
naturaleza
y libertad
revista de filosofía

Para la publicación de este número se ha contado con la ayuda
financiera de las siguientes instituciones:
**Departamento de Filosofía y Lógica y Filosofía de la Ciencia
de la Universidad de Sevilla**
Asociación de Filosofía y Ciencia Contemporánea. Madrid

DEBATE SOBRE LA INTELIGIBILIDAD DE LA CONCIENCIA

Número Monográfico de
NATURALEZA Y LIBERTAD
Revista de estudios interdisciplinares

Número 7

Málaga, 2016
ISSN: 2254-9668

Esta revista es accesible *on-line* en el siguiente portal:
http: //grupo.us.es/naturalezayl

Redacción y Secretaría:
Naturaleza y Libertad. Revista de estudios interdisciplinares. Departamento de Filosofía y Lógica. Calle Camilo José Cela s.n. E-41018 Sevilla. Depósito Legal: MA2112-2012
ISSN: 2254-9668
☎ 954.55.77.57 Fax: 954.55.16.78. E-mail: jarana@us.es

ÍNDICE

MÁQUINAS COMPUTACIONALES Y CONCIENCIA ARTIFICIAL

Gonzalo Génova Fuster
Universidad Carlos III de Madrid

Vivimos en una época en la que nos enorgullecemos de las máquinas pensantes
y desconfiamos de los hombres que intentan pensar.
Juan Génova, *in memoriam*

Resumen: En este ensayo argumento que la noción de máquina incluye necesariamente el hecho de ser diseñada con un fin. Por tanto, no basta con ser sistema mecánico para ser máquina. Puesto que el método científico-experimental excluye metodológicamente la consideración de la finalidad, resulta que también es insuficiente para entender cabalmente qué son las máquinas. Por el contrario, para entender una máquina es necesario ante todo entender su finalidad (y también su estructura), en claro paralelismo con la causa final (y la formal) de Aristóteles. Obviamente, la finalidad y la estructura no son componentes de la máquina que puedan interaccionar físicamente con otros componentes, y sin embargo son esenciales para comprender su funcionamiento. Esto arroja una interesante luz sobre la relación entre la mente y el cuerpo: de modo análogo a como la finalidad de un artefacto explica su funcionamiento, la conciencia es la explicación del comportamiento específicamente humano. Máquinas y seres humanos tienen en común que, para entenderlos, es necesario acudir al principio de finalidad. Pero mientras en las máquinas es una finalidad dada, los seres humanos nos caracterizamos porque podemos proponernos nuestros propios fines.

Palabras clave: máquina, algoritmo, finalidad, diseño, conciencia, libertad.

Computational machines and artificial consciousness

Abstract: In this essay I argue that the notion of machine necessarily includes its being designed for a purpose. Therefore, being a mechanical system is not enough for being a

machine. Since the experimental-scientific method methodologically excludes the consideration of finality, it is then also insufficient to fully understand what machines are. On the contrary, in order to understand a machine it is first required to understand its purpose (and also its structure), in clear parallel with Aristotle's final and formal causes. Obviously, purpose and structure are not machine components that can physically interact with other components, however they are essential to understand its operation. This casts an interesting light on the relationship between mind and body: in analogous way as an artifact's finality explain its operation, consciousness is the explanation of specifically human behavior. Machines and human beings have in common that, in order to understand them, it is necessary to appeal to the principle of finality. Yet, while finality is given in the case of machines, we human beings are characterized by our possibility to self-propose our own ends.

Keywords: machine, algorithm, finality, design, consciousness, freedom.
Recibido: 31/08/2016 **Aprobado**: 12/09/2016

1. Introducción

Para comenzar, tengo que decir que, aunque estoy en un departamento de informática y enseño principalmente informática, las reflexiones que me han conducido aquí, y a dialogar con el libro de Juan Arana, *La conciencia inexplicada* (2015), se originan más bien en mi actividad como profesor de ética para ingenieros. Esta es una actividad fascinante, porque los estudiantes de ingeniería manifiestan, quizá como nadie, esa mentalidad "naturalista" que ya ha sido aludida varias veces en este seminario, y que considera, entre otras cosas, que el método científico-experimental es el camino privilegiado, incluso el único camino, para alcanzar el conocimiento de la realidad (Génova & González, 2012). Una mentalidad que puede resumirse también en los dos

tópicos de nuestra cultura mediática que identifica Ignacio Quintanilla (2012): que las ciencias naturales tienen el monopolio del saber racional y objetivo; y que con el tiempo suficiente, y sin las rémoras y los prejuicios de la tradición, la técnica podrá solucionar todos nuestros problemas. No obstante, la propia actividad científica e ingenieril contiene elementos que rompen este esquema y reclaman la apertura de la inteligencia a otras formas de razonar que no pueden ser encerradas dentro de los procedimientos del método científico y la producción técnica (Génova & González, 2016), abriendo así la puerta a otros modos de pensamiento y en particular a la reflexión ética.

El esfuerzo por enseñar ética a personas con mentalidad de ingenieros me ha ayudado a alumbrar algunas ideas, que luego han confluido con reflexiones sobre la inteligencia artificial (Génova & Quintanilla, 2016), y sobre el principio de finalidad en las máquinas diseñadas (Génova, 2016). De todas formas, si me permiten la comparación, en mí está todo mezclado, afortunadamente; y, como decía Steve Jobs, "me gusta vivir en la intersección entre las humanidades y la tecnología" (Isaacson, 2011:13). Por eso me resulta tan natural pasar de la ciencia y la ingeniería a la filosofía, y vuelta.

2. Puntos de acuerdo con el libro

Antes de entrar en el núcleo de mi comentario, me gustaría señalar algunos puntos de acuerdo fundamental con el profesor Arana. Me gustaría en

primer lugar agradecer una visión más positiva de Descartes (45/15)[1], tratando de recuperar sus aportaciones más interesantes, lejos de una visión maniquea que lo presenta como el origen de todos los males modernos. Personalmente no me cabe duda del gran mérito que corresponde al máximo iniciador de la modernidad, aunque sólo sea por poner el dedo en la llaga en problemas que el aristotelismo renacentista decadente resolvía de modo formulario, sin hacer verdadera filosofía. No obstante, en el capítulo de errores cartesianos, estoy de acuerdo con Arana en que el principal desacierto de Descartes no fue distinguir materia y espíritu (puesto que es pertinente distinguirlos), sino sustancializarlos e independizarlos (67/23, 70/24, 192/61). Por lo tanto me parece que es muy acertado, y hay que insistir mucho en ello, todo el esfuerzo por hablar del cuerpo y la mente como "dimensiones" o aspectos de la realidad, no como "cosas" reales.

Pareja a la sustancialización de la materia está el reduccionismo epistemológico a las dimensiones materiales y medibles de la realidad (42/12, 68/24), que desemboca en la absolutización del método científico: desde esta perspectiva (a mi juicio, equivocada), la materia *es* lo que conocemos de ella mediante el método científico-experimental: lo medible, lo repetible, lo verificable, lo regular (la extensión). Y a la inversa, si no es verdad que el método científico-experimental lo puede explicar todo, tampoco es verdad que la esencia última de la realidad material sea la extensión. Descartes sustancializa

[1] Los números se refieren a la página/epígrafe del libro, ya que el autor desea ser citado por estos últimos.

no solo la materia: sustancializa la cantidad, que es sólo un aspecto de la materia. Quizás este punto se podría haber enfatizado o desarrollado más.

Suscribo igualmente la tesis de Arana, formulada en sus precisos términos (125/42), según la cual la conciencia no es naturalizable, al menos en el sentido de "naturaleza" que usa el autor, sentido que acepto porque también coincido en que no vale la pena pelearse por las palabras, tratando de hacernos dueños de sus significados. No obstante, no debemos olvidar que "natural" tiene múltiples sentidos; por ejemplo, cuando se dice que es natural que haya un reparto equitativo de tareas en la familia: obviamente, es ésta una *naturalidad* que no es la que estudian las ciencias naturales. Podríamos decir así que "natural" no se dice sólo en el sentido de *origen o fuente* de "lo que ocurre" en los fenómenos de la naturaleza, sino también en el sentido de *meta o destino* de las acciones libres, "lo que debería ocurrir". La confusión en los diversos usos de "lo natural" es fuente de múltiples equívocos (no digamos cuando se trata de naturalizar la ética); por otra parte, no considero que estos diversos sentidos sean completamente equívocos, sino parcialmente análogos.

Pero permítanme que no siga mencionando puntos de acuerdo con el libro de Arana, puesto que la lista sería interminable.

3. La noción de máquina

Vayamos a mi aportación específica, que es a la vez mi principal punto de divergencia con lo que Arana expone en su libro, y que ojalá sirva para enriquecer sus tesis. Aunque he dicho que estas reflexiones entroncan con la enseñanza de la ética para ingenieros (asunto sobre el que volveré al final de

esta exposición), empiezo ahora por una cuestión más ingenieril: *¿qué entendemos por máquina?* Me parece que el concepto de máquina engloba dos aspectos diferentes (lo sometido a leyes mecánicas, lo diseñado con un fin), pero Arana aparentemente sólo considera el primero de ellos.

El autor dice que "sabemos en líneas generales qué es una máquina" (86/28), y ha definido antes: "Máquina es toda conjunción de elementos materiales que realiza *acciones exclusivamente físicas*" (57/19, el énfasis es mío). Un poco más adelante cita la definición de su amigo Martín López Corredoira, que parece aprobar[2]: "Máquina es lo mismo que *sistema mecánico*; es un sistema cuyas acciones tienen una causa o razón de ser *explicable por la interacción mecánica* de los distintos componentes que la forman (maquinaria)" (65/22, el énfasis es mío).

A mi modo de ver, en ambas definiciones, la de Arana y la de López Corredoira, falta un elemento absolutamente fundamental. Según estas definiciones, que sólo consideran aspectos puramente físicos o mecánicos, el Sistema Solar es una máquina. Pero, ¿lo es? Me parece bastante obvio que no. El Sistema Solar no es una máquina. ¿Por qué? Porque no ha sido *diseñado*. Salvo que (una hipótesis que quizás no queramos descartar del todo) pensemos que el sistema solar ha sido construido por una poderosa civilización extraterrestre, que ha calculado y ha establecido el tamaño del Sol y las órbitas de los planetas para que la zona de habitabilidad sea una determi-

[2] Primero recoge una definición López Corredoira que critica, pero luego en nota al pie se recoge otra definición más ajustada y que Arana sí parece aprobar.

nada, y entonces pueda surgir la vida en el planeta Tierra, etc. Entonces sí, el Sistema Solar sería una máquina, porque habría sido diseñado con un fin.

Por lo tanto, si no ha sido diseñado con un fin, entonces el Sistema Solar no es propiamente una máquina, aunque metafóricamente podamos referirnos a él como tal, quizás como herencia de una tradición que todavía ve el cosmos y sus elementos como artefactos de su Creador (por tanto dotados de una finalidad "natural"). Pienso que hablar de "máquinas naturales", queriendo decir "sistemas mecánicos no finalizados", resulta hoy día más inapropiado que antaño, cuando todavía pervivía de modo general la idea de que el cosmos es creación. Incluso referirse a ellos como "sistemas" es impropio, porque se está colando implícitamente la idea de ordenación, de finalidad. Es muy difícil pensar la realidad excluyendo de verdad y radicalmente toda finalidad.

Así pues, si la finalidad y el diseño son esenciales a la noción de máquina, decir que "una máquina es una conjunción de elementos materiales que realiza acciones exclusivamente físicas" es incompleto, y me atrevo a decir que incorrecto, puesto que el elemento omitido es esencial. Aún más incorrecto es pretender que el funcionamiento de la máquina es explicable sólo por la interacción mecánica de sus componentes. Me parece que es más sabio el diccionario de la Real Academia, que también cita Arana (64/22), cuando dice que una máquina es, según su tercera acepción, "un agregado de diversas partes ordenadas entre sí y *dirigidas a* la formación de un todo". La clave está en el "dirigidas a", es decir, la finalidad, el diseño.

Es verdad que el termostato no se da cuenta de que mide la temperatura (137/46), pero su fabricante y su usuario sí se dan cuenta: ahí está la finali-

dad, la intencionalidad, el significado. La Mettrie se preguntaba si la materia de un reloj puede marcar las horas (51/17), a lo que Arana responde muy acertadamente que marcar las horas es una *interpretación* del movimiento del reloj, y como tal interpretación corresponde a un acto mental que está fuera del propio reloj: las agujas del reloj no marcan las horas, el que marca las horas es el que las mira, o el que ha fabricado el reloj para que las marque. En otras palabras: las acciones de una máquina no son *exclusivamente físicas* ni se pueden explicar sólo como interacciones mecánicas: son *acciones intencionales*, dirigidas a un propósito, que alguien ha establecido desde fuera. Por tanto, es imposible entender una máquina sin entender su propósito. Pretendo ilustrar esta tesis con un sencillo y barato experimento mental.

4. Un experimento mental en el desván

Una de las tareas habituales a las que nos enfrentamos los profesionales de la ingeniería consiste en lo que denominamos el proceso de ingeniería inversa (*reverse engineering*); es decir, dado un artefacto, tratar de averiguar cómo funciona y para qué sirve, con el fin de mejorarlo o, si es el caso, imitarlo. En un primer momento puede parecer que el tradicional método científico-experimental es adecuado para resolver esta tarea, pero una mirada más atenta revela que esto no es así. En efecto, descubrir *para qué sirve* un artefacto, es decir, su *finalidad*, es tanto como descubrir la *intención* con la que fue fabricado (o incluso la intención con la que es utilizado, que puede no coincidir con la primera: piénsese en tantos objetos que usamos como pisapapeles…).

Tomemos el ejemplo de un artefacto de forma irreconocible que encontramos en un viejo desván. Lo observamos atentamente, y descubrimos que tiene un movimiento cíclico con un periodo de 24 horas más un minuto, es decir, *parece un reloj*. Hasta aquí todo bien, esta observación de un movimiento regular está al alcance del método científico-experimental. Ahora bien, de esta observación no podemos concluir *cuál fue la intención* que tenía el fabricante que construyó tal artefacto. Al menos se nos presentan tres posibilidades para esta intención: (a) la finalidad del artefacto era manifestar un movimiento cíclico con un periodo de 24 horas, pero resulta que por diversas causas el periodo es algo mayor, es decir, hay un *funcionamiento defectuoso*; (b) la finalidad del artefacto era manifestar un movimiento cíclico con un periodo de 24 horas más un minuto, y resulta que *funciona perfectamente*, pues ese es el periodo que manifiesta; (c) la finalidad del artefacto era una que no tenía *nada que ver* con el movimiento cíclico y que no se nos ha ocurrido todavía —si hay movimiento cíclico observable es pura casualidad, o efecto no intencionado del fabricante.

La cuestión importante, para este ensayo, es que *no hay ningún experimento imaginable* que sirva para determinar cuál de estas opciones es la correcta. La única forma de saber cuál era la intención del fabricante, y por tanto para qué sirve el artefacto, es *preguntar al propio fabricante* (una forma de preguntarle es leer el manual de instrucciones o atender a otras manifestaciones públicas de su intención). Puede ocurrir incluso que la finalidad del artefacto fuera ser regalado como motivo decorativo, por tanto su exactitud para medir el tiempo es irrelevante. De lo que concluimos que el método científico-experimental, en sentido estricto, es *insuficiente para averiguar para*

qué sirve un artefacto, es decir, para hacer ingeniería inversa; pero no olvidemos, que, como ya he dicho, ésta forma parte habitual de las tareas de los ingenieros.

De aquí podemos concluir que es necesario abandonar un paradigma de pensamiento que es *empobrecedor* para la propia ciencia e ingeniería. El método científico-experimental no tiene la respuesta a todas las preguntas, en particular las preguntas acerca de la finalidad de los artefactos que construyen los ingenieros. El mundo humano de las intenciones es tan real, o más, que el mundo de las relaciones físico-mecánicas. Para hacer buena ingeniería *no basta con dominar las leyes de la naturaleza*; es también esencial la hermenéutica de la intencionalidad y el contexto del simbolismo de los artefactos.

Como decía al final del epígrafe anterior, las acciones de una máquina no son *exclusivamente físicas*: son *acciones intencionales*, dirigidas a un propósito, por lo que resulta imposible entender una máquina sin entender su finalidad. Es imposible entender qué es una máquina desde el reduccionismo del método científico-experimental (el cual excluye de partida el análisis de la finalidad); o, dicho de otra forma, es imposible entender una máquina si la reducimos a sus aspectos físico-mecánicos. Con esto no quiero decir que el método científico-experimental esté equivocado en su olvido deliberado de la finalidad, sino sólo que es insuficiente para entender cabalmente aquellas realidades caracterizadas por tener finalidad, como son los actos libres, y los productos de esos actos libres, tales como los artefactos humanos. El método científico es legítimo tal cual es, olvidadizo y desconsiderado con la finalidad; lo que no es legítimo es absolutizarlo, el reduccionismo. No me propongo

desautorizar el método científico, sino tan sólo destronarlo de un lugar que no le corresponde.

Que los artefactos tienen finalidad es algo obvio y que admiten sin dificultad incluso gente con mentalidad tan poco filosófica como los ingenieros. No es ese el punto, sino que esa finalidad no se descubre con el método científico entendido en sentido estricto. Es decir, las explicaciones finalísticas no son verificables ni falsables. Los experimentos científicos pueden decir lo que ocurre, y con qué regularidad. Pueden establecer rigurosamente una regularidad, una ley. Pero no pueden asegurar que esa regularidad responde a un diseño intencionado, ni cuál era ese diseño. No podemos saber con certeza la intención, salvo interrogando al autor. Propiamente hablando, no puede haber evidencia *empírica* de que existe diseño.

Con esto no quiero decir que el razonamiento hacia atrás, hacia el diseño, sea ilegítimo; no quiero decir que no podamos conocer (quizás mejor decir "adivinar") la finalidad, que no podamos hacer ingeniería inversa: ¡lo hacemos los ingenieros todos los días! Lo que quiero decir es que los experimentos no son el camino adecuado, suficiente, para lograrlo.

En resumen, es imposible entender lo que realmente es una máquina sin entender para qué sirve, su propósito. Y para entenderlo no basta el método científico-experimental.

5. Finalidad y estructura de las máquinas

Uno de los puntos de arranque más importantes del pensamiento renacentista es su rechazo al principio de finalidad: la búsqueda de una

explicación finalista es estéril, porque explicar, lo que se dice explicar, no explica nada. En palabras de Francis Bacon (1623: III, 5): *nam causarum finalium inquisitio sterilis est, et, tanquam virgo Deo consecrata, nihil parit*. Quizás sea verdad que el principio de finalidad es estéril para comprender *lo que pasa* en la naturaleza; sin embargo, es esencial para comprender *lo que debería pasar* en los artefactos humanos.

Suele decirse que el método científico-experimental y la perspectiva naturalista dejan fuera el arte, la metafísica, la ética y la religión. Pues bien, creo haber demostrado que también deja fuera la ingeniería, puesto que, repito, y perdonen la machaconería, es imposible entender una máquina sin entender su propósito, su finalidad, su función. Y, por cierto, también es imposible entenderla sin entender su estructura. Si algo hacemos los ingenieros es diseñar la finalidad y la estructura de las máquinas (Kroes, 2010). A mis alumnos de ingeniería informática lo que les enseño es eso: finalidad y estructura de las máquinas computacionales, de los sistemas de información.

Es obvio el paralelismo con la causa final y la causa formal de Aristóteles, que por cierto él mismo explica en un contexto tecnológico: el carpintero que fabrica una mesa de madera. No estoy empeñado en recuperar a Aristóteles, y me da igual si sus causas son tres, cuatro o cinco, o si se llaman de una forma o de otra. Ahí coincido también en que pelearse por las palabras no es demasiado sensato. Sin embargo, no deja de ser curioso, incluso paradójico, que se llame mecanicismo a una filosofía que nace y vive de prescindir de las causas finales, cuando no hay nada que menos se entienda sin las causas finales que un mecanismo. Un mecanismo, cualquier artefacto, no se entiende

sin entender para qué sirve. Como dice un gran escritor, cuyas palabras quizás recuerden:

> Para ver una cosa hay que comprenderla. El sillón presupone el cuerpo humano, sus articulaciones y partes; las tijeras, el acto de cortar. ¿Qué decir de una lámpara o de un vehículo? El salvaje no puede percibir la biblia del misionero; el pasajero no ve el mismo cordaje que los hombres de a bordo. (Borges, 1975).

No podemos comprender un artefacto si no comprendemos su finalidad. Si nosotros no estuviéramos aquí, si la civilización humana hubiera desaparecido, y llegaran unos extraterrestres, pero no extraterrestres humanoides, sino extraterrestres gelatinosos, y apareciera uno por el salón de mi casa, no podría entender el mobiliario, las sillas, mesas, sillones y estanterías, porque el mobiliario sólo se entiende si se conoce primero el cuerpo humano. Ni siquiera verían, propiamente hablando, cada uno de los elementos del mobiliario, como elementos diferenciables. *Para ver una cosa hay que comprenderla.* Y comprenderla es entender para qué sirve, su finalidad.

Así pues, las máquinas no se entienden sin la finalidad, es más, una máquina se define principalmente por su finalidad. Centrándonos en la teoría de la computación, este es un principio que está perfectamente asumido, y que fue además sentado por Alan Turing y los fundadores de las ciencias de la computación (Turing, 1936, 1948): un algoritmo (o *computación efectiva*) es un procedimiento basado en reglas que obtiene *un resultado deseado* en un número finito de pasos. Es decir, uno de los elementos esenciales de la definición de algoritmo es su finalidad, el objetivo o resultado deseado que tiene que perseguir (Hill, 2016).

Por lo tanto, una máquina que "hace cosas" no es propiamente una máquina (una vez más, no basta ser sistema mecánico para ser máquina). Si la máquina no hace lo que yo quiero, la máquina no funciona bien. Es más, si no sé lo que tiene que hacer la máquina, ni siquiera puedo saber si la máquina funciona bien o mal, no puedo someterla a control de calidad (que es otra de las tareas habituales de los ingenieros). Primero tengo que saber para qué está la máquina, su finalidad.

A mi modo de ver, lo interesante de este planteamiento es que la finalidad no es un componente físico de la máquina: la finalidad no *interactúa* con los otros componentes. La finalidad del reloj no interactúa con las agujas. La finalidad no ejerce ningún tipo de fuerza por contacto, ni a distancia tampoco, con los elementos de la máquina; y, por cierto, la estructura tampoco interactúa con ellos. Y, sin embargo, como ya hemos establecido, la finalidad y la estructura son esenciales para entender qué es una máquina en general, y qué es esta máquina en concreto. La finalidad (y la estructura) *influyen* en el funcionamiento de un artefacto, pero obviamente no *interaccionan* físicamente con sus elementos.

Algo análogo ocurre de modo muy particular en las máquinas computacionales, con su conocida estratificación en *hardware* y *software*, el nivel físico y el nivel lógico. El *software* no es causa eficiente del comportamiento del *hardware*. El *software* no es físico, luego no puede *interactuar* con el *hardware*. Y, sin embargo, ningún informático dudará lo más mínimo de que el *software influye* de modo muy real en el *hardware*. ¿Cómo es esa influencia? Usando un lenguaje muy antiguo, podemos decir que el *software* no es causa eficiente, sino causa formal: información, interpretación. El software es *el-*

nombre-que-damos-a lo que ocurre en el *hardware*: es una interpretación de las interacciones físico-mecánicas.

Pienso que esto arroja una interesante luz sobre el problema de la relación entre la mente y el cuerpo: por supuesto que la mente no interacciona con el cuerpo, en el sentido de las leyes físicas. Y, sin embargo, la mente es indispensable para entender el ser pensante entero. La mente no es un componente más del ser humano, sino su dimensión o aspecto intencional, sin el cual es imposible comprenderlo cabalmente, al igual que ocurre con la finalidad y estructura de una máquina.

6. Causa y explicación

En lenguaje aristotélico diríamos que la finalidad de un artefacto es *causa* de su funcionamiento. Si bien esta formulación puede ser aceptable, actualmente corre el riesgo de ser mal entendida. Puesto que la finalidad no es física, ni interactúa físicamente, no puede ser causa eficiente. Obvio para cualquiera que conozca el lenguaje aristotélico: la finalidad no es causa eficiente, sino, precisamente, causa final. Pero para quien no domine este lenguaje, la expresión puede volverse confusa. Si la finalidad no es causa eficiente, entonces no es causa de ninguna manera, como consecuencia de la eliminación moderna de la finalidad como categoría causal aceptable.

Por eso me parece más pertinente hablar de explicación que de causa, respetando por otra parte el sentido original del termino griego empleado por Aristóteles (αἰτία), que hoy día resulta oscurecido cuando se traduce unívocamente con el término latino "*causa*", en razón precisamente de la reducción

semántica de este último. Por tanto, en lugar de causa, es mucho mejor decir que la finalidad es la *explicación* del funcionamiento del artefacto. Dicho así, es mucho más fácil de entender, y transmite mejor lo que se quiere decir.

Por lo mismo, no debemos decir que la conciencia, o mente, es causa del comportamiento, en cuanto fenómeno observable, sino que lo explica. *La conciencia no es causa, sino explicación.* Es decir, no es causa en el sentido cartesiano, de causalidad eficiente, pero sí es causa en el sentido aristotélico, de causa final. Empeñarse en decir que la conciencia causa el comportamiento, muy en la línea dualista cartesiana, resbala peligrosamente hacia un terreno que no queremos pisar, es decir, conduce directamente al intento de convertir la conciencia en otro fenómeno natural, observable, en una "cosa" que supuestamente "interacciona" con esa otra cosa que llamamos cuerpo. En definitiva, el intento de naturalizar la conciencia. Todo este embrollo puede evitarse simplemente diciendo que la conciencia no es causa, sino explicación.

Lo cual da una interesante pista para entender *por qué la conciencia es inexplicada*: porque *la conciencia no se entiende como explicada* por otra cosa (menos aún, explicada mediante causas/explicaciones eficientes), *sino como explicación* intencional del comportamiento específicamente humano. Para entender la conciencia no hay que intentar explicarla, sino intentar entender qué es lo que ella misma explica, y cómo lo explica.

7. La finalidad en las máquinas y la finalidad en los humanos

Entonces, ¿cuál es la diferencia entre las máquinas y los humanos? Como dije al principio, ésta es una cuestión que puede abordarse desde la perspectiva de la enseñanza de la ética para ingenieros. De hecho, en mi experiencia, es muy útil hacerlo así para que los estudiantes de ingeniería, familiarizados con el concepto de máquina, entiendan mejor, por contraste, lo que significa "ser libre".

Lo característico de una máquina es precisamente que tiene finalidad, y una finalidad *dada*. En cambio, los seres humanos nos caracterizamos porque podemos proponernos nuestros propios fines. Podemos proponernos planes, proyectos, objetivos vitales; *podemos decidir qué queremos ser en la vida*. No somos producto exclusivamente de nuestra biología, ni tampoco de nuestra educación: tenemos fines, pero no estamos "programados". Podemos rebelarnos, no somos esclavos de nuestra herencia genética o social. Esto, por cierto, es una idea muy interesante para ser explotada cuando uno habla con estudiantes, que a menudo aceptan el tópico de que la ética es una imposición social asumida inconscientemente: ¿Quién no se ha rebelado contra sus padres, contra su educación, contra la herencia recibida? Siempre diciendo que estamos sometidos-a. ¿Sometidos-a? Pero, ¿quién no se ha rebelado?

Por el contrario, una máquina (y en particular una máquina algorítmica o computacional) no puede decidir qué objetivos quiere perseguir, porque dejaría de ser una máquina. Que es precisamente lo que les pasa a los robots que se hacen humanos en la ciencia ficción: los replicantes de *Blade Runner* (Scott, 1982) ya no son robots, son humanos, aunque sea en una forma de

humanidad que nos desconcierta y no sabemos precisar bien. Hay otras películas que tratan este tema, en las que la humanización del robot no es tan radicalmente completa. Por ejemplo, en *Inteligencia Artificial* (Spielberg, 2001) el protagonista es un niño-robot, David, que "vive" continuamente movido por un objetivo que otros han programado en él: necesita ser querido, necesita el afecto de una "madre", y no descansará hasta encontrarlo. Este comportamiento quizás sí podría ser realizado por una máquina; al menos yo no encuentro una objeción tan fuerte en su contra. En cambio, un ente que se cuestiona, como el replicante Roy Batty, por qué estoy aquí, cuánto tiempo me queda, y no me da la gana de hacer lo que me dice mi creador, este ente no es ni puede ser un robot programado. Por el contrario, es un ente que ha completado su transición a la humanidad, un ente que se usa en la película como metáfora de nosotros mismos. Por lo tanto reivindico la desobediencia como característica esencial de lo humano: el ser humano es "desobediente" en el sentido de que es autónomo, no heterónomo; en el sentido de que no está completamente controlado por una ley que no es él mismo.

Quizás en un futuro indeterminado seamos capaces de producir en el laboratorio un tipo de robots no algorítmicos (no dirigidos hacia un fin dado) que propiamente puedan ser calificados como autoconscientes, capaces de hacer "lo que les dé la gana", de proponerse sus propios objetivos; pero seguramente no sería adecuado seguir llamándolos robots. Serían "humanos" en el sentido de autodeterminados, verdaderamente libres, aunque quizás su estructura física (¿biológica?) fuera muy diferente a la nuestra. Pero, ¿de qué serviría esto? ¿Para qué fabricar máquinas que no harán lo que queremos,

sino lo que les dé la gana? ¿En qué sentido puede decirse que siguen siendo máquinas? No digo que sea imposible fabricar estos seres "replicantes", sino que no serían propiamente "máquinas". Serían entidades con *conciencia artificial* (fabricada), pero no serían *máquinas computacionales*, puesto que no estarían gobernadas por una finalidad dada y extrínseca. La autodeterminación, por su misma definición, queda fuera del paradigma computacional clásico y del concepto general de máquina.

8. Conclusión

Me gustaría concluir con un juego de palabras y una reflexión sobre el fin último de la vida. *¿Cuál es el destino de los novios?* Si esta pregunta la formulo en inglés, puedo elegir dos formas: *What is the fate of the just-married*, a qué están abocados. O también, *What is the destination of the just-married*, a dónde van a ir, a dónde han decidido ir. Lo cual me sirve para ilustrar que nuestra finalidad no nos es simplemente dada, para que ahí nos la encontremos, como si fuéramos máquinas. *Tenemos un destino, pero también nos forjamos nuestro destino*.

Una de las constantes preguntas que se plantea la humanidad es acerca de su destino. ¿Tiene el ser humano (individual y comunitariamente) un destino? El antiguo fatalismo griego pretendía que el destino humano estaba controlado por los dioses del Olimpo. Esta tendencia resurge hoy día, atribuyendo el control a los genes: somos, en el fondo, esclavos de nuestra programación biológica. Contra la tendencia fatalista se rebela la afirmación radical de la libertad humana, la afirmación de que lo característico de los

seres humanos es que nos proponemos nuestros propios fines, decidimos lo que queremos ser. Considero que esta capacidad de autoproponerse los fines es lo más característico de la diferencia entre humanos y máquinas, y por lo tanto de lo que podría y no podría hacer una supuesta conciencia artificial. Son precisamente los que no se atreven a afirmar radicalmente la libertad los que más fácilmente caerán en la tentación de considerar que los humanos no son en último término otra cosa que complicados robots biológicos.

Bibliografía

J. Arana, *La conciencia inexplicada*, Madrid, Biblioteca Nueva, 2015.

F. Bacon, *De Augmentis Scientiarum*, 1623.

J. L. Borges, "There are more things", en *El libro de arena*, Buenos Aires, Emecé, 1975.

G. Génova, M. R. González, "Cuatro problemas del método científico-experimental que reclaman la apertura a la inteligencia meta-metódica". In Manuel Oriol (ed.), *Inteligencia y filosofía*. Marova, 2012: 661-680.

—, "Teaching Ethics to Engineers: A Socratic Experience". *Science and Engineering Ethics*, 22(2): 567-580, April 2016.

G. Génova, "¿Para qué sirve la filosofía a los ingenieros? Descubriendo el principio de finalidad en los artefactos". *X Jornadas de la Asociación Española de Personalismo. ¿Qué es filosofía? ¿Y para qué sirve?* Madrid, 4-6 de mayo de 2016. Universidad Francisco de Vitoria.

G. Génova & I. Quintanilla, "¿Are Human Beings Humean Robots?", enviado para revisión, 2016.

R. K. Hill, "What an algorithm is". *Philosophy & Technology*, 29(1): 35-59, 2016.

W. Isaacson, *Steve Jobs*. Madrid, Debate, 2011.

P. Kroes, "Engineering and the dual nature of technical artefacts". *Cambridge Journal of Economics*, 34(1):51-62, 2010.

I. Quintanilla, *Techné: la filosofía y el sentido de la técnica.* Madrid, Common Ground España, 2012.

R. Scott, *Blade Runner*, Warner Bros., 1982.

S. Spielberg, *Inteligencia Artificial*, Warner Bros., 2001.

A. M. Turing, "On computable numbers, with an application to the Entscheidungsproblem", *Proceedings of the London Mathematical Society*, 2(42): 230-265, 1936.

—, "Intelligent Machinery". National Physical Laboratory Report, 1948. In Meltzer, B., Michie, D. (eds), *Machine Intelligence* 5. Edinburgh University Press, 1969.

Gonzalo Génova Fuster
ggenova@inf.uc3m.es