Límites éticos de la inteligencia artificial

Se describe a lo largo del artículo la inteligencia artificial como actividad, destacando los aspectos en que aparecen problemas éticos según el campo específico de aplicación. Posteriormente la exposición se centra en los sistemas de toma de decisión, y concretamente en aquellos en los que hay que considerar múltiples criterios conflictivos. Por último, en las conclusiones finales se trata de recoger criterios que permitan establecer los límites éticos de la inteligencia artificial.

Fernando de Cuadra*

Introducción

¿Se podrá automatizar cualquier tarea actualmente desempeñada por seres humanos? Creo que sí. ¿Y las tareas intelectuales más abstractas, que sólo llevan a cabo personas muy capaces y

* Profesor de Informática en la Universidad Pontificia Comillas. Madrid.

Tomo 243 (2001) RAZÓN Y FE PP. 498-508

especializadas? También. Las máquinas ya reemplazan con éxito a campeones de ajedrez o a pilotos de avión. Y lo más espectacular está aún por llegar.

En la práctica, los límites técnicos no son críticos. Los avances de la tecnología rompen las barreras establecidas a mayor velocidad de lo que la sociedad y el individuo puedan asimilar. Los límites físico-teóricos están demasiado lejos para que lleguen realmente a restringir posibilidades. Debemos mucho antes encontrarnos con otros obstáculos, de naturaleza económica, sociojurídica y –especialmente– ética.

Parece que hay cierta diferencia esencial entre un piloto automático y un juez automático (o médico, o policía). Nos resignamos a poner nuestra vida en manos del primero, pero seguramente pondríamos reparos a someternos a las decisiones del segundo. ¿Por qué? Los dos son sistemas automáticos, que parten de una cierta información para tomar decisiones que pueden ser críticas. Este artículo trata de analizar la naturaleza de la barrera que separa estos dos tipos de sistemas automáticos.

Ingeniería

En general, el objetivo de la ingeniería es el diseño, desarrollo y explotación de sistemas más o menos complejos para mejorar las condiciones de vida de las personas de una forma racional y controlada. Los sistemas están formados por estructuras, máquinas y personas que colaboran activa o pasivamente en la consecución de los fines establecidos.

Los límites técnicos de la ingeniería se establecen dinámicamente según los avatares teóricos de la ciencia (modelos y principios contrastados experimentalmente) y los avatares prácticos de la tecnología (capacidad de diseño, fabricación y control). Otros límites importantes son los de la imaginación creativa. Pero debe distinguirse entre imaginar un sistema visto desde fuera, como un comportamiento deseable, e imaginar un sistema visto desde dentro, como un comportamiento factible y controlable. Lo primero es un ejercicio de creatividad especulativa o artística, necesaria pero no suficiente para desarrollar un sistema. Si no somos capaces de ir más allá, nos quedamos en el terreno de la ciencia-ficción. Lo segundo es un ejercicio de creatividad práctica propio de la ingeniería, que en general podemos llamar diseño.

Cuando los sistemas que se desea desarrollar son complejos, los proyectos de ingeniería también lo son. Esto obliga a que las decisiones –y los pro-

blemas éticos asociados a las mismas– se tomen en sucesivas etapas siguiendo una estructura jerárquica. Simplificando, podemos identificar cinco etapas fundamentales:

1. Aplicación de políticas generales y estudios preliminares.
2. Especificación.
3. Diseño.
4. Desarrollo.
5. Explotación.

Tomemos como ejemplo la construcción de una central nuclear. En la primera etapa debe decidirse sobre el conflicto entre rentabilidad y peligros de la energía nuclear y, si se decide que compensa globalmente, hay que adoptar un conjunto de normas de seguridad para minimizar estos peligros hasta límites tolerables. En la segunda etapa se describe en detalle el conjunto de características técnicas de la central, como su localización, potencia y tipos de tecnología, pero también se detalla cómo se deben aplicar las normas de seguridad decididas en la primera etapa. En la tercera se describen con detalle todos los subsistemas de la central, las conexiones y relaciones entre ellos y cómo se plasman en todas y cada una de sus piezas las normas y procedimientos de seguridad de la segunda etapa. En la fase de desarrollo se construye, y en la de explotación se mantiene el sistema bajo control para respetar en todo momento las normas emanadas de las etapas anteriores.

Las decisiones con mayor contenido ético se plantean en las fases iniciales de un proyecto. Frecuentemente, aparece un conflicto entre beneficio económico y daño al medio ambiente, entre bienestar a corto plazo y posibles desastres a medio plazo, o también entre molestias a una minoría frente a supuestos beneficios para la mayoría. A partir de la decisión inicial, básicamente hay que seguir escrupulosamente las normas y no cometer negligencias, errores ni fraudes. Debe tenerse en cuenta además que en muchos proyectos de ingeniería ni siquiera son relevantes las decisiones iniciales (por ejemplo, el desarrollo de un nuevo modelo de televisor).

Algo fundamental en los proyectos de ingeniería es que siempre existe un responsable de cada aspecto del proyecto: diseño, seguridad, pruebas, puesta en marcha, mantenimiento, etc. Si se produce un fallo, se supone culpable al responsable, que tiene que demostrar su inocencia. La responsabilidad es muy importante, pues condiciona críticamente el grado de autonomía de las máquinas inteligentes.

Inteligencia artificial

BAJO este nombre recojo un conjunto bastante diverso de disciplinas que persiguen objetivos similares: resolver automáticamente problemas, y/o realizar tareas, que por su complejidad requieren un cierto grado de inteligencia. Entre estas disciplinas incluyo la programación matemática, la investigación operativa, la informática avanzada, la ingeniería del conocimiento y el control inteligente. En cuanto a su vertiente aplicada, considero la inteligencia artificial como una rama más de la ingeniería.

Muchos autores restringen el término «inteligencia artificial» a un conjunto limitado de tipos de aplicaciones informáticas. Esto se debe a diversas causas: razones históricas, modas intelectuales; prejuicios gremiales, sutilezas técnicas, ventajas comerciales, el estado actual de la tecnología, o simple ignorancia. Los argumentos más respetables que se manejan para hacer esta distinción son de dos tipos:

1. La naturaleza del problema o tarea que se desea automatizar. Algunos ejemplos son la visión artificial, el aprendizaje, algunos problemas complejos de control, el diagnóstico a partir de síntomas, el diseño «creativo», y el procesamiento de lenguaje natural.

2. La técnica (o tecnología) empleada. Por ejemplo, algunos tipos de lenguajes de programación, la búsqueda heurística, las redes neuronales, la lógica borrosa, los sistemas expertos, y el reconocimiento de patrones.

Según mi experiencia, toda frontera que se quiera establecer entre inteligencia artificial e informática convencional es superficial, difusa, y dependiente del estado del arte. Por esto creo que resulta más adecuado eliminar de raíz dicha frontera.

Aunque hay sistemas electromecánicos analógicos capaces de realizar tareas bastante complejas, la inteligencia artificial está estrechamente relacionada con los sistemas digitales. Los sistemas digitales están compuestos por máquinas físicas y virtuales (*hardware* y *software*) cuyo comportamiento puede describirse mediante álgebra –o lógica– binaria. La capacidad potencial de los sistemas digitales para implantar comportamientos complejos es inmensa. A ello contribuyen la inversión en tecnología avanzada en *hardware,* la posibilidad de abstracción del *software* y la propia automatización de técnicas de diseño y verificación. Los límites prácticos se deben fundamentalmente al estado de la tecnología de los dispositivos físicos (almacenamiento limitado de información, tiempo de respuesta, tamaño y consumo), y a las restricciones impuestas por el mercado en cuanto rentabilidad de las inversiones.

¿Pueden las máquinas ser inteligentes, o bien sólo parecerlo? Ésta es una cuestión que puede suscitar discusiones interminables (y muy entretenidas), pero es irrelevante desde el punto de vista –siempre práctico– de la ingeniería. Si un sistema funciona como si fuera inteligente, ¿qué importa si realmente lo es? Otra cuestión bien distinta es la responsabilidad. Una máquina no es libre, y por tanto no es responsable de sus decisiones ni de sus actos. Libertad, responsabilidad y conciencia son características propias de la condición humana, y estrechamente relacionadas entre sí.

Aspectos éticos según tipos de aplicaciones

Se considera aquí que las decisiones problemáticas desde un punto de vista ético son sólo aquellas que puedan perjudicar directa o indirectamente a las personas. Por lo tanto se debe cuestionar si el desarrollo y explotación de un sistema inteligente puede afectar a la vida, la salud (lo que incluye el medio ambiente), la libertad, la intimidad, el puesto de trabajo o la dignidad de alguien, y en general a su bienestar físico y mental.

Hay áreas de trabajo que son en sí mismas una fuente inagotable de preocupación ética. En este grupo incluyo los sistemas de armamentos, la genética y la industria de alto riesgo medioambiental. Dado su carácter general y su amplia difusión, este tipo de problemas éticos no es objeto de este artículo.

Dos tipos de aplicaciones de la inteligencia artificial también se van a dejar deliberadamente fuera de esta discusión. Uno es el recreativo y artístico, que incluye el campo de los videojuegos y la realidad virtual. El otro es el de las herramientas avanzadas de trabajo y los equipos técnicos especializados, utilizadas en actividades industriales, burocráticas o contables.

La amenaza planteada por el primer tipo de aplicaciones está relacionada con problemas psicológicos de alienación, adicción, incomunicación personal o educación. No se discute aquí por falta de conocimiento y por tanto de opinión fundada sobre el tema. La amenaza planteada por la automatización de actividades laborales se relaciona con la destrucción de puestos de trabajo. Esta amenaza se ignora por considerar que la dinámica del mercado de trabajo es una realidad ya muy conocida y globalmente beneficiosa (aunque pueda afectar localmente a algunos colectivos, y esto debe resolverlo la legislación laboral). Quien argumenta, por ejemplo, que una excavadora

quita el trabajo a muchos obreros, simplemente ignora los puestos de trabajo generados por su diseño, fabricación, comercialización, transporte, manejo, mantenimiento, formación de personal especializado y otras actividades asociadas.

No es fácil clasificar los campos de aplicación de la inteligencia artificial, pues tienden a ser todos los imaginables. Para identificar grandes familias de aplicaciones y sus problemas éticos asociados se ha optado aquí por un criterio sencillo y bastante claro, basado en las funciones más elementales que realiza un sistema digital.

A diferencia de otras especialidades de ingeniería, la materia prima de la inteligencia artificial es la información, en la que se materializa algo más abstracto, que es el conocimiento. Hay cuatro funciones básicas que se realizan en cualquier proceso o sistema complejo: almacenamiento, transporte, proceso y control. En función de estas cuatro tareas básicas, identificamos los siguientes tipos de aplicaciones informáticas:

1. *Gestión de información.* La información se almacena físicamente en las memorias, y a más alto nivel en las bases de datos. En estas aplicaciones, la función principal es el almacenamiento y recuperación de información.
2. *Comunicación.* La información se transporta mediante sistemas y redes de comunicaciones. En estas aplicaciones, el transporte es la función más importante.
3. *Control.* El control de los sistemas consiste en la sincronización de las tareas que realiza cada uno de los subsistemas para garantizar la consecución de una tarea común más compleja. En estas aplicaciones, la función de control es dominante.
4. *Razonamiento automático.* El proceso o transformación de información corre a cargo de procesadores (a nivel físico) y sistemas de cálculo automático en general. En este tipo de sistemas, la función de proceso de información es la fundamental.

Cualquier sistema digital complejo reúne todas estas funciones básicas. Pero habitualmente es fácil identificar cuál es la función dominante de una aplicación, y esto nos permite caracterizar sus distintos problemas éticos asociados de una forma sencilla.

En las aplicaciones de gestión de información y de comunicación, los problemas éticos están relacionados con el derecho a la intimidad y a un cierto grado de anonimato de las personas. No existe una garantía real de seguridad (informática) ante el acceso de personas no autorizadas a nuestros datos

más íntimos. Existe el riesgo de opresión por un estado policial, o por otro tipo de organizaciones. También se puede poner en peligro el derecho a la rehabilitación, haciéndonos esclavos de por vida de nuestra historia pasada.

En las aplicaciones de control (robótica, transportes automatizados, control de procesos industriales), los prolemas éticos son los propios de la ingeniería clásica: cómo se garantiza la seguridad (física) de las personas y del medio ambiente, quién se hace responsable de ella en caso de fallos y, por tanto, de posibles desastres. El problema se reduce al análisis, gestión y aceptación social del riesgo asociado al progreso tecnológico.

En las aplicaciones de razonamiento automático, los problemas éticos son quizá más complejos y sutiles. A su discusión se dedica el siguiente apartado.

Razonamiento automático y toma de decisiones

AUNQUE hay diversas técnicas alternativas para abordar este tipo de desarrollos, se va a emplear aquí un enfoque propio de la programación matemática y de la investigación operativa. Esto es suficiente porque todas las técnicas persiguen los mismos objetivos y, por tanto, también plantean los mismos problemas éticos. Para facilitar la exposición se empleará un proyecto real de logística, como es la asignación óptima de turnos de trabajo en una empresa.

Centrémonos en cómo resolver el problema de forma automática. La solución del problema se modela como un valor concreto de un conjunto de variables de decisión. En el ejemplo, algunas variables de decisión pueden ser el número total de turnos, las horas de comienzo de cada turno, o las horas de descanso programadas. La solución tendrá que cumplir una serie de restricciones; por ejemplo, que haya en todo momento un número suficiente de trabajadores (restricción técnica) o que ningún turno supere ocho horas de duración (restricción legal, pactada, o simplemente racional).

Toda solución que cumpla las restricciones es una solución factible. Pero si hay más de una solución factible, se querrá seleccionar la mejor (o las mejores) de todas, y eso se determina mediante atributos, objetivos o criterios de optimización. Un atributo podría ser minimizar el número total de trabajo, y otro podría ser respetar al máximo las preferencias de horario de cada trabajador, o intentar centrar las horas de descanso dentro de cada turno de trabajo.

Si se desea obtener sólo una solución factible, basta con buscar una que respete todas las restricciones. Si además se desea obtener la solución óptima, hay que seleccionar la mejor de entre las factibles. Esto último exige definir artificialmente qué se entiende por «mejor». Una práctica habitual es sopesar los distintos atributos reduciéndolos a una sola función que se desea minimizar (función objetivo). Por ejemplo, una suma ponderada de atributos donde cada factor de ponderación refleja la importancia relativa de cada uno frente a los demás. La solución óptima será la que corresponda a un valor mínimo de la función objetivo.

Si los atributos son realmente conflictivos entre sí, será difícil incluirlos en una única función objetivo. En el ejemplo empleado, es difícil ponderar el número total de horas de trabajo (coste) frente al grado de satisfacción de los trabajadores. Otro ejemplo habitual en ingeniría es el conflicto entre minimizar el coste de una inversión industriall frente a minimizar su impacto medioambiental. Lo más correcto (práctica y éticamente) en estos casos es aplicar técnicas de optimización multiatributo, que permiten obtener un conjunto óptimo de soluciones no condicionado por una determinada ponderación.

Este conjuno de soluciones cumple la propiedad de que ninguna es objetivamente peor ni objetivamente mejor que ninguna otra. La herramienta automática ayuda a la toma de decisiones proporcionando un conjunto óptimo de alternativas y la información necesaria para ponderarlas, pero la decisión final –y por tanto la responsabilidad– se deja al usuario de la herramienta.

Razonamiento automático y ética

El problema ético más inmediato en un sistema de razonamiento automático es el mismo que en otros tipos de sistemas: hasta qué punto el objetivo fundamental de la aplicación es ético. Ya se ha comentado que hay áreas de tabajo éticamente delicados, como el armamento, la genética o ciertos tipos de industria. En el ejemplo anterior de los turnos de trabajo el objetivo es opinable, porque puede verse como un arma empleada para explotar al trabajador (se emplea para negociar convenios, pero sólo dispone de ella la empresa) o como un método racional para obtener beneficios mejorando al mismo tiempo las condiciones de trabajo.

En las aplicaciones de toma de decisiones, la definición y forma de uso de la herramienta dependen en gran medida de la labor del ingeniero de

especificación y diseño, frente a otros tipos de proyecto controlados más rígidamente por el cliente. Existe la posibilidad de definir restricciones con flexibilidad, de hacer explícitos los criterios conflictivos y de mostrar las posibilidades reales de la herramienta como ayuda, y no como elemento frío de decisión. Los objetivos finales de la aplicación se pueden modificar parcialmente durante el proyecto, y en algunos casos esta influencia puede considerarse un deber desde un punto de vista ético.

Una cuestión más general es hasta qué punto debe confiarse a una máquina la toma de decisiones, cuando nunca puede ser responsable de las mismas. Antes de tratar este tema, vamos a excluir de la discusión una gran cantidad de aplicaciones: todas aquellas en las que la solución de un problema es única por naturaleza. Esto puede ocurrir porque la restricciones permitan sólo una solución o porque no existan realmente atributos conflictivos entre sí.

En el caso de problemas con solución única, la toma automática de decisiones no es éticamente discutible. Una solución errónea sería simplemente un fallo, pero no se delega la responsabilidad. Se debe tratar como una aplicación de tipo «control», y por tanto debe desarrollarse bajo las normas de seguridad correspondientes a la gravedad de las consecuencias del fallo. En casos en que esté en juego la vida humana, esto incluye pruebas sistemáticas, auditorías, redundancia de sistemas, monitorización y funciones de autocomprobación. El sistema debe funcionar además bajo la supervisión de personas que son responsables legal y moralmente de su buen comportamiento.

En problemas que incluyen atributos éticamente conflictivos, la decisión final debe tomarla una persona que se haga responsable de las consecuencias. Una herramienta automática puede generar el conjunto de soluciones óptimas, para que el usuario de la herramienta escoja entre ellas. En el ejemplo de la distribución de turnos de trabajo, la empresa puede natualmente elegir la solución de mínimo coste ignorando el bienestar de los trabajadores (respetando sólo mínimos legales o de convenio). Pero se dan dos circunstancias muy interesantes: primero, se conocen los costes objetivos de ceder más o menos en la negociación, pues está disponible todo el abanico de soluciones óptimas; segundo, en ninguna de las soluciones óptimas se perjudica a los trabajadores innecesariamente, sino sólo en caso de que realmente esto represente un beneficio significativo para la empresa.

Esta última propiedad es muy importante. Por ejemplo, mediante los resultados obtenidos por la herramienta de razonamiento automático, se comprobó que se podía aumentar bastante el tiempo de descanso de los trabajadores sin que la empresa se viera afectada en absoluto. Curiosamente, el

problema de la duración del descanso había ocupado horas de negociación en años anteriores, al igual que otros puntos conflictivos sólo en apariencia.

Resumen y conclusiones

NO es la tecnología la que restringirá la aplicación de sistemas automáticos a cualquier tipo de actividad. Mejor será entonces que sean sus problemas éticos asociados los que establezcan los límites prácticos.

Un sistema automático inteligente puede desarrollarse para cualquier tipo de finalidad. Por tanto, una fuente de problemas éticos es la valoración de la finalidad en sí. Ejemplo de áreas de trabajo éticamente discutible son la industria de armamento, los procesos industriales contaminantes o la manipulación genética.

Un campo en el que lo discutible no son los objetivos, sino el uso de la Inteligencia Artificial en sí, es el de los juegos y el arte. Es discutible su uso por la posibilidad de adicción, y por la deshumanización –o alienación– en unas actividades tan ligadas tradicionalmente a las relaciones personales.

Clasificando los tipos de sistemas inteligentes por sus funciones principales (almacenamiento, transporte, control y proceso), se han identificado tres grandes familias de problemas éticos:

- En las aplicaciones de tipo «control», los problemas éticos son similares a los de otras especialidades de ingeniería. Ante el posible riesgo de desastres y accidentes, hay que establecer las normas, procedimientos y controles de seguridad necesarios. Además hay que contar con la supervisión de personas que se hagan responsables legal y moralmente del comportamiento del sistema, siempre bajo el supuesto de que no existe un sistema absolutamente infalible.
- En las aplicaciones de gestión masiva de información y comunicaciones, se puede amenazar seriamente la intimidad, libertad y dignidad de las personas. Esto plantea problemas éticos especiales, en particular porque es muy difícil responsabilizar directamente a una persona u organización de los posibles abusos y usos indeseables del poder de la información.
- En las aplicaciones de tipo «razonamiento automático», hay que distinguir entre razonamientos de solución única (que se tratarían éticamente como aplicaciones de control), y razonamientos que emplean argumentos éticamente conflictivos, en los que se puede generar auto-

máticamente un conjunto de decisiones óptimas. La presencia de argumentos conflictivos en la toma de decisiones debería dejar a la máquina el papel de generar datos objetivos y precisos sobre la verdadera medida del conflicto, para consulta y ayuda de la persona responsable. Esto tampoco garantiza –evidentemente– que la decisión final sea éticamente correcta, pero al menos puede evitar errores innecesarios, como sería el adoptar una decisión que es peor que otra desde cualquier punto de vista.

La falta de libertad y, por tanto, de responsabilidad de las máquinas se presenta aquí como factor crítico para limitar su acción y autonomía. Naturalmente, podemos cuestionar si somos nosotros realmente libres y realmente responsables de nuestros actos y decisiones. Tal vez sólo seamos máquinas programadas de forma tan compleja como para crear nuestra propia ilusión de libertad. Y tal vez la conciencia sea un mecanismo de supervivencia social, adquirido y transmitido de generación en generación. Está bien, no podemos saberlo con certeza. Pero sí sabemos con certeza que alguien tiene que hacer de juez, de médico o de policía. Y que ese alguien tiene que hacerse responsable de sus decisiones, por lo que no puede ser una máquina creada por nosotros. Tal vez en el fondo no sepamos lo que somos, pero sí que estamos seguros de lo que nuestras máquinas son: nada más que máquinas.