo en términos de una analogía con las prácticas de quienes sí lo utilizamos. El nivel macro hasta el cual debemos relacionar los microprocesos cerebrales para comprenderlos como psicológicos es, en un sentido más amplio, el nivel de interacción, desarrollo y evolución del organismo y su entorno. Este nivel incluye la interacción social como una parte particularmente importante, pero, aun así, una parte propia. No hay forma de capturar las propiedades semánticas de las cosas (palabras, diagramas, impulsos nerviosos, estados cerebrales) mediante una micro reducción. Las propiedades semánticas no son solo relacionales, sino, podríamos decir, superrelacionales, ya que la relación que un vehículo de contenido particular, o token, debe tener para poseer contenido no es solo una relación con otras cosas […] sino una relación entre el token y la vida entera —y la vida contrafactual— del organismo al que «sirve» y los requisitos de supervivencia de ese organismo, así como su ascendencia evolutiva. (Dennett, IS, 65, mi traducción)

4.6. Intencionalidad original Vs. derivada

Entendiendo esto así, podemos por fin hablar de un tema central en la teoría de la actitud intencional de Dennett: la distinción entre intencionalidad original e intencionalidad derivada. Dennett la introduce explícitamente en el capítulo *8. Evolution, Error and Intentionality*, cuando pregunta: "¿Son la intencionalidad original y la intencionalidad intrínseca lo mismo?" (Dennett, IS, 289)

Buena parte de la filosofía de la mente contemporánea asumió que existe una gran diferencia entre aquellos sistemas que poseen significado o contenido por sí mismos y aquellos que solo lo poseen porque nosotros se lo atribuimos. Esta es precisamente la posición defendida, de maneras distintas, por autores como Searle y Fodor.

Recordemos que para Searle los estados mentales humanos poseen una *intencionalidad intrínseca u original*. Mis creencias, deseos y percepciones se refieren al mundo por sí mismas, no necesitan que alguien más las interprete para que tengan contenido. En cambio, las palabras de un libro, los símbolos de un computador o las señales de tránsito poseen únicamente una *intencionalidad derivada*. Significan algo porque los seres humanos les otorgamos ese significado. Un computador puede manipular símbolos, pero no comprende aquello que los símbolos representan[28]. De la misma manera, Dennett menciona que Fodor comparte una preocupación similar, aunque tenga una perspectiva distinta. Si las creencias y pensamientos tienen contenido, este contenido debe ser original y no simplemente prestado por un observador externo. Es así como surgió el desafío con el que iniciamos esta sección.

Dennett afirma que:

> […] discrepamos cuando afirmo aplicar exactamente la misma moral, las mismas reglas pragmáticas de interpretación, al caso humano. En el caso de los seres humanos (al menos), Fodor y compañía están seguros de que tales hechos más profundos existen, aunque no siempre podamos encontrarlos. Es decir, suponen que, independientemente de la capacidad de cualquier observador o intérprete para descubrirlo, siempre hay un hecho fundamental sobre lo que una persona (o su estado mental) realmente quiere decir. Ahora bien, podríamos llamar a esta creencia compartida una creencia en la intencionalidad intrínseca, o quizás incluso en la intencionalidad objetiva o real. (Dennett, IS, p. 294)

Dennett sospecha que toda esta manera de plantear el problema está equivocada. El error consiste en asumir desde el comienzo que debe existir algún tipo especial de intencionalidad original que sirva como fundamento último de todas las demás formas de significado.

Alejándose de esta perspectiva, para Dennett la pregunta relevante no es dónde aparece la intencionalidad auténtica, sino bajo qué condiciones resulta útil y justificado atribuir estados intencionales a un sistema y ahí es donde entra la actitud intencional. Dennett mismo sabe que su propuesta no va en línea con una larga tradición de filósofos (y seguramente del sentido común también) que llevan creyendo por siglos que el ser humano es único y que su mente ha de ser la cuna de la intencionalidad original; por ello suele recurrir a ejemplos que intenta mostrar, de una manera pedagógica, lo difícil que resulta localizar esa supuesta intencionalidad original y lo aparentemente fácil que resultaría reconocernos como sistemas con intencionalidad derivada. Tal vez su ejemplo más ilustrativo ocurre cuando pide que imaginemos el siguiente escenario:

[28] Esta es justamente la intuición que motiva experimentos mentales como el de la habitación china.

> Suponga que usted decide, por las razones que sean, que desea experimentar cómo es la vida en el siglo XXV; y suponga que la única manera conocida de mantener su cuerpo vivo durante ese largo tiempo es poniéndolo en un aparato de hibernación, donde descansará, desacelerado y comatoso, por el tiempo que usted desee. Usted puede hacer los arreglos necesarios para introducirse en esta cápsula de sostenimiento, ser puesto a dormir, y luego ser despertado automáticamente en el año 2401. (Dennett, IS, 295, traducción mía)

Cumplir con este deseo es mucho más complejo que sencillamente encontrar una cápsula tal e introducirse en ella. La cápsula y su diseño son ya un reto tecnológico gigante, pero más allá de ello, se requiere de poder ubicar la cápsula en un lugar en donde ya se haya provisto que contará con la energía, la protección a las inclemencias del clima y los nutrientes con los que ha de alimentar a su cuerpo por cerca de cuatrocientos años. Usted no puede simplemente confiar en que sus nietos y tataranietos cuidarán de la cápsula. ¿Qué hacer entonces? Dennett dice que hay dos estrategias posibles. La primera de ellas es escoger una cierta ubicación y "plantar" allí tanto la cápsula como los nutrientes, los generadores de energía y todo lo demás que se necesite para que el proyecto funcione por cuatro siglos. El gran problema de esta estrategia es que este "campamento" no se puede mover en caso de que haya algún problema. Suponga que, con el tiempo, aparece una falla geológica en ese punto que amenaza con abrir una grieta que destruirá el campamento. O, de manera menos trágica, que muchos años en el futuro el alcalde local desea construir una autopista en esa zona y pasa con sus máquinas de demolición por ese lugar. Esta, como estrategia, parece bastante susceptible a fallar. Pero queda aún otra opción, descrita por Dennett de la siguiente manera:

> La segunda alternativa es mucho más sofisticada, pero evita el gran inconveniente de la primera: diseñe una instalación móvil que albergue a la cápsula con los sensores y dispositivos de alerta temprana necesarios para que pueda desplazarse lejos del peligro y pueda buscar nuevas fuentes de energía cuando lo necesite. En breve, construya un robot gigante e instale la cápsula (con usted dentro) en él. (Dennett, IS, 296, traducción mía)

Esta segunda estrategia es en extremo compleja, pero no por ello imposible. Aparte del diseño de la cápsula, el diseño del robot presenta un nivel elevadísimo de complejidad. Ya que usted estará en un estado de coma, no puede estar atento al funcionamiento del robot ni guiar sus acciones, así que ha de dotarlo con una programación y unas herramientas que le permitan navegar su ambiente, detectar peligros, tomar ciertas decisiones acerca de dónde asentarse, cuándo ha de moverse, cómo explotar fuentes de energía, etc.:

> Usted ha de diseñarlo de manera tal que sepa generar sus propios planes en respuesta a circunstancias cambiantes. Debe "saber" cómo "buscar" y "reconocer" y luego explotar fuentes de energía, cómo moverse a un territorio más seguro y cómo "anticipar" y luego evitar peligros. (IS, 296, traducción mía)

Como si el reto no fuera lo suficientemente complejo ya, Dennett pide que suponga que este robot puede no estar solo en su empresa. Puede que haya otros robots, diseñados por personas como usted, que también quieren saber cómo es vivir en el siglo XXV, y que es probable que, en un futuro, si escasean los recursos y los nutrientes, haya una cierta competencia o rivalidad entre ellos. Así, Dennett sugiere que al diseñarlo tenga en cuenta esta posibilidad y dé a su robot otras capacidades adicionales:

> Sin duda sería prudente diseñarlo con la suficiente sofisticación en su sistema de control para permitirle calcular los riesgos y beneficios de cooperar con otros robots, o de formar alianzas para el beneficio mutuo. […]
>
> El resultado de este proyecto de diseño sería un robot capaz de exhibir autocontrol, ya que usted debe ceder el control en tiempo real a su artefacto una vez se disponga a dormir. Así, será capaz de derivar sus propias metas subsidiarias de sus evaluaciones de su estado actual y el peso de su fin último (que es preservar su vida). Estas metas secundarias podrían llevarlo a terrenos lejanos en proyectos que duren cientos de años. (IS, 297, traducción mía)

Nada de esto parece descabellado. Al ver los sistemas de inteligencia artificial disponibles hoy en día es perfectamente viable ver que se les puede dar una tarea como su meta última, y pedirles que diseñen estrategias para afrontar situaciones inesperadas y que tomen decisiones acerca de qué hacer respetando su fin último y tal vez algunas otras precauciones que se le asignen. Pero, ¿dónde se ubicarían Fodor y Searle frente a este ejemplo que propone Dennett?

> Sin embargo, de acuerdo con Fodor et al., este robot no tendría intencionalidad original en absoluto, sino sólo la intencionalidad que deriva de usted en su papel artefactual como su proyector. Su simulacro de estados mentales sería justamente eso -no habría decisiones reales, ni visión real, ni cuestionamiento ni planeación reales- sino sólo *como si* decidiera, viera, se preguntará y planeara. (Dennett, IS, 297, traducción mía).

Pero ¿en qué se basa esta distinción? ¿Cómo defender que esta intencionalidad es simplemente derivada? El robot fue diseñado por un humano, es verdad, pero fue hecho de manera tal que pudiera percibir su ambiente, que pudiera juzgar las situaciones, evaluarlas y tomar sus propias decisiones ante circunstancias imprevistas. Para cuidar su meta principal puede diseñar planes que tengan metas subsidiarias y embarcarse en esos planes por décadas. Se asume que ha de tener alguna capacidad de evaluar sus propios estados para repararse o saber qué necesita y que ha de poder aprender de sus experiencias para no embarcarse repetidas veces en un plan que no da los resultados esperados. Pero Dennett no se embarca en una defensa de la intencionalidad original de este robot. De manera sorpresiva, da un giro a la historia y, con él, a su estrategia argumentativa (o, más bien, revela por fin su verdadera estrategia argumentativa):

> Nuestra propia intencionalidad es exactamente como la del robot, pues la historia de ciencia ficción que acabo de contar no es nueva; sólo es una variación de la visión que tiene Dawkins de nosotros (y de otras especies biológicas) como "máquinas de supervivencia" diseñadas para prolongar el futuro de nuestros genes egoístas. Somos, en efecto, artefactos diseñados a través de los eones como máquinas de supervivencia para genes que no pueden actuar rápida e informadamente en favor de sus propios intereses. […] ¡Así, nuestra intencionalidad es derivada de la intencionalidad de nuestros genes "egoístas"! (Dennett, IS, 298, traducción mía).

Como se puede ver, la estrategia de Dennett está lejos de comprar la distinción de Fodor y Searle de que los seres humanos tenemos intencionalidad original y los demás artefactos sólo derivada. Su estrategia es, en cambio, defender que nuestra propia intencionalidad es también derivada, y, en este sentido, no sería diferente de la del robot. Esta es una brillante jugada porque voltea la carga de la prueba. Si Fodor y Searle desean defender que los seres humanos tenemos intencionalidad original han de poder mostrarlo de alguna manera diferente a la mera conducta de

los seres humanos. Porque en este ejemplo de Dennett, la conducta de esta máquina y de un humano son virtualmente indistinguibles.

El problema parece ser más claro si se expone el punto de vista de la realización múltiple (multiple feasibility). Si dos sistemas se comportan de manera funcionalmente equivalente, pueden ser comprendidos exitosamente desde la actitud intencional y responden de manera semejante a su entorno, parece arbitrario afirmar que uno posee contenido genuino mientras que el otro solo posee contenido derivado. Después de todo, desde la actitud intencional ambos sistemas exhiben exactamente aquello que nos lleva a atribuirles creencias, deseos y objetivos. Si uno de ellos merece ser considerado un sistema intencional, resulta cada vez más difícil explicar por qué el otro no. Explicar esta diferencia exige entonces señalar una propiedad adicional que posea el sistema con intencionalidad original y de la que carezca el sistema con intencionalidad derivada. Sin embargo, este es precisamente el punto donde Dennett considera que la posición de Searle y Fodor se vuelve problemática. Cuanto más complejos son los sistemas que imaginamos, más difícil resulta identificar qué es aquello que falta. Si el robot percibe su entorno, almacena información, modifica su comportamiento, desarrolla estrategias de supervivencia y puede ser interpretado exitosamente desde la actitud intencional, la afirmación de que carece de intencionalidad auténtica comienza a parecer una mera estipulación. En otras palabras, se afirma que existe una diferencia profunda, pero resulta cada vez más complicado mostrar en qué consiste exactamente esa diferencia.

Dennett no está intentando demostrar que los sistemas artificiales poseen intencionalidad original, más bien está cuestionando la utilidad misma de esa distinción. Si nosotros mismos podemos ser comprendidos como el resultado de un largo proceso de diseño evolutivo, entonces la exigencia de una fuente especial de significado intrínseco comienza a perder fuerza. Lo que llamamos intencionalidad no sería una propiedad misteriosa o mágica que aparece de repente en ciertos sistemas, sino una característica que emerge en organismos y artefactos suficientemente complejos para ser comprendidos desde la actitud intencional.

Por ello, la diferencia entre una intencionalidad supuestamente original y una derivada deja de ser el centro de la discusión. Lo verdaderamente importante pasa a ser qué sistemas admiten una interpretación intencional exitosa y por qué dicha interpretación funciona. Desde esta perspectiva, los seres humanos no ocupan una posición especial o única dentro del mundo natural, sino que constituyen un caso más bien sofisticado de los mismos procesos que permiten hablar de objetivos, creencias y racionalidad en otros sistemas.

5. Ejemplo

Pensemos en dos personajes, Iris y Josh, una pareja que lleva junta poco más de dos años. Se conocieron en un supermercado y deciden pasar el fin de semana en la casa de campo de uno de sus amigos. Iris es insegura, piensa constantemente en su relación y está dispuesta a hacer prácticamente cualquier cosa por Josh. Busca su aprobación, teme decepcionarlo y parece preocuparse por el futuro de la relación. Como cualquier otra persona, interpreta las acciones de quienes la rodean, formula expectativas sobre lo que ocurrirá y actúa con base en ellas. Nada en su comportamiento inicial nos lleva a pensar que se trata de algo distinto a un ser humano.

Todo esto ocurre en la película *Compañera Perfecta* [*Companion*] (2025), cuyo giro principal consiste en revelar que Iris es, en realidad, un robot de compañía. Un robot que mata, huye, se desespera y miente. Aunque la película muestra que esto sucede porque Josh modificó ilegalmente parte de su código, ello no implica que simplemente le hubiera descargado un hipotético software de "conciencia v2.0". Lo que altera son ciertos parámetros de su comportamiento, como la inteligencia, el nivel de agresividad, la capacidad de respuesta emocional y otros rasgos similares. A partir de estas modificaciones, y especialmente cuando descubre su propia condición de robot, Iris adquiere una nueva comprensión de la situación en la que se encuentra.

Antes de que se revele que Iris no es humana, sino un producto tecnológico, y que todos sus recuerdos forman parte de la narrativa implantada por la empresa fabricante (como, por ejemplo, cómo conoció a Josh, su graduación y gran parte de su historia personal), interpretamos sus acciones de una manera bastante común. La vemos como un ser humano más, y no dudamos, por un segundo, que sea un ser con vida mental, intencionalidad, creencias y deseos. Cuando se enfrenta a un peligro grave, estamos convencidos de que mata al hombre que intenta abusar de ella porque se siente amenazada, porque se encuentra incómoda en un entorno hostil y porque desconfía de quienes la rodean. En otras palabras, explicamos su conducta apelando a sus creencias, deseos e intenciones. Esto muestra que nuestra atribución de una vida mental, de creencias, deseos, anhelos, etc. A otros se basa en su conducta, como bien lo estableció Ryle en *The Concept of Mind* en 1949. Claro, cuando se nos revela que no es humana sino un robot, parecemos trastabillar un poco, pero, de acuerdo con Dennett, ello se debería más a un prejuicio nuestro, de acuerdo con el cual -igual que Fodor y Searle- creemos que los únicos con intencionalidad originaria somo los humanos y las máquinas no pueden tenerla. Pero ya hemos visto que no es tan fácil defender esta diferencia. Más aun, y si nunca se hubiera revelado que Iris es una máquina -ni siquiera a Josh-, ¿qué hubiera pasado? Josh se habría casado con ella, seguramente habrían formado una vida juntos, convivido muchos años y si Josh muriera joven en un accidente ¿podríamos defender que su relación fue una gran mentira? ¿Qué nunca nadie fue se compañera fiel, que no lo quiso, que no cuidó de él, que fue engañado por años? No parece ser un asunto tan claro de dilucidar.

Esto plantea una pregunta interesante: ¿los deseos, creencias e intenciones que atribuimos a Iris pertenecen realmente a ella o pertenecen a quienes la diseñaron y programaron? ¿Son producto de

la empresa que la fabricó o de Josh, quien modificó su configuración? Por un lado, parecería que debemos seguir una posición cercana a la defendida por Kripke cuando afirma:

> ¿Cómo se determina cuando ocurre un fallo? [...] Dependiendo de la intención del diseñador, cualquier fenómeno particular puede o no contar como un fallo de la máquina [...]. Si alguna vez una máquina falla y, de ser así, cuando, no es una propiedad de la máquina en sí como objeto físico, sino que es bien definido solo en función de su programa, según lo estipulado por su diseñador. (Kripke, Wittgenstein on Rules and Private Language, p. 33-34)

A esto Dennett nos confronta con:

> Aquí, creo, encontramos una expresión tan poderosa y directa como podría ser de la intuición que hay detrás de la creencia en la intencionalidad original. […]. Algo tiene que ceder. O debes abandonar el racionalismo del sentido —la idea de que eres diferente al gallo incipiente no solo por tener acceso, sino también por tener acceso privilegiado a tus significados— o debes abandonar el naturalismo que insiste en que, al fin y al cabo, eres solo un producto de la selección natural, cuya intencionalidad es por tanto derivada y por tanto potencialmente indeterminada. (IS, p. 313)

Es así como, en el caso de Iris, por más que se intente desligarla de la categoría de agente racional bajo la premisa de que simplemente sigue comandos, y aunque parezca que toda su intencionalidad es únicamente el resultado de un diseño previo y de una serie de modificaciones realizadas por terceros, esta explicación no termina de encajar con las acciones que lleva a cabo ni con el desenlace de la película. Quien desconozca que Iris es un robot la tratará como a cualquier otra persona, no solo porque físicamente lo parece, sino porque actúa como una. Y defender, sin mayores argumentos, que el único ser capaz de tener intencionalidad es el ser humano es, como diría John Martin Fischer "megalomanía metafísica" (Fischer, 2007, 67).

Aquí hay un punto importante que hasta ahora no hemos resaltado lo suficiente, pero que resulta fundamental para comprender la propuesta de Dennett. La intencionalidad, y todo aquello que deriva de ella (acciones, decisiones, comentarios, expectativas, etc.), no necesariamente ha de ser una característica exclusivamente humana. Cuando atribuimos intencionalidad a una rana, una persona o una máquina, no estamos proyectando sobre ellos una versión reducida de la intencionalidad humana. Más bien, estamos apelando a una noción más amplia, de la cual los seres humanos constituimos solo un caso particular.

Algo similar ocurre con la inteligencia. No tendría sentido definir qué es la inteligencia únicamente a partir de los rasgos mediante los cuales los humanos la manifestamos, pues ello convertiría nuestras capacidades en la medida de todas las demás. Del mismo modo, no deberíamos determinar qué cuenta como intencionalidad tomando como referencia exclusiva la experiencia humana. La cuestión relevante no es si un sistema piensa exactamente como nosotros, sino si su comportamiento puede ser comprendido y explicado mediante la atribución de creencias, deseos, objetivos y expectativas.

De hecho, cuando Dennett discute la distinción entre intencionalidad original e intencionalidad derivada, la conclusión a la que apunta es mucho más radical. Si la intencionalidad existe, no hay razones para pensar que los seres humanos poseamos una versión privilegiada o metafísicamente superior de ella. Ser humano no implica poseer una forma especial de intencionalidad que otros sistemas no puedan tener. Por el contrario, la propia historia evolutiva nos sitúa dentro de la misma

continuidad natural que el resto de los sistemas capaces de exhibir conductas intencionales. Conviene enfatizar este punto porque el caso de Iris podría inducir una intuición engañosa. La película presenta un robot prácticamente indistinguible de un ser humano. Mientras los demás personajes ignoran su naturaleza artificial, la juzgan y se relacionan con ella como si fuera una persona. Sin embargo, una vez descubren que es un robot, muchos de los rasgos que antes parecían evidentes —sus creencias, sus deseos, sus miedos o sus proyectos— pasan a ser puestos en duda. La diferencia de trato no surge necesariamente de un cambio en su comportamiento, sino del supuesto previo de que una máquina no puede ser un agente intencional en el mismo sentido en que lo es un ser humano.

De hecho, cuando Iris decide matar a Josh ocurre lo siguiente:

> *JOSH: ¡Maldita sea! ¡Joder! ¡Casi me disparas!*
>
> *IRIS: Y me siento fatal por ello.*
>
> *(Entonces)*
>
> *Mira eso. Mi primera mentira. Qué divertido. Entiendo por qué.*
>
> Hancock (2025, mi traducción)

Este momento es particularmente interesante porque la mentira parece requerir mucho más que una simple respuesta automática. Para mentir, Iris debe reconocer una diferencia entre aquello que ocurrió y aquello que decide comunicar. Debe comprender que Josh espera una determinada respuesta y, además, debe modificar deliberadamente la información que entrega. Cuando afirma "Comprendo el atractivo", no está simplemente describiendo una nueva función adquirida, sino reconociendo que ha descubierto una herramienta social que antes no utilizaba.

Desde la actitud intencional, esta escena resulta perfectamente comprensible. Atribuimos a Iris creencias sobre la situación, deseos respecto de su propia supervivencia y objetivos que intenta alcanzar mediante determinadas acciones. No explicamos su conducta apelando al movimiento de circuitos o al estado de sus procesadores. Tampoco recurrimos a las intenciones originales de los programadores. La explicación más natural consiste en afirmar que Iris comprendió algo acerca de la situación y actuó en consecuencia. Es decir, la tratamos exactamente como trataríamos a cualquier otro agente racional.

Así pues, volvemos nuevamente a lo planteado por Brentano, pero esta vez exponiéndolo frente a Darwin, Dennett lo hace de la siguiente manera:

> La intencionalidad, según Brentano, se supone que es la "marca de lo mental" y, sin embargo, la principal belleza de la teoría darwiniana es su eliminación de la Mente de la explicación de los orígenes biológicos. ¿Qué propósito serio podría tener entonces una metáfora tan flagrantemente engañosa? El mismo desafío podría plantearse a Dawkins: ¿Cómo puede ser sensato animar a la gente a pensar en la selección natural como un relojero, mientras añade que este relojero no solo es ciego, sino que ni siquiera intenta fabricar relojes? Podemos ver con mayor claridad la utilidad, de hecho, la utilidad ineludible de la postura intencional en biología observando otros casos de su aplicación. Los genes no son los únicos micro agentes que reciben poderes aparentemente conscientes por biólogos sobrios. (IS, p. 314, mi traducción)

Precisamente por ello, Iris constituye un ejemplo útil para los propósitos de este trabajo. Si seguimos la caracterización propuesta por Rescher, encontramos en ella los tres elementos que hemos identificado como centrales para hablar de error. En primer lugar, existe vulnerabilidad, Iris puede equivocarse acerca de las personas que la rodean, sobre sus intenciones o sobre las consecuencias de sus acciones. En segundo lugar, existe intencionalidad, pues sus acciones son comprensibles únicamente si le atribuimos creencias, deseos, expectativas y objetivos. Su comportamiento es muy complejo como para ser comprendido únicamente desde la actitud física o incluso desde la de diseño. Finalmente, existe retroalimentación. Iris modifica su comportamiento a medida que obtiene nueva información sobre Josh, sobre sí misma y sobre las circunstancias en las que se encuentra. Aprende de lo que ocurre y reorganiza sus decisiones futuras a partir de ello.

Lo importante es que ninguna de estas características depende de que Iris sea biológicamente humana. De hecho, ese es justamente el punto que Dennett intenta defender. La intencionalidad no es un privilegio exclusivo de nuestra especie, sino una estrategia explicativa aplicable a cualquier sistema cuyo comportamiento resulte comprensible mediante la atribución de creencias y deseos. Por ello, cuando Iris se equivoca, corrige sus acciones, engaña, aprende o modifica sus objetivos, la explicación que mejor funciona sigue siendo una explicación intencional.

Ahora bien, podría objetarse que todo esto depende de un escenario ficticio que exagera las capacidades tecnológicas actuales. La objeción es razonable. Sin embargo, el objetivo del ejemplo no es demostrar que hoy existan sistemas exactamente como Iris, sino aislar conceptualmente los elementos que nos interesan: intencionalidad, error, aprendizaje y corrección. La ciencia ficción nos permite observar estas relaciones sin las limitaciones técnicas del presente. Una vez identificadas, podemos regresar a los sistemas reales y preguntarnos hasta qué punto estas mismas dinámicas comienzan ya a aparecer en ellos.

6. Críticas

En este trabajo no se pretense seguir ciega o acríticamente a Dennett. Por ello vale la pena explorar algunas de las críticas que se le han formulado a la teoría de la actitud intencional.

6.1. Racionalidad y creencia

Siguiendo a Zawidzki, una de las objeciones más importantes proviene de Stephen Stich y puede resumirse en una preocupación bastante intuitiva: la relación entre racionalidad y creencia. Los seres humanos son agentes que poseen creencias; en este sentido, son creyentes. Asimismo, suelen ser considerados agentes racionales. Por ello tienen creencias, deseos y actúan, al menos en principio, de acuerdo con ellos. Sin embargo, la dificultad aparece cuando observamos que los seres humanos constantemente realizan acciones irracionales. Entregan mal el cambio, olvidan información evidente, buscan las gafas sin darse cuenta de que las llevan puestas o toman decisiones que van en contra de sus propios intereses. Si esto ocurre, ¿debemos concluir que dejan de ser agentes racionales? Y si dejan de serlo, ¿debemos suspender también la atribución de creencias y deseos? Dennett responde a esta preocupación de la siguiente manera:

> Dicho sin rodeos, en mi opinión no se puede decir en qué cree realmente alguien cada vez que comete un error cognitivo. Stich sugiere que esto es obviamente falso. Presumiblemente, cada error de este tipo es un error particular, del que parece derivarse que siempre habrá (aunque no podamos encontrarlo) una historia de creencia completa que contar al respecto. (Dennett, IS, p. 103, mi traducción)

Sin embargo, la preocupación de Stich no desaparece tan fácilmente. Resulta importante recordar que, para Dennett, un sistema intencional debe satisfacer ciertas normas mínimas de racionalidad. En términos generales: (i) posee las creencias que debería poseer dadas las circunstancias; (ii) posee los deseos que debería poseer de acuerdo con sus necesidades y objetivos; y (iii) actúa de una manera que resulta razonable a la luz de esas creencias y deseos. El problema es que los seres humanos parecen desviarse constantemente de estos criterios.

Ahora bien, esta dificultad no es exclusiva de Dennett. De hecho, varias de las teorías revisadas anteriormente enfrentan problemas similares. Leibniz, por ejemplo, sostiene que incluso los agentes racionales, dotados de percepciones genuinas y deseos orientados hacia el bien, pueden equivocarse. La razón es que toda criatura finita posee percepciones limitadas y grados variables de claridad. El error no elimina la racionalidad del agente. Simplemente pone de manifiesto que esta nunca es perfecta. Algo semejante ocurre en Spinoza. Las ideas inadecuadas no convierten automáticamente a alguien en irracional. Por el contrario, forman parte de la condición ordinaria de una mente finita. Lo relevante no es la ausencia total de error, sino la capacidad de avanzar hacia formas más adecuadas de comprensión.

Incluso Locke puede interpretarse en esta dirección cuando analiza casos extremos. Un niño pequeño, una persona gravemente enferma o alguien afectado por condiciones que limitan

severamente sus capacidades racionales pueden actuar de maneras erráticas o equivocadas. Sin embargo, esto no significa necesariamente que carezcan de deseos, creencias o intenciones. Más bien muestra que existen grados y condiciones bajo las cuales estas facultades pueden operar de forma defectuosa.

Por ello, volver a Rescher resulta particularmente útil. Si el error exige una vulnerabilidad necesaria, una intencionalidad reflexiva y un mecanismo de retroalimentación, entonces equivocarse no implica dejar de ser un agente intencional. Al contrario, parece ser una consecuencia natural de ser uno. Quien entrega mal el cambio, busca unas gafas que ya lleva puestas o toma una decisión que posteriormente lamenta sigue actuando en función de ciertos deseos y creencias. El error no consiste en la ausencia de intencionalidad, sino en una desviación entre aquello que el agente cree, aquello que desea y aquello que efectivamente ocurre. Más aún, se muestra por qué el componente de retroalimentación es importante. Cuando el agente descubre su equivocación, puede reconocer que la acción realizada no produjo el resultado esperado, revisar las creencias que la motivaron y modificar su comportamiento futuro. Precisamente por ello, el error no destruye la racionalidad; la presupone. Solo un sistema capaz de reconocer la distancia entre sus objetivos y sus resultados puede aprender de sus equivocaciones.

6.2. Describir creencias y tener creencias

Esta crítica puede resumirse de la siguiente manera: que un sistema pueda ser descrito exitosamente desde la actitud intencional no implica necesariamente que posea creencias reales. Después de todo, también describimos termostatos, programas de ajedrez o sistemas de navegación en términos de objetivos, preferencias e información disponible, sin por ello asumir que poseen una vida mental comparable a la humana. En este sentido, el éxito predictivo de la actitud intencional parecería mostrar únicamente que se trata de una herramienta útil de descripción, pero no que las creencias constituyan fenómenos objetivamente o científicos.

Sin embargo, esta objeción pierde parte de su fuerza cuando introducimos el problema del error. Pues no estamos hablando únicamente de sistemas cuyo comportamiento puede predecirse, sino de sistemas capaces de equivocarse, reconocer la equivocación, modificar su conducta y aprender de ella. La cuestión ya no es simplemente si actúan como si tuvieran creencias, sino si existe una explicación alternativa capaz de dar cuenta de estos procesos con la misma eficacia explicativa.

Y, si se deseas forzar el asunto, siempre estamos en el filo de navaja de caer en el problema de las otras mentes. La manera como atribuimos a otros seres humanos la posesión de creencias es mediante una inferencia que tiene como premisas su conducta y la similitud biológica con nosotros. Si se comporta como yo, y tiene una constitución biológica como la mía, ha de tener mí misma vida mental. Por contrapartida, un comportamiento similar, pero con estructura no-biológica, o biológica muy diferente, no debería tener una vida mental como la mía. Pero este argumento, falaz, encierra una premisa escondida, a saber, que sólo organismos con mi misma disposición biológica podrían tener vida mental, o que toda vida mental o intencionalidad y creencia, ha de ser similar a la humana, premisa que no se ha probado pero el argumento acepta como verdadera.

Por otro lado, el autor Matthew Elton, nos presenta dos críticas más[29].

6.3. ¿Escoger una posición?

Elton sostiene que la teoría de Dennett enfrenta una dificultad respecto a cómo se selecciona una actitud explicativa. Su preocupación es que las tres actitudes —física, de diseño e intencional— parecen presentarse como opciones que el observador escoge deliberadamente según le resulte conveniente. Si esto es así, nada impediría que alguien adoptara una actitud inadecuada para el sistema que pretende explicar, generando descripciones excesivas, innecesarias o simplemente equivocadas.

Para ilustrar esta preocupación, Elton recurre al ejemplo del termostato (y, de manera similar, al de una alarma). Allí afirma que para Dennett: "[...] Los sistemas intencionales versátiles no son tan diferentes de los termostatos. Señala que las razones para considerar un termostato como un sistema intencional no impresionan, pero no son cero" (Elton, 2003, p. 43). La crítica consiste en señalar que, si seguimos esta línea, podríamos adoptar la actitud intencional frente a objetos para los cuales dicha estrategia parece innecesaria. Podríamos decir que el termostato "cree" que la temperatura es demasiado baja y que "desea" aumentarla, pero este vocabulario no parece aportar nada relevante para explicar por qué funciona o por qué se avería. Bastaría con una explicación desde la actitud de diseño o incluso desde la actitud física.

Sin embargo, el ejemplo no resulta completamente favorable para Elton. El autor parece llevar la propuesta de Dennett a uno de sus extremos e ignora un aspecto de la teoría: Dennett nunca afirma que debamos utilizar **siempre** la actitud intencional. Por el contrario, insiste en que las distintas posturas poseen ámbitos de aplicación diferentes y que, cuando una explicación más simple resulta suficiente, no existe razón para adoptar una más compleja. Como afirma Dennett:

---

[29] Elton presenta originalmente seis críticas si se incluye la discusión desarrollada en "Clarifying the Design Stance" (2003, p. 38). Allí sostiene que la actitud intencional asume la racionalidad del sistema, mientras que la actitud de diseño asume su correcto funcionamiento. En consecuencia, ambas parecen conducir a predicciones similares, diferenciándose únicamente en el tipo de explicación que ofrecen: una en términos de creencias y deseos, la otra en términos de funciones y propósitos.
No podemos seguir al autor en este punto por dos razones. Primero, porque la semejanza entre ambas actitudes no constituye una objeción para Dennett, sino precisamente una característica central de su propuesta. La actitud intencional y la actitud de diseño no son estrategias completamente independientes, sino niveles explicativos relacionados. Segundo, porque reducir la diferencia entre ellas a una mera variación de vocabulario corre el riesgo de ignorar aquello que la actitud intencional permite explicar y que la actitud de diseño no siempre captura adecuadamente, especialmente en sistemas complejos cuyo comportamiento puede describirse de forma más eficiente mediante la atribución de creencias, deseos y objetivos.
Por esta razón, consideramos que la discusión puede incorporarse dentro de la crítica general sobre la elección de una actitud explicativa y no requiere un tratamiento independiente.
Tampoco desarrollaremos extensamente la crítica sobre la distinción entre intencionalidad original e intencionalidad derivada ("Intrinsic, as if, and derived intentionality"), ya que fue abordada anteriormente y Elton no introduce elementos sustancialmente nuevos frente a los debates ya examinados entre Dennett, Fodor y Searle.
Finalmente, Elton retoma la discusión entre holismo y atomismo, vinculada principalmente a las objeciones de Fodor. La preocupación consiste en que dos agentes difícilmente podrían compartir exactamente una misma creencia, dado que cada creencia depende de una red más amplia de creencias relacionadas. Sin embargo, cuando analizamos la respuesta de Dennett a Fodor ya quedó expuesto por qué una perspectiva holista no constituye una dificultad decisiva para la actitud intencional. Lo relevante no es la identidad perfecta entre contenidos mentales internos, sino la capacidad explicativa y predictiva que se obtiene al atribuir determinados estados intencionales a un sistema.

> ¿Por qué debería descalificar al atril? Para empezar, la estrategia no se recomienda en este caso, ya que no obtenemos de ella ningún poder predictivo que no tuviéramos previamente. Ya sabíamos lo que el atril iba a hacer —es decir, nada— y adaptamos las creencias y los deseos para que encajaran de una manera bastante poco ética. En el caso de las personas, los animales o las computadoras, sin embargo, la situación es diferente. En estos casos, a menudo la única estrategia práctica es la intencional; nos brinda un poder predictivo que no podemos obtener por ningún otro método. (Dennett, IS, p. 23, mi traducción)

Por ello, estamos de acuerdo en que artefactos como termostatos, alarmas o licuadoras podrían ser descritos desde la actitud intencional. No obstante, bajo la lógica de la cascada explicativa que hemos presentado, no parece necesario ascender hasta ese nivel cuando la actitud de diseño ya proporciona una explicación satisfactoria. En estos casos, la descripción intencional es posible, pero no especialmente útil. No estamos obligados a usarla. Las tres actitudes no compiten necesariamente entre sí, sino que funcionan como niveles explicativos distintos, donde las explicaciones más complejas solo resultan justificadas cuando las más simples dejan de ser suficientes.

6.4. Entradas y Salidas[30]

La segunda crítica de Elton sostiene que Dennett no presta suficiente atención al papel de los inputs y outputs en la atribución de intencionalidad. Según el autor, "Cuando adoptamos las diferentes posturas, no estamos simplemente adoptando una postura hacia el funcionamiento interno del sistema. También lo estamos adoptando en relación con nuestra comprensión de sus entradas y salidas" (p. 43). Por ello, considera que ni la actitud física ni la de diseño son suficientes para explicar el significado de ciertos comportamientos; haría falta un paso interpretativo adicional que permitiera comprender esos inputs y outputs como portadores de contenido.

Sin embargo, esta objeción parece devolvernos a una discusión ya conocida. En el fondo, recuerda los problemas planteados por Turing y por el debate entre sintaxis y semántica. La crítica de Elton parece reintroducir la exigencia de una semántica original o intrínseca, una exigencia que Dennett intenta precisamente evitar al desplazar la atención hacia el éxito explicativo y predictivo de la actitud intencional. Tal y como hicimos atrás en el caso de la objeción de Searle a Turing, sería un error pedir que sólo una parte del sistema "comprendiera", aislándola del resto del sistema como un todo. En la habitación china no es la persona encerrada dentro quien ha de comprender el chino, sino el sistema entero (la habitación, las notas escritas que entran y salen, el libro con el manual de combinaciones etc.). Pedir que el input sea, *per se*, significativo o portador de contenido es nuevamente pedir que una parte del sistema sea un sistema completo, que mis neuronas occipitales "vean" cuando es el cerebro entero, más mis ojos y mi cuerpo el que está en capacidad de hacerlo.

****

[30] Conviene tener cuidado con el propio lenguaje de inputs y outputs. Hablar en estos términos ya supone una determinada manera de conceptualizar el sistema. O bien estamos frente a un artefacto cuya conducta todavía no sabemos si debe describirse intencionalmente, o bien estamos frente a un sistema que ya hemos abstraído como un agente capaz de procesar información para alcanzar ciertos fines. En ambos casos, los inputs y outputs no constituyen por sí mismos una objeción a la teoría de Dennett, sino una dimensión adicional que debe ser explicada dentro de ella.

En términos generales, las críticas revisadas muestran dificultades importantes para la teoría de la actitud intencional, especialmente cuando se intenta precisar la objetividad de las creencias, la relación entre los distintos niveles explicativos o los criterios para atribuir racionalidad. Sin embargo, ninguna parece comprometer directamente el ejemplo desarrollado a partir de Compañera perfecta. La objeción de Stich, por ejemplo, cuestiona qué ocurre cuando un agente se comporta de manera irracional. Pero justamente Iris sigue siendo interpretable mediante creencias y deseos incluso cuando se equivoca, miente o actúa de manera contradictoria. Del mismo modo, la preocupación sobre la objetividad de las creencias no afecta el análisis realizado, pues nunca dependimos de demostrar que los estados mentales de Iris existieran como entidades científicas observables; bastó con mostrar que atribuirlos permite comprender y predecir exitosamente su conducta.

Algo similar ocurre con las críticas de Elton. La preocupación sobre la elección de una actitud no representa un problema para nuestro caso, ya que precisamente mostramos por qué la descripción física o la de diseño resultan insuficientes para explicar ciertas acciones de Iris, mientras que la actitud intencional ofrece una explicación más económica y potente. La discusión sobre Entradas y Salidas, tampoco altera la conclusión alcanzada, pues incluso si aceptamos que existe un proceso de interpretación adicional, seguimos describiendo exitosamente el comportamiento del sistema mediante creencias, deseos y objetivos. Finalmente, que Iris pueda describirse simultáneamente como un artefacto físico, un producto de diseño y un sistema intencional no implica que estemos hablando de entidades distintas, sino de distintos niveles de análisis sobre el mismo agente. Por ello, las objeciones examinadas funcionan más como advertencias metodológicas que como refutaciones. Nos obligan a ser cautelosos respecto a qué estamos atribuyendo y por qué lo hacemos, pero no eliminan el hecho central que intentamos mostrar: que sistemas artificiales como Iris pueden ser comprendidos de manera fructífera mediante la actitud intencional (que de hecho es la mejor posición que podemos tomar) y que, bajo estas condiciones, conceptos como creencia, deseo, aprendizaje y error conservan un poder explicativo que difícilmente puede simplificarse o ignorarse.

7. Conclusiones

En años recientes se presentó un caso en el que un robot industrial encargado de levantar y organizar cajas aplastó a un trabajador. El incidente tuvo una enorme repercusión mediática. Nos encontrábamos en pleno auge de las nuevas tecnologías y no faltaron las reacciones alarmistas: "esto es Terminator en la vida real", "no falta mucho para que Blade Runner deje de ser ficción". Sin embargo, este caso no parece encajar del todo con lo que hemos expuesto hasta ahora. El robot simplemente ejecutaba las instrucciones para las que había sido diseñado y, por ello, nadie atribuyó responsabilidad alguna a la máquina. El error se ubicó en el diseño, en los protocolos de seguridad o en la supervisión humana, pero nunca en el agente artificial mismo.

Ahora bien, ¿hasta cuándo puede mantenerse esta posición? No parece imposible imaginar sistemas artificiales que actúen con niveles crecientes de autonomía, aprendizaje y adaptación. No estamos tan lejos de escenarios en los que la explicación de una acción ya no pueda agotarse únicamente apelando a la programación inicial o al diseño del sistema.

Algo similar ocurre en el relato de Isaac Asimov, “Todos los males del mundo”; allí el padre de un adolescente aparece registrado por Multivac como un futuro delincuente. Ante la negativa del padre de estar planeando algún crimen, su propio hijo responde: "Tal vez sea algo subconsciente, papá; una forma de resentimiento hacia tu jefe". El hijo confía más en la capacidad de procesamiento de la máquina que en el conocimiento que su padre tiene de sus propias intenciones. Cuando su madre manifiesta desconcierto ante esta situación, él responde: "Verás, mamá..., es que Multivac no se equivoca nunca".

Aquí hablar de equivocación es casi una figura retórica. Lo que realmente se afirma es que Multivac no es como nosotros. Los seres humanos se equivocan; la máquina no. De este modo se refuerza la idea de que el error es una categoría propia de los agentes humanos y que no resulta aplicable a sistemas artificiales, incluso cuando estos alcanzan niveles extraordinarios de complejidad.

Más adelante, cuando las cosas comienzan a salirse de control, los operadores de Multivac tampoco atribuyen el problema a un error de la máquina. La explicación vuelve a recaer sobre los seres humanos. Como los menores de edad son registrados junto a los datos de sus padres, Multivac no puede distinguir adecuadamente entre ambos conjuntos de información. El problema no es que la máquina haya errado, sino que los datos fueron organizados de una manera que conduce a una clasificación equivocada. Nuevamente, la falla se localiza en el diseño del sistema y no en su funcionamiento.

Se puede observar entonces que incluso en la ciencia ficción, donde encontramos algunas de las máquinas más avanzadas imaginadas por la literatura, el error suele reservarse para describir la falibilidad humana. Cuando algo sale mal, la equivocación aparece en quien programa, diseña o alimenta el sistema, mientras que la máquina permanece como símbolo de precisión y perfección. Sin embargo, esa imagen resulta cada vez más difícil de sostener. A medida que los sistemas artificiales adquieren mayores capacidades de aprendizaje, adaptación y autonomía, la diferencia

entre un simple fallo mecánico y un error en sentido fuerte comienza a difuminarse. Si un sistema puede interpretar información, actuar con base en ella, corregir su comportamiento y modificar estrategias futuras, entonces la pregunta por el error deja de ser una cuestión exclusivamente humana.

Este trabajo no pretende resolver todas las dificultades que surgen de esta posibilidad. El objetivo es moderado: mostrar que el error no parece ser un fenómeno exclusivamente humano y que las herramientas filosóficas con las que tradicionalmente lo hemos pensado permiten también abordar ciertos comportamientos artificiales. Si esto es correcto, entonces algunas preguntas que antes pertenecían a la ciencia ficción comienzan a convertirse en problemas filosóficos, jurídicos y políticos de primer orden.

¿Si podemos eventualmente atribuir error a los agentes artificiales, podremos alguna vez atribuirles también responsabilidad? ¿Necesitaremos nuevas categorías conceptuales para comprender sus acciones? ¿O bastará con ampliar las categorías que ya utilizamos para explicar nuestros propios errores? Si aceptamos que ciertos sistemas artificiales pueden exhibir vulnerabilidad, intencionalidad y mecanismos de retroalimentación, entonces estas preguntas dejan de ser especulaciones sobre el futuro para convertirse en problemas del presente.

Todavía quedan muchas incógnitas por resolver. Sin embargo, hay algo que parece claro: debemos comenzar a responderlas ahora, antes de que nuestras teorías sobre el error, la racionalidad y la responsabilidad queden tan desactualizadas como aquellos personajes de Asimov que confiaban ciegamente en que Multivac nunca podía equivocarse.

8. Referencias

Agustín de Hipona. (1982). *Contra los académicos* (V. Capánaga, Trad.). Biblioteca de Autores Cristianos. (Obra original ca. 386–388)

Agustín de Hipona. (2010*). Confesiones* (A. Custodio Vega, Trad.). Biblioteca de Autores Cristianos. (Obra original ca. 397–400)

Aristóteles. (1970). *Metafísica* (V. García Yebra, Trad.). Editorial Gredos.

Aristóteles. (1985). *Ética Nicomáquea / Ética Eudemia* (J. Pallí Bonet, Trad.; E. Lledó Íñigo, Introd.). Editorial Gredos. (Obras originales ca. 350 a.C.)

Asimov, I. (1958). "All the troubles of the world". Super-Science Fiction, 2(3). [Publicado en español como «Todos los males del mundo» en I. Asimov (1990), *Cuentos completos I* (pp. 579–596). Ediciones B.]

Augustine. (2002). *On the Trinity*, Books 8–15 (G. B. Matthews, Ed.; S. McKenna, Trad.). Cambridge University Press. (Obra original ca. 400–416)

Bennett, J. (1984). *A study of Spinoza's Ethics*. Cambridge University Press.

Bennett, J. (1986). "Spinoza on error". *Philosophical Papers*, 15(1), 59–73. https://doi.org/10.1080/05568648609506252

Bolyard, C. (2018). "Augustine on error and knowing that one does not know". En A. Speer y M. Mauriège (Eds.), *Irrtum – Error – Erreur* (pp. 3–18). De Gruyter. https://doi.org/10.1515/9783110592191-003

Brentano, F. (1995). *Psychology from an empirical standpoint* (A. C. Rancurello, D. B. Terrell y L. L. McAlister, Trads.; P. Simons, Pról.). Routledge. (Obra original 1874) [Psicología desde un punto de vista empírico]

Burgos, J. (2026). "Daniel Dennett: Intencionalidad original, evolución, estrategia intencional y eliminativismo". *Ideas y Valores*, 75(190), 34–62.

Chavarriaga, R., Sobolewski, A., y Millán, J. D. R. (2014). "Errare machinale est: The use of error-related potentials in brain-machine interfaces". *Frontiers in Neuroscience*, 8, 208. https://doi.org/10.3389/fnins.2014.00208

Chen, N., Mohanty, S., Jiao, J., y Fan, X. (2021). "To err is human: Tolerate humans instead of machines in service failure". *Journal of Retailing and Consumer Services*, 59, 102363. https://doi.org/10.1016/j.jretconser.2020.102363

Choi, S., Mattila, A. S., y Bolton, L. E. (2021). "To err is human(-oid): How do consumers react to robot service failure and recovery?" *Journal of Service Research*, 24(3), 354–371. https://doi.org/10.1177/1094670520978798

Churchland, P. M. (1999*). Materia y conciencia: Introducción contemporánea a la filosofía de la mente* (M. N. Mizraji, Trad.; 2.ª ed.). Editorial Gedisa. (Obra original revisada 1988)

Cicerón, M. T. (s.f.). *Philippics 12. Lexundria*. Recuperado el 24 de junio de 2026, de https://lexundria.com/cic_phil/12/y

Dennett, D. C. (1978). *Brainstorms: Philosophical essays on mind and psychology*. Bradford Books / MIT Press.

Dennett, D. C. (1987). *The intentional stance*. MIT Press.

Dennett, D. C. (1998). *La actitud intencional* (D. Zadunaisky, Trad.). Editorial Gedisa. (Obra original 1987)

Descartes, R. (2009). *Meditaciones acerca de la filosofía primera seguidas de las objeciones y las respuestas* (J. A. Millán Alba, Trad.). Alianza Editorial. (Obra original 1641)

Elton, M. (2003). *Daniel Dennett: Reconciling science and our self-conception*. Polity Press.

Epicuro. (2002*). El arte de la felicidad* (C. García Gual, Trad.). Editorial Edaf.

Epicuro. (2012). *Obras completas* (J. Vara, Ed. y Trad.). Cátedra. (Obras originales ca. siglos IV–III a.C.)

Favaretti Camposampiero, M., Priarolo, M., y Scribano, E. (Eds.). (2016). "I volti dell'errore nel pensiero moderno: Da Bacone a Leibniz" [Número especial*]. Rivista di Storia della Filosofia*, 71(4).

Fischer, J. M. (2007). "Semi-compatibilism" En: Fischer, J.M., Pereboom, D., Kane, R. & Vargas, M. *Four Views on Free Will*. 2007. Malden: Blackwell.

Fodor, J. A. (1978). "Propositional attitudes". *The Monist*, 61(4), 501–523. https://doi.org/10.5840/monist197861444

Hancock, D. (2025). *Companion* [Guion de película]. Script Hive. https://www.scripthive.com/database/script/429411b9-c4b2-4347-ba5b-247562d48400

Hernández Girardi, R. (2011). "La percepción como criterio en Epicuro". *Euphyía*, 5(9), 46–61.

Honig, S., y Oron-Gilad, T. (2018). "Understanding and resolving failures in human-robot interaction: Literature review and model development". *Frontiers in Psychology*, 9, 861. https://doi.org/10.3389/fpsyg.2018.00861

Kripke, S. (1982). Wittgenstein on Rules and Private Language. Harvard University Press.

Leibniz, G. W. (1989). *Philosophical papers and letters* (2.ª ed.; L. E. Loemker, Ed. y Trad.). Springer Netherlands. (Obras originales 1666–1716)

Leibniz, G. W. (2004). *Nuevos ensayos sobre el entendimiento humano* (J. Echeverría Ezponda, Trad.). Alianza Editorial. (Obra original 1765)

Locke, J. (2005). *Ensayo sobre el entendimiento humano* (E. O'Gorman, Trad.). Fondo de Cultura Económica. (Obra original 1689)

LoLordo, A. (2008). "Locke's problem concerning perceptual error". *Philosophy and Phenomenological Research*, 77(3), 686–705. https://doi.org/10.1111/j.1933-1592.2008.00216.x

Lucrecio. (1993). *De la naturaleza de las cosas* (F. Socas, Trad.). Editorial Gredos. (Obra original ca. 55 a.C.)

Maitra, K. (2002). "Leibniz's account of error." *International Journal of Philosophical Studies*, 10(1), 63–73. https://doi.org/10.1080/09672550110103073

Mayr, E. (2011). *Understanding human agency*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199606214.001.0001

Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., y Tscheligi, M. (2017). „To err is robot: How humans assess and act toward an erroneous social robot". *Frontiers in Robotics and AI*, 4, 21. https://doi.org/10.3389/frobt.2017.00021

Putnam, H. (1967). "Psychological predicates". En W. H. Capitan y D. D. Merrill (Eds.), *Art, mind, and religion* (pp. 37–48). University of Pittsburgh Press. [Reimpreso como «The nature of mental states» en W. G. Lycan (Ed.), *Mind and cognition: An anthology* (2.ª ed., pp. 27–34). Blackwell, 1999.]

Rescher, N. (2007). *On error: Researches into the nature and causes of wrongheadedness*. University of Pittsburgh Press.

Rist, J. M. (1972). *Epicurus: An introduction*. Cambridge University Press.

Ryle, G. (2009). *El concepto de lo mental* (D. C. Dennett, Introd.; E. Rabossi, Trad.). Paidós. (Obra original 1949)

Spinoza, B. (2000). *Ética demostrada según el orden geométrico* (Edición bilingüe; A. Domínguez, Ed. y Trad.). Trotta. (Obra original 1677)

Spinoza, B. (2006). *Tratado de la reforma del entendimiento y otros escritos* (A. Domínguez, Ed. y Trad.). Alianza Editorial. (Obras originales 1677)

Stich, S. P. (1981). "Dennett on intentional systems". *Philosophical Topics*, 12(1), 39–62. https://doi.org/10.5840/philtopics198112142

Strawson, Peter (1963). "Freedom and Resentment". *Proceedings of the British Academy* 48:187-211.

Turing, A. M. (1950). "Computing machinery and intelligence". *Mind*, 59(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Zawidzki, T. (2007). *Dennett.* Oneworld Publications.

Zhou, L. (2022). "Descartes on the source of error: The Fourth Meditation and the correspondence with Elisabeth". *British Journal for the History of Philosophy,* 30(6), 992–1012. https://doi.org/10.1080/09608788.2022.2132908