¿Quién puede mentirle a un robot? Roboética e imperativo categórico

Ignacio Quintanilla Navarro, Comunidad de Madrid, España

Resumen: *Este trabajo analiza la incidencia que nuestro contexto tecnológico tiene en la evolución y el contenido de nuestra argumentación moral. Las principales referencias históricas con las que se dialoga, desde un caso concreto de roboética, son Kant y Scheler. En una primera parte, se propone un esbozo sistemático e histórico de la interacción entre tecnología y argumentación ética. La segunda parte presenta un ejemplo preciso: el desarrollo de la argumentación kantiana sobre la mentira expuesta en el opúsculo de 1797 "Sobre un presunto derecho a mentir por filantropía" en el contexto de la interacción ser humano-robot. En la tercera parte se hacen algunas consideraciones sobre un programa general de investigación en roboética.*[1]

Palabras clave*: mentira, inteligencia artificial, inteligencia objetiva, ética formal, imperativo categórico, empatía*

Abstract: *This paper analyzes the impact of our new technological system in the evolution and structure of our ethical argumentation. Considering the possibility of the deception in the relationship human-robot, the paper discusses some of the implications of the ethical doctrines of Kant and Scheler in a situation of human-robot communication. After a historical analysis of the relationship between technology and ethical argumentation the paper proposes an interpretation of the moral argument exposed by Kant in his essay "On a Supposed Right to Lie Because of Philanthropic Concerns", published in 1979, in the context of the interaction human being-robot. The final section offers some suggestions on general roboethics.*

Keywords: *Roboethics, Deception, Artificial Intelligence, Objective Intelligence, Formal Ethics, Cathegorical Imperative, Empathy*

I

Al hablar de ética y tecnología pensamos habitualmente en la evaluación de algún tipo de riesgo tecnológico; riesgo físico, político o moral. Esta evaluación es imprescindible, pero también insuficiente para comprender el alcance de la relación entre ética y tecnología y elude algunos de sus aspectos filosóficamente más interesantes. El riesgo, por otra parte, es inherente a la convivencia y a la condición humana. Asumir riesgos –en ocasiones enormes y fatales– de manera más o menos compartida ha sido siempre un aspecto esencial de nuestra realidad biográfica, social e histórica. Por eso, para una perspectiva filosófica, en el cambio tecnológico hay otros temas fundamentales además de la mera suma de riesgos de tales o cuales usos concretos de la ingeniería genética o de las nuevas TIC.

En una de sus escasas incursiones en el terreno de la técnica Popper hace esta llamativa observación incidental: *cada invento nuevo refuta una cosmovisión* (Popper, 1994: 14). Y es que, en efecto, al hacer reales formas posibles de existencia, el cambio tecnológico involucra y cuestiona permanentemente lo más fundamental de nuestras cosmovisiones e ideas. Pretender que la ética queda fuera del rango de esta afirmación es un prejuicio peligroso. No hay una ética *previa* a la tecnología con la que evaluar y encauzar una tecnología *posterior* a la ética. Al menos, no de entrada y si por

[1] En 2007 el Ministerio de Comercio, Industria y Energía de Corea del Sur creó el primer departamento gubernamental de roboética. Un buen estado de la cuestión roboética puede verse en la página del Etiske Råd (Consejo danés de ética): http://etiskraad.dk, dentro del proyecto *Homo-Artefakt*. Cf. también Veruggio, 2005; Veruggio y Operto, 2010; Lin, 2012; Moon, 2012.

Revista Internacional de Tecnología, Ciencia y Sociedad
Volumen 3, Número 1, <http://tecnociencia-sociedad.com>, ISSN 2530-4895

Republicado de *Revista Internacional de Tecnología, Conocimiento y Sociedad 3*(1), 2014 (pp. 1-13)

ética entendemos una forma de saber humano.[2] La ética es ya también efecto de la tecnología, tanto como la tecnología lo es de la ética, y aquí está, precisamente, lo más interesante de la cuestión.

El programa de investigación en el que se inscribe este texto se resume en la siguiente pregunta: ¿qué novedades esenciales introduce en la estructura de nuestra argumentación ética el actual contexto tecnológico? Pero valorar el alcance de esta pregunta requiere considerar, en primer lugar, una dimensión argumental y, en segundo, una dimensión histórica del problema.

Por lo que se refiere a la dimensión argumental es preciso distinguir entre cuatro niveles de argumentación ética sobre la tecnología. La argumentación *tipo 1* o *instrumental* plantea la preferencia ética entre opciones tecnológicas concretas desde una perspectiva realmente técnica y no propiamente axiológica –del tipo, por ejemplo, si seguimos comiendo carne animal con tantos antibióticos nuestras propias bacterias se harán resistentes y nuestros tratamientos médicos serán ineficaces. Aunque este nivel de debate puede implicar cuestiones morales concomitantes –como la grave responsabilidad ética que tiene el ciudadano de implicarse en el debate sobre nuestros los modelos de desarrollo– se trata, en principio, de falsas cuestiones éticas. En efecto, nadie defiende, en ese caso, que los tratamientos médicos humanos se vuelvan más ineficaces. La discrepancia está en si, de hecho, nuestros hábitos de alimentación y/o los intereses de la industria cárnica suponen o no un riesgo grave para la nuestra salud. Pero esto no constituye un problema moral sino epistemológico y/o técnico. El trabajo duro en este nivel de argumentación le corresponde, pues, más bien, al científico, al técnico y al periodista.

La argumentación *tipo 2* es la que podemos denominar *de aplicación* o *clásica* y se refiere a los casos en los que la novedad técnica no exige realmente una novedad en la argumentación ética sino que moviliza viejos recursos éticos en nuevos contextos de aplicación. Por ejemplo, para determinar si es justo que el beneficio de un accionista de industrias cárnicas ponga realmente en peligro la salud del género humano basta apelar a la prioridad del bien común sobre el beneficio particular. Una prioridad que está ya bien consolidada en nuestra cultura desde muchos siglos antes de que surgiera la preocupación medioambiental. En este caso, la novedad técnica exige trabajo ético pero no novedad o creación ética. El interés filosófico de las cuestiones tipo 2 no siempre es significativo y suele ser más bien al político, al educador o al jurista a quienes les toca bregar aquí.

La argumentación *tipo 3* es aquella en la que la novedad técnica suscita un problema moral que sí exige una genuina novedad ética, ya sea en los conceptos básicos implicados, ya en la estructura misma de la argumentación, y, en consecuencia, la innovación técnica remite a una innovación ética real.[3] Veremos un ejemplo en este escrito. La argumentación *tipo 4*, finalmente, es la que trata de realizar una valoración global del cambio tecnológico de una sociedad en su conjunto o, incluso, de la historia entera de la técnica humana. Aquí encontraríamos las grandes interpretaciones teóricas sobre sentido de la técnica al estilo de un Martin Heidegger, Ortega y Gasset o Lewis Mumford.

Estos cuatro niveles de argumentación son todos de vital importancia para nuestra sociedad y la frontera entre ellos no siempre está clara. Sin embargo, una correcta distinción entre estos cuatro planos de debate parece necesaria por muchas razones, y, a nuestros efectos, porque la existencia misma de las argumentaciones tipo 3 y 4 sería cuestionada por algunos importantes enfoques éticos. Además, parece evidente que el lugar de una *roboética* que estuviera realmente justificada como disciplina filosófica sería el de las argumentaciones tipo 3.

Este trabajo defiende, pues, la importancia de prestar más atención a las argumentaciones del tipo 3 y aporta un ejemplo, entre otros posibles, de relevancia de las mismas. Y es aquí donde debe hacerse

[2] Se asume la paradoja de que el formalismo ético rechazaría, de entrada, esta afirmación. Veremos seguidamente hasta qué punto la vigencia del propio formalismo kantiano puede tener que ver con nuestro cambio tecnológico. En este sentido, hay que destacar la importancia de una aproximación ética al propio diseño tecnológico, muy descuidada en relación con la del beneficio, uso y riesgos. Cf.: Mitcham y Duval, 2000.

[3] Puede darse, por supuesto, la situación inversa en que una genuina innovación ética suscite innovación técnica, o que ambas innovaciones, técnica y ética, sean aspectos parciales de alguna otra novedad en otro nivel de la cultura. Cf. Quintanilla, 2006: 147.

nuestra segunda consideración, de carácter histórico. En lo que sigue no se pretende justificar un determinismo tecnológico. La revisión del cambio tecnológico humano sugiere que nuestras ideas condicionan nuestra técnica tanto como la técnica nuestras ideas. Por otra parte, el *sociologismo* en materias de argumentación moral es a menudo un error metodológico y casi siempre una reducción parcial de los problemas. Dicho esto, sin embargo, debemos destacar que los rasgos fundamentales de la argumentación moral en Occidente se corresponden con más claridad histórica con las principales fases de nuestro desarrollo tecnológico que con cualquier otra dimensión relevante de nuestra cultura, incluida la propia argumentación política, metafísica, religiosa o científica (Quintanilla, 2012: 71).

La socialización del afecto, en general, y de la violencia en particular y su conformación simbólica y ritual parece ser la primera concreción de la moralidad en las comunidades humanas. Este proceso de asentamiento de una afectividad social común y una estimativa compartida - que sigue siendo fundamental en nuestras sociedades postmodernas - caracteriza lo que podemos considerar como primer estadio de nuestra deliberación moral y se corresponde con el primer estadio humano de evolución tecnológica: el de la *técnica prehistórica,* previa a la *polis*. La identificación de la conciencia individual con la tradición del grupo y la íntima imbricación, a su vez, entre moralidad, religión, y un ecosistema muy concreto, hacen que una reflexión moral explícita, en forma de ética, tal y como la concebimos desde sus orígenes griegos, esté virtualmente ausente en estas cosmovisiones que Platón –por ejemplo en el *Protagoras*– nos presenta como un estadio ético-político *primitivo* que la *politeia* debe superar.[4]

Hacia el 3500 a.C. surge una nueva fase tecnológica humana –el período *clásico* de la civilización– asociada al asentamiento de grandes núcleos de población estables con estructura social compleja. Nace, en suma, la *polis* y, allí donde las circunstancias favorecen la deliberación entre iguales como instancia de resolución, una reflexión racional sobre la *praxis* del ciudadano que constituye la *ética* en el sentido más propio de nuestra tradición. La figura de Sócrates encarna claramente este tránsito de la moral simbólico-ecológica de la tribu a la moral autónomo-dialógica del ciudadano, con lo que, a fin de cuentas, la acusación formulada en su día contra él –la de introducir *dioses extraños*– se revela rigurosamente exacta. Sócrates descubre cuatro niveles de argumentación práctica que a sus contemporáneos –y a los nuestros– les costaba distinguir: el de lo que es moralmente *bueno*, el de lo que tiene *legitimidad* política o ética, el de lo que es *legal* en un momento dado y, finalmente, el de lo que es socialmente *correcto*. El fascinante terreno de indagación que abre esta cuádruple distinción, al que asistimos maravillados todavía hoy al leer los diálogos platónicos, inaugura una etapa nueva en nuestra reflexión ética. Es la etapa *clásica* que se estructura en torno a la *praxis* o realización social del ciudadano mediante un proyecto individual y colectivo de *vida buena*.[5]

Con todo, las virtudes y los hábitos que conforman –como una *segunda naturaleza* - esta vida buena siguen estando íntimamente vinculados a la contemplación, comprensión e interiorización biográfica de cierto orden natural e inmutable de cosas (*natural* e *inmutable* son todavía esencialmente sinónimos) que debe entenderse y aceptarse. Vale decir que, en este segundo estadio de argumentación moral, la ética sigue siendo el corolario práctico de la naturaleza y mantiene un carácter claramente "aristocrático" en el que el humano mejor en un sentido natural y el humano mejor en un sentido ético coinciden. La vigencia de este esquema de argumentación –que se puede apreciar tanto en Platón como en Nietzsche– es fácil de constatar.[6] A lo que no se presta suficiente atención es a que, nacido de la madurez racional de la polis antigua, el final del monopolio de este

[4] *Protágoras*, 320c-322d. Este nivel de praxis humana –en el que la distinción entre *comunidad* y *organización* resulta virtualmente inexistente– facilita enormemente la integración afectiva y funcional del sujeto en su contexto natural y social. De aquí la fascinación que ejerce sobre el discurso moderno –que entronca con nuestra tradición cínica– la figura del salvaje y su estilo de moralidad. En el salvaje, la plena integración del individuo en la comunidad y en un hábitat natural coinciden plenamente, es decir, que deber, tradición y "ecología" (siempre en un sentido parcial e idiosincrático) son virtualmente indistinguibles. Este estrato de la moralidad nunca deja de estar vigente en una sociedad y puede contemplarse, de hecho, el surgimiento de nuestra actual cultura ecológica como la búsqueda de un nuevo hábitat o ecosistema que se ajuste a nuestras nuevas formas de moralidad.

[5] El estudio más influyente de este estadio de la praxis y su vigencia actual es el de Arendt, 1958.

[6] Un trabajo clásico en este sentido es MacIntyre, 1981.

estadio de argumentación moral –que también pervive hoy, lógicamente, como un segundo estrato de moralidad– coincide con la consolidación de Revolución Industrial.

En efecto, este esquema *clásico* de la ética sufre su primera rectificación esencial en la obra de Kant. La ética de Kant, y su apelación radical a una buena voluntad individual como origen de toda moralidad genuina, integra por vez primera en el terreno de la teoría moral de Occidente una dimensión rigurosamente personal de la acción humana que la tradicional ética de las virtudes no podía asumir en ninguna de sus variantes (Paton, 1971). Los motivos teológicos y políticos de esta radical transformación del discurso ético parecen bastante claros. Pero, junto a estos motivos, es un hecho que la ética kantiana coincide con una revolución –la Revolución Industrial– cuya impronta es, desde luego mucho más difusa, pero también mucho más crucial que la de la propia Reforma o la Revolución Francesa. La filosofía de Kant es un gozne filosófico entre dos grandes etapas de la condición tecnológica humana y lo que mejor delimita esta inflexión es el hecho de que la noción de *naturaleza* deja de funcionar como sin embargo se sigue esperando que funcione:

> Aun cuando por particulares enconos del azar o por la mezquindad de una naturaleza madrastra le faltase por completo a esa voluntad la facultad de sacar adelante su propósito… sería esa buena voluntad como una joya brillante por sí misma…[7]

Cuando Kant presenta su obra como una gran *revolución copernicana* en el mundo de la filosofía está expresando una profunda verdad. Los puntos de tensión argumental fundamentales en la *Crítica de la razón pura* siguen siendo, a fecha de hoy, los puntos de tensión cruciales en nuestra cultura y se resumen en dos novedades que, desde una perspectiva *tecnológica*, aparecen perfectamente claras: la incapacidad práctica de la noción moderna de *Naturaleza* para fundamentar la racionalidad de nuestra acción moral y la incapacidad teórica de la noción moderna de *Naturaleza* para justificar la racionalidad de nuestro conocimiento científico. Es decir, el orden natural del universo pierde tanto su tradicional función moral, avalando la praxis humana, como su función teórica primordial dando sentido a la realidad y justificando su carácter intrínsecamente racional.[8]

Mientras Kant escribía sus últimas obras, Napoleón Bonaparte movía sus ejércitos por Europa más o menos a la misma velocidad que lo hacía Julio César. Pero eso estaba a punto de acabar. En los continentes europeo y americano los elementos fundamentales, tanto en tecnología como en ética, inalterables desde la polis griega, acabarían hacia finales del siglo XVIII. Las generaciones posteriores a Kant tendrán que afrontar esta *revolución copernicana* y ensayar, desde esta nueva situación, diversas estrategias; ya sea buscando una nueva racionalidad desde lo histórico-social, ya asumiendo con más o menos rigor una ingenierización de todos los aspectos del espíritu; es decir, basando la cultura entera en la tecnología punta disponible –que era la mecánica–[9] o ya, sencillamente, explorando el fin de la racionalidad occidental tal como se concibe desde Sócrates o Platón.

El idealismo alemán, tanto como el positivismo, el irracionalismo o el marxismo serán, pues, filosofías de la etapa industrial y, como tales, tendrán que afrontar sus grandes fenómenos adyacentes. Fenómenos como el surgimiento de la *cultura de masas*, es decir, de amplios grupos humanos de productores-consumidores insertos en un sistema de procesos y productos

---

[7] Kant, 1785, Cap. 1, BA p.3.

[8] Por ello, las dos novedades más llamativas de la obra de Kant son su esfuerzo por colocar el ámbito de la reflexión ética al margen de la naturaleza –humana o cósmica– y, por tanto, de cualquier clase de experiencia moral natural, y su negativa a aceptar la posibilidad de un saber científico de la cosa en sí. Esta segunda tesis, por cierto, la del *idealismo trascendental*, comporta, en la obra de Kant, la negación de la posibilidad un conocimiento científico de "lo artificial" o de una genuina ciencia de lo artificial en un titánico , y por ahora exitoso, esfuerzo por preservar un paradigma epistemológico griego, a saber, el que circunscribe el ámbito del conocimiento objetivo al ámbito de *lo natural*, solamente hay verdadero saber objetivo de lo natural, y por tanto no puede haber genuina ciencia de lo artificial. La primera crítica a este planteamiento la encontramos en Simon, 1969.

[9] La figura de Franz Reuleaux encarna paradigmáticamente ese colosal proyecto decimonónico de mecanizar la totalidad de la cultura.

estandarizados; la primacía de la organización sobre la comunidad como modelo de interacción social; la preponderancia de las categorías *de trabajo, producción o creación* productiva en todos los contextos –económico, estético, científico etc.– con secuelas tan inevitables como la mística del "genio" y/o del "triunfador social" –que desbanca al "sabio", al "santo" y al "héroe"–; o, finalmente, la prevalencia del un modelo de la *lucha y competición* como paradigma de interacción social humana.[10]

La crisis de las ciencias europeas de Husserl, como tantas otras obras a caballo entre el siglo XIX y el XX, recoge, así, la confrontación entre la concepción de la inteligencia y la que caracteriza la cosmovisión clásica y la nueva situación de una cultura que se encuentra con que *la naturaleza no es, para el hombre, más que el punto de partida que tiene la responsabilidad de modificar.*[11] El impacto de este nuevo estatus tecnológico determina, así, categorías y argumentos ético-políticos todavía vigentes.[12]

Ahora bien, si la situación radical del hombre moderno interioriza el drama de quien todavía espera de la naturaleza unas funciones lógicas y éticas que la naturaleza ya no puede cumplir, la del estadio post-moderno interioriza el *vértigo-serenidad* de la renuncia definitiva a esta demanda. Ni la naturaleza ni la historia, como gran metarrelato que prolonga el devenir de la naturaleza en la cultura –o el retorno a ella tras alguna escisión traumática–, se muestran ahora capaces de habilitar una racionalidad generalmente compartida. Sucede todo ello en torno a las dos grandes guerras y en torno la segunda, precisamente, asistimos al surgimiento de una nueva situación tecnológica –la etapa postindustrial o "tercera ola"– definida por el acceso a energías y materiales virtualmente inagotables, por la incorporación de la genética y la biología celular al entramado de producción tecnológica y, sobre todo, por el surgimiento de las nuevas TIC. El ideal mismo de una experiencia básica común para todo el género humano, así como de un horizonte ontológico inflexible, se desactiva o pasa a tener que conquistarse mediante complejos ejercicios argumentales previos. Como señala Lewis Mumford, desde comienzos del siglo XX asistimos a la interiorización del nuevo paradigma existencial de que *ya todo es posible*.[13]

La reflexión ética apenas ha comenzado a afrontar explícitamente la incidencia de esta situación en el núcleo de su argumentación y a explorar el alcance de sus implicaciones éticas. Una de ellas es la que queremos abordar ahora con un ejemplo concreto: la producción de un sistema artificial en cuya interacción con nosotros sea posible la mentira. Esta situación, además de suponer una alternativa interesante al test de Turing, en el plano teórico, obliga a repensar nuestra propia condición de agentes morales y la estructura misma de la argumentación ética.

II

En 1797 Benjamin Constant publica uno de sus primeros escritos: *Des réactions politiques,* donde, entre otras cosas, critica la tesis de que tengamos la obligación absoluta de decir siempre

[10] Tampoco puede olvidarse que en la fase industrial de la argumentación ética la categoría de *novedad* aparece como una cualidad intrínsecamente positiva que moviliza la justificación ético-política en torno a los ejes de *progreso* y a tareas de *liberación*. Sin embargo, que aunque esta asimilación intrínseca entre novedad, progreso y bondad se lleva a efecto, principalmente, en la mentalidad colectiva de los EEUU desde los comienzos del siglo XIX, en Europa no generalizará en la población hasta comienzos del siglo XX.

[11] Hegel, 1817: I, 95.

[12] No sólo en la literatura filosófica, sino en la literatura en general. Motivos como el desarraigo, la ausencia del padre, la falta de hogar al que volver o la ausencia de parámetros narrativos en los que habilitar la figura de un héroe, refieren directamente nuestra cultura oficial –al menos la académica– a la fase industrial de la tecnología.

[13] Un efecto de esta nueva situación tecnológica es que la humanidad entera, sin distinción de culturas, comunidades políticas o sociedades, pasa a vivir inmersa en un único vector común de transformación técnica. Además, por primera vez en la historia, el consumo privado del ciudadano medio en tiempos de paz se convierte en un motor de cambio tecnológico de mayor entidad que la guerra o el consumo propio de instituciones políticas, religiosas o industriales. El hogar moderno y la vida cotidiana se convierten en el entorno en el que se toma contacto directo con el cambio tecnológico. También, por primera vez en la historia de la ética, los principales problemas de moralidad pública y privada se definen y se dilucidan colectivamente como aceptación o rechazo de opciones tecnológicas concretas (Quintanilla, 2012: 203).

la verdad, sostenida por cierto filósofo alemán a quien no cita. Constant considera que semejante principio es moralmente irrealizable e incompatible con una sociedad viable y defiende la alternativa de que sólo tenemos la obligación de decir la verdad a quien es merecedor de ese derecho. Aunque Constant no se refería precisamente a Kant, Kant se dio por aludido pues la tesis del deber incondicionado de decir siempre la verdad había sido ya expuesta, por ejemplo, en pasajes muy significativos de la *Fundamentación de la metafísica de las costumbres*.

Kant recoge, pues, el guante y publica ese mismo año uno de sus más celebres opúsculos ético-jurídicos: *Sobre un presunto derecho a mentir por filantropía*. La refutación Kantiana de las tesis de Constant tiene varios niveles de argumentación que, en principio, no nos interesan.[14] En definitiva, la tesis kantiana es que la veracidad es un deber formal de la comunicación humana con independencia del perjuicio o de las consecuencias positivas que de ella se deriven. Al mentir dañamos, en nuestra propia persona, a toda la humanidad en su conjunto porque inutilizamos la fuente misma de cualquier pacto, contrato o acuerdo humano; en realidad la fuente de la propia comunicación y de la vida social, en lo que supone una contradicción directa de la norma moral consigo misma.[15]Se trata, pues, de un desarrollo –con algunas variaciones– de la argumentación ya expuesta al final del capítulo primero de la *Fundamentación de la metafísica de las costumbres.*

Ahora bien, si releemos la argumentación de Kant desde el punto de vista de una filosofía de la técnica nos encontramos con algunos motivos y argumentos cuya actualidad debe subrayarse. A mi juicio son fundamentalmente dos. El primero es la extraordinaria coherencia con la que Kant justifica –sin conciencia de ello, claro está–, que la cualidad moral del diseño e innovación tecnológicos no puede fundamentarse en la previsión de sus consecuencias. El consecuencialismo moral resulta tan inaceptable en el ámbito del cambio tecnológico como en cualquier otro ámbito de acción humana susceptible de valoración moral. Una evaluación ética de la novedad tecnológica sólo es factible, para Kant, desde el ejercicio de una racionalidad pura, esto es, *a priori* y en virtud de principios o máximas que se apliquen de antemano en la propia acción innovadora.

En primer lugar porque, como señala Kant respecto de la mendacidad, *a pesar de cuanta astucia me precie de tener, las consecuencias no son tan fácilmente previsibles que no pueda suceder…*[16]. En segundo lugar, porque sólo la acción que tiene el propio deber por motivo posee un verdadero significado moral. La prudencia, por tanto, por mucha astucia y previsión que ejercite, no puede legitimar jamás un uso mendaz de la información y, por idénticas razones, tampoco está en condiciones de fundamentar *–de facto* y *de jure–* la valoración moral de la producción o empleo de ningún nuevo artefacto. Valdría decir de ningún *otro* artefacto si consideramos la propia mentira, a fin de cuentas, como un artefacto comunicativo.

Cuando se contempla la naturaleza y rapidez de nuestro cambio tecnológico se constata la imposibilidad creciente de abarcar en nuestros cálculos todos los efectos posibles que nuestra innovación puedan tener sobre el conjunto de la realidad. Ello sucede ya, además, con carácter irreversible. Esta constatación fáctica proporciona ahora una interesante perspectiva sobre las virtualidades del formalismo ético en nuestro contexto tecnológico. Parece cada vez más plausible que sólo un enfoque kantiano de evaluación moral pueda proporcionarnos estrategias para habilitar una racionalización moral compartida sobre nuestro cambio tecnológico. En consecuencia, la normativización *a priori* en función de imperativos podría ser, en este momento, la estrategia argumental más sólida para orientar el cambio tecnológico de todas las habilitadas por nuestra tradición ética.

Esta situación constituye, por sí misma, una novedad importante en nuestro nivel de argumentación tipo 3. Ahora bien, esta situación general de la argumentación moral frente al cambio tecnológico se hace especialmente patente en lo que se refiere a la ética y diseño de cualquier

[14] Como la crítica al concepto de *derecho a la verdad* o la incoherencia en el uso que Constant hace de su noción de *principio intermediario,* en el contexto de una interpretación de la Revolución Francesa.

[15] *La no veracidad es la lesión del deber para consigo mismo*, indica Kant en nota a pie, apartando, de momento, el desarrollo de esta idea en aras del carácter jurídico que quiere darle al texto. Kant, 2006: 62.

[16] Kant, 1785, Cap. I, BA p. 18.

sistema de inteligencia artificial. Para empezar, es legítimo deducir de lo anterior que no hace falta esperar a ver qué cosas es capaz de hacer un robot en tal o cual fecha para establecer los principios fundamentales de la roboética. Para seguir, es posible que el enfoque ético kantiano sea susceptible de concreción algorítmica real en la programación de un sistema de inteligencia artificial.

En efecto, se ha llamado ya la atención sobre el hecho de que, al menos la primera formulación del imperativo categórico, posee una virtualidad algorítmica que la podría hacer efectivamente integrable en la programación de un robot.[17] Esta posibilidad es la que nos introduce en la consideración del segundo de nuestros motivos de reflexión.

Este motivo es la constatación de que la condición de objeto-sujeto moral se amplía y complica enormemente en el esquema de eticidad formal que Kant nos propone, cuando lo desarrollamos en un contexto de inteligencia artificial. Como es sabido, cuando una persona no actúa *por* deber, sino sólo *conforme* al deber, no actúa, según Kant, como un verdadero sujeto moral sino con una motivación patológica que, a todos nuestros efectos, nos equipara al funcionamiento de cualquier otro sistema natural (o artificial). Además, cuando mentimos, tratamos al interlocutor como simple medio, como un objeto, y no como una verdadera persona (valga decir, de momento y hablando, como se verá, impropiamente, que robotizamos al interlocutor). Pero, sobre todo, debemos prestar atención, a la hipótesis kantiana subyacente de que, allí donde se da una situación de interacción racional o comunicativa, surge la responsabilidad moral con independencia de que alguno de los agentes implicados tenga la condición de sujeto moral o agente libre propiamente dicho.

En efecto, un supuesto operativo en la argumentación kantiana que nos ocupa es que los fenómenos de *simpatía, compasión* o *percepción de otro yo ajeno como correlato psicológico-emocional* del *yo propio* no son inherentes al fenómeno del comportamiento moral en relación a la mentira. Además, en el formalismo kantiano, la condición de agente moral es totalmente independiente de cualquier intuición, convicción o presupuesto trascendental acerca de la existencia o índole de cualquier *naturaleza* y es, por tanto, anterior a esta. En otros términos: la pregunta: *¿es natural?* no desempeña ningún papel lógico en la determinación de nuestra conducta moral con relación a ningún agente. Ningún argumento sustancial en la argumentación kantiana impediría, por ejemplo, atribuir a un sistema artificial el ejercicio de una voluntad política que, de hecho, ya dejamos en manos de tribunales cuyos protocolos de toma de decisiones pueden formalizarse y distanciarse *ad infinitum* del ejercicio actual de decisiones humanas puntuales.

De este modo, para refutar el argumento de la legitimidad condicional de la mentira, expuesto por Constant, Kant establece tres principios. El primero es que el ejercicio de la mentira es inaceptable ya en nuestra propia y exclusiva persona con total independencia de nuestro interlocutor. El segundo es que el sujeto perjudicado en toda mentira es la humanidad entera, al margen de la materia y situación que la originen, pues, como se ha indicado, la mentira daña a la humanidad entera en nuestra propia persona. El tercero, finalmente, es que, en la determinación del carácter inmoral de la mentira, la figura de un interlocutor engañado o de un tercero perjudicado son realmente irrelevantes porque, en realidad, lo que se infringe es un deber sagrado de la razón para consigo misma. Conviene no perder de vista esta idea cuando, un poco más adelante, traigamos a colación al *entendimiento agente* aristotélico como una de las versiones clásicas más célebres de lo que sería una *inteligencia objetiva*.

Kant intenta hacer más asequibles todas estas tesis con un célebre ejemplo. Un asesino pregunta a alguien por su víctima y ese alguien trata de evitar el crimen con una mentira. Dice que la posible víctima no está en su casa cuando sí lo está. Si, como consecuencia de esa mentira, el crimen tiene lugar pese a todo –por ejemplo porque la víctima, que ha escapado inadvertidamente de la casa, se topa con el asesino en la calle–, el mentiroso es también, piensa Kant, corresponsable del asesinato. Por el contrario, si el interrogado dice la verdad y la víctima se encuentra en la casa y muere por

[17] Cf. Powers, 2011. Debe consignarse aquí la distinción convencional entre *Machine Ethics* (o ética de la maquina) y *Roboetics* (roboética). La primera se refiere al eventual comportamiento moral de máquinas inteligentes, la segunda, a la moralidad en el comportamiento de los seres humanos cuando diseñan o interactúan con máquinas inteligentes. Las reflexiones en torno a la denominada *Friendly artificial intelligence*, guardan una estrecha afinidad con la primera en su tarea de promover una inteligencia artificial benevolente con el género humano.

ello, el declarante veraz no tiene responsabilidad moral alguna ni tampoco jurídica ante ningún tribunal. Además siempre es posible que la víctima haya escapado de casa mientras tiene lugar la conversación y todo acabe bien. Se trata, como vemos, de un incipiente ejemplo de aplicación de teoría de juegos en la filosofía moderna y la conclusión del razonamiento kantiano es bien clara: *ser veraz (sincero) en todas las declaraciones es, pues, un sagrado mandamiento de la razón, incondicionalmente exigido y no limitado por conveniencia alguna*.[18]

La clave aquí reside en analizar la fuerza de la argumentación kantiana y hasta qué punto permanece total o parcialmente intacta en el caso de que uno o más de los tres personajes de nuestra historia fuera un robot (lo que en realidad no está tan lejos de la verdad si quién nos pregunta es en realidad un hombre arrebatado por la ira, según los postulados del propio Kant); y ello con total independencia de las cualidades intelectuales, morales o políticas que pensemos que puede llegar a tener de hecho o intrínsecamente un ser inteligente artificial.

Si asignamos un orden convencional a los tres personajes de nuestra historia en los papeles de 1: quien pregunta en la casa si allí está una persona; 2: el inquilino de la casa que debe responder y 3: el fugitivo que se ha refugiado en ella, y dejamos abierta la motivación del perseguidor –sea esta criminal o no–, se nos ofrece una interesante casuística que debería considerar ocho situaciones de interacción posibles. A saber: que los tres personajes sean personas humanas: PPP, que los tres sean robots- pongamos que con aspecto humano –es decir, androides, para redondear la situación con un elemento estético que no es desdeñable aunque no se pueda explorar en este escrito–, es decir: AAA; y luego los casos APP, PAP; PPA, AAP, APA; PAA. A su vez, podemos imaginar distintas variantes, y matices en cada uno de los casos que resultan significativas a la hora de sacar conclusiones acerca de la moralidad de la interacción, especialmente si tenemos en cuenta que moralidad puede abarcar, en este caso, legitimidad, justicia o bondad.

Pensemos, por ejemplo, en el caso de un robot policía en un Estado legítimo, y que el amigo que se esconde en la casa sea, como tantos personajes de películas de Hitchcock, víctima de un terrible malentendido (o eso nos dice él) que lleva al robot a catalogarlo –según sus protocolos de decisión– como un terrorista en plena ejecución de un horrible atentado. Esta situación plantea claramente las tres cuestiones de si es posible y/o legítimo y/o bueno mentirle a ese robot. Un robot que bien podría expresar en sí mismo –como los tribunales revolucionarios del *22 pradial*– la *voluntad general* de la república.

O bien, consideremos el bloqueo que, en el caso PAP, tendría un mayordomo androide kantiano a la hora de elegir entre la primera de las clásicas leyes robóticas de Asimov: *un robot no puede hacer daño a un ser humano o, por inacción, permitir que un ser humano sufra daño*, y la argumentación que Kant nos presenta en su opúsculo y que se expresa netamente en la primera formulación del imperativo categórico.[19] En cuanto a la situación AAA, parece precipitado descartarla de nuestra consideración si tenemos en cuenta la posibilidad de que la programación de un sistema de inteligencia artificial contenga más inteligencia ética que otro en los términos que seguidamente se indicará. La casuística es variada, como se ve, y siempre interesante. Sin ánimo de agotar la indagación que aquí se nos abre, quisiera llamar la atención sobre tres problemas filosóficos asociados a esta simulación.

El primero es la sospecha de que el bloqueo axiológico del robot mayordomo del caso PAP podría sugerir una inconsistencia interna entre la primera y la tercera formulación del imperativo categórico en el propio Kant; es decir, entre el principio de elevar la máxima de nuestra acción a ley universal y el principio de tomar siempre a una persona como fin en sí misma y no sólo como medio. El segundo toma como base la el hecho de que entre el *conocer con verdad lo que se debe*

[18] Kant, 2006: 64.

[19] Claro está que, en este caso, Kant podría ofrecerle al robot el argumento de que quien dice la verdad es causa pero no agente, mientras que si miente y sucede que, a causa de esa mentira, acontece una desgracia sería agente además de causa. Sobre la consideración moral de las tres célebres leyes de la robótica propuestas por Asimov por primera vez en el relato *Runaround*, de 1942, puede verse Anderson, 2008.

hacer y el *ser bueno* existe una diferencia cuya problemática relación surca toda la historia de la ética (Palacios: 2008). Trasladada a nuestra situación, esta distinción subraya una faceta moral específica en la programación y ejecución de nuestros sistemas de inteligencia artificial. Con independencia de la índole moral que otorguemos a un agente o sistema, puede y debe haber verdadera *inteligencia moral* en el diseño y la ejecución de todo sistema de inteligencia artificial.

En tercer lugar, podemos llevar todavía un poco más lejos el análisis de esa moralidad kantiana que, como hemos visto y al menos en los términos de la primera formulación del imperativo categórico, no presupone ninguna vivencia empática ni de otro yo ni de la naturaleza, y preguntarnos, de el enfoque la simpatía scheleriena, por el estatuto de la "otreidad" misma –de la función de ser "otro yo"– en una situación de interacción ética como la descrita. Valgan estas líneas del propio Scheler para demarcar el problema roboético al que apuntamos:

> El problema del individuo y la comunidad y el del "yo" y los "otros" como sujetos psíquicos es, finalmente, y en un sentido más fundamental, también un problema de valor, tanto un problema ético como un problema jurídico… En la forma más exclusivista, pero también la más clara y más rigurosa atacó ya J.T. Fichte la cuestión por este lado. Argumentaba aproximadamente así: partiendo de una primitiva conciencia del deber, o de una pura conciencia de lo que debe ser (que para él es no sólo, con Kant, el supuesto de toda aprehensión de un valor y de todas las decisiones prácticas, sino también de toda afirmación y negación teórica de hechos objetivos, en el sentido de su "primado de la razón práctica sobre la teórica"), en cuanto núcleo y esencia del "yo" puro hay que exigir la forzosa existencia de sujetos extraños con el carácter de un yo "para los cuales" yo pueda tener deberes (cualesquiera).[20]

Valga decir que, desde este planteamiento, el problema no es: de la índole ontológica *–a posteriori–* del robot dependerá el deber que yo tenga para con él, sino precisamente el contrario: de la índole *–a priori o trascendental–* de mis deberes para con un robot dado dependerá su índole ontológica.

III

La obra de Kant nos enseña que es posible sugerir formas de responsabilidad moral que no precisan ni una percepción moral de la mentira, ni una percepción empática del yo ajeno –que, con todo, permanece también abierta en la interacción humano-robot–[21] ni una percepción de *lo natural* en ninguna de sus variantes. Es decir, pudiera ser que nada de esto fuera estrictamente necesario en la situación comunicativa de mentirle a un robot –o de ser mentidos por él. Esta situación, en términos de inteligencia artificial, suscita el problema de si la mentira es definible o no en términos exclusivamente algorítmicos o pragmáticos, o en meros términos de disfunción o incoherencia entre acciones (Wagner y Atkin, 2010).

Es, incluso, posible que la verdad, como el error o la falsedad, sí fueran definibles en relación a secuencias pragmáticas de algoritmos pero la mentira, como tal, no lo fuera. Lo que es tanto como decir que el criterio de *comunicación eficaz bienintencionada* podría anular en cualquier apreciación de la mentira en un contexto comunicativo. De ahí esa tendencia "natural" a la mentira que puede detectarse ya en la interacción entre robots como simulación de una estrategia adaptativa (Floreano, 2007).

Sin embargo, más allá del desarrollo de la rica casuística argumental que nos abre la robotización del ejemplo de Kant, es al propio horizonte argumental en el que esta casuística se inscribe, al que queremos hacer alguna referencia para finalizar.

Si retomamos nuestro nivel 3 de argumentación moral parece oportuno preguntarse, tras nuestra breve experiencia kantiana, si la interacción con un sistema de inteligencia artificial puede hacernos más buenos o más malos. Siguiendo una inveterada tradición del gremio filosófico no voy

[20] Scheler, 2004:289. El propio Scheler amplía luego la perspectiva del problema al revisar la función de la *certeza intuitiva*, es decir, *la aprehensión inmediata de algo que no se puede experimentar,* desempeña en nuestro acceso a otro agente como un *yo* (Scheler, 2004: 297).

[21] Una aproximación psicológica real a la interacción humano-robot en este tipo de contextos puede verse en Kahn, 2012.

a ofrecer una respuesta definitiva a esta pregunta. Pienso, con todo, haber aportado elementos suficientes para concluir que, en efecto, es una pregunta muy importante, éticamente hablando, y mucho más difícil de contestar de lo que parece a primera vista. Para empezar, de lo que llevamos dicho se desprende: 1) que no hay un consenso tan consolidado como parece en la tradición de nuestra argumentación ética para responder negativamente, y 2) que resolver esta cuestión implica habilitar recursos argumentales novedosos en la historia de la argumentación ética.

La interacción del ser humano con cualquier objeto forma parte de la propia condición humana y posee ya alguna índole de humanidad (de hecho casi todas las tradiciones religiosas del mundo otorgan algún tipo de divinidad o sacralidad a numerosos objetos o lugares con cuya interacción nos hacemos mejores o peores). La interacción del hombre con el robot es, además, interacción del hombre consigo mismo a través de un robot. Pero además, en el caso del robot, esta índole de humanidad cobra un significado especial porque en el proyecto racional del comprensión del universo y de sí mismo que la cultura moderna pone en marcha, la producción de un ser que le iguale en condición y capacidad es un hito intrínseco, y no extrínseco, al propio ser humano.[22] "El robot", como proyecto indeterminado o como objeto preciso, forma parte del alma humana y es un pliegue material más (mecánico, electrónico, cuántico, biológico) de la inteligencia humana su tarea de comprensión de sí misma. Más aún, se trata también de un proyecto inherente a su vocación de *Imago Dei* y, en este sentido, está intrínsecamente vinculado a cuanto el hombre asuma como sagrado.[23]

Surge así un interesante dilema lógico. Un dilema que podemos expresar en estos términos: si la inteligencia –que es siempre teórica y práctica a la vez– es una cualidad meramente humana, es decir, patrimonio propio de nuestra especie biológica; como es inherente a la condición humana el humanizar, a su vez, el universo y la materia que habita; no se puede negar, a priori, la posibilidad de que el ser humano integre en diversos niveles de eticidad a nuevos agentes morales producidos total o parcialmente por él. La creación, también en el plano material y no sólo jurídico o mental, de un entorno material progresivamente más ético constituiría una dimensión ineludible en la conformación del hábitat humanizado. Podríamos decir que introducir ética en la materia es una condición trascendental de cualquier progreso técnico objetivo.

En este sentido, incluso admitiendo el supuesto de que cualquier sistema de inteligencia artificial producible por el hombre no sea sino una formidable estructura de decisiones "enlatadas" de algún modo –lo que ya es bastante suponer–, no podemos sucumbir al espejismo de que la inteligencia natural sea más "humana" que la inteligencia artificial. La inteligencia artificial es, precisamente, la que los humanos comprendemos, la que sabemos hacer y la que nos refleja y expresa como seres sociales. La inteligencia natural, en cambio, la de nuestro par de hemisferios neuronales, sigue siendo un formidable y extraño misterio, sólo podrá dejar de serlo en la medida en que podamos reproducirlo. En realidad, si refleja el rostro de alguien, será, desde luego, el de la divinidad, al menos inicialmente; no el nuestro.

Ahora bien, también es posible que, según una rica y venerable tradición de pensamiento que nos remonta al *nous* de Anaxágoras*,* al *logos* de filón de Alejandría, al entendimiento agente de Aristóteles (al menos en su versión averroísta), o a la gramática universal de Raimundo Lulio o Leibniz, la inteligencia –también la inteligencia ética, sea esta o no bondad moral– no remita

[22] Obviamente, esta reflexión no es sólo válida en el estricto caso de robots o androides. La expansión de la corporalidad humana, inherente a nuestro sistema tecnológico, abarca otros fenómenos como el cyborg o la corporalidad virtual. Una aproximación literaria a esta última cuestión la encontramos en Hayles, 1999. Debe recordarse que las prolongaciones artificiales de la percepción, el deseo o la decisión humanas , en diversas modalidades de entramados tecnológicos, desempeñan un destacado –aunque más o menos implícito– papel en obras tan dispares como las de Munford, Heidegger, Deleuze, Baudrillard, Ihde, etc.

[23] Como las figuras del androide o el cyborg en el cine, la del autómata humanoide, desempeña un papel importante en nuestro imaginario colectivo desde la Antigüedad. La leyenda que de Descartes habría construido un autómata para suplir a su única hija, Francine, fallecida a los 5 años, y que el embarco de este autómata en su viaje a Suecia habría impedido la navegación hasta que el capitán descubre al monstruoso pasajero y ordena arrojarlo por la borda, recoge los principales elementos de esta figura . El origen de esta leyenda parece estar en la obra del cartesiano Bonaventure d´Argonne *Mêlanges d´Historie et de Litterature*, de 1700. Puede verse, a este respecto, la obra de Kang, 2011.

exclusivamente a una herramienta biológica de cierta especie concreta, sino también a una cualidad real de acontecimientos, procesos y sistemas en el universo, es decir, a una *inteligencia objetiva* de la que, en principio, pueden participar, con modalidades propias, infinitos agentes inteligentes.

Este sería el segundo término de nuestro dilema y, como sucede con todos los dilemas, tanto por el primero como por el segundo de sus términos alternativos se llega a la misma conclusión.

REFERENCIAS

Anderson, S. L. (2008). "Asimov´s "three laws of robotics" and machine metaethics". *Journal AI & Society* - Special Issue: *Ethics and artificial agents archive*, 22(49), 477-493.

Arendt, H. (1958). *The Human Condition*. Chicago: Chicago University Press.

Floreano, D., Mitri, S., Magnenat, S., Keller, L. (2007). "Evolutionary conditions for the emergence of communication in robots". *Current Biology* 17(6), 514-9.

Hayles, N. K. (1999). *How we Became Posthuman, virtual bodies in Cybernetics, Literature and Informatics*. Chicago: Chicago University Press.

Hegel, G. W. F. (1990) [1817]. *Enciclopedia de las ciencias filosóficas*. Trad. Francisco Larroyo. México: Porrúa.

Kahn, P. H., *et al.* (2012). "Do people Hold a Humanoid Robot Morally Accountable for the Harm it Causes?", *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM, Nueva York, 33-40.

Kang, M. (2011). *Sublime Dreams of Living Machines: The Automaton in the European Imagination.* Cambridge, MA: Harvard University Press.

Kant, I. (1956) [1785]. *Grundlegung zur Metaphysik der Sitten* (*Fundamentación de la metafísica de las costumbres*). Wiesbaden: Suhrkamp.

— (1797) "Sobre un presunto derecho de mentir por filantropía" (Trad. Juan Miguel Palacios). En *Teoría y práctica*. Madrid: Tecnos.

Lin, P., *et al.* (2012). *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge MA: MIT.

MacIntyre, A. (1981). *After Virtue*. Notre Dame: University of Notre Dame Press.

Mitcham, C. y Duval, R.S. (2000). *Engineer's Toolkit: Engineering Ethics*. Upper Saddle River, NJ: Prentice Hall.

Moon, A., *et al.* (2012). "Open Roboethics: Establishing an Online Community for Accelerated Policy and Design Change", *Proceedings of We Robot 2012*. Miami: Miami University Press.

Palacios, J.M. (2008). *Bondad moral e inteligencia ética*. Madrid: Encuentro.

Paton, H. J. (1971). *The Categorical Imperative; a study in Kant's moral philosophy*. University of Pennsylvania Press.

Popper, K. R. (1994) [1987]. *Sociedad Abierta, universo abierto*. Madrid: Tecnos.

Powers, T. M. (2011). "Prospects for a Kantian Machine". *Machine Ethics*. En Anderson, M. y Anderson, S. (eds). Cambridge, MA: Cambridge University Press.

Quintanilla, I. (2006). "El cambio tecnológico: cuatro apuntes desde la filosofía", *Empresa y Humanismo*, IX, 141-164.

— (2012). *Techné. La filosofía y el sentido de la técnica.* Madrid: Common Ground Publishing.

Scheler, M. (2004) [1923]. *Esencia y formas de la simpatía*. Trad. José Gaos. Buenos Aires: Losada.

Simon, H. A. (1968). *The Sciences of the Artificial*. Cambridge, MA: MIT Press.

Veruggio, G. (2005). "The Birth of Roboethics", *IEEE- ICRA,Workshop on Roboethics*, 1-4.

Veruggio G. y Operto F. (2010). "Ethical and societal guidelines for Robotics". En Gerhard Dabringer (ed.), *Ethical and Legal Aspects of Unmanned Systems*. Viena: Austrian Ministry of Defence and Sports.

Wagner, A. R. y Atkin R. C. (2010). *Robot deception: Recognizing when a Robot Should Deceive.* Georgia: Georgia Institute of Technology.

SOBRE EL AUTOR

Ignacio Quintanilla Navarro: Nacido en 1960, doctor en filosofía y licenciado en psicología industrial ha impartido docencia en la Universidad Complutense de Madrid y el la Universidad Nacional de Educación a Distancia y actualmente es colaborador habitual de la Escuela de Filosofía de Madrid. Sus áreas de investigación habituales son teoría del conocimiento, historia de la filosofía moderna y contemporánea y filosofía de la técnica. En todas ellas ha publicado diversos artículos y monografías. La última es *Techné, la filosofía y el sentido de la técnica*, Common Ground Publishing, 2012.