Actas de las Jenui, vol. 7. 2022.
Páginas: 143-150

Cuentos robóticos como recurso didáctico para enseñar ética a ingenieros informáticos

Gonzalo Génova Fuster
Departamento de Informática
Universidad Carlos III de Madrid
ggenova@inf.uc3m.es

María del Rosario González Martín
Departamento de Estudios Educativos
Universidad Complutense de Madrid
marrgonz@ucm.es

Resumen

En este trabajo se presenta la experiencia con el uso de un recurso didáctico particular en un curso universitario de Ética para Ingenieros. El recurso consiste en un breve relato de ciencia ficción («Dilemas robóticos») que sirva para estimular las preguntas sobre qué sea la conciencia moral, siguiendo una larga tradición de uso de la literatura para la educación ética. El género de la ciencia ficción tiene en general bastante aceptación entre el público juvenil, y si son estudiantes de carreras tecnológicas aún más. En este caso se trata de un relato protagonizado por un robot dotado de una «mente» artificial, lo que pretende acercar la temática especialmente a estudiantes de ciencias de la computación. Tras la lectura, comentario y debate del texto, se observa que los alumnos han sido capaces de comprender con mayor hondura algunas nociones éticas fundamentales.

Abstract

This paper presents the experience with the use of a particular didactic resource in a university course of Ethics for Engineers. The resource consists of a short science fiction story («Robotic Dilemmas») that serves to stimulate questions about what is moral conscience, following a long tradition of using literature for ethics education. The science fiction genre is generally quite popular among young audiences, and if they are students of technological careers even more so. In this case it is a story starring a robot with an artificial «mind», which aims to bring the subject especially to students of computer science. After reading, commenting and discussing the text, it can be observed that the students have been able to understand more deeply some fundamental ethical notions.

Palabras clave

Recurso didáctico, dilemas morales, ciencia ficción, ética, conciencia moral.

1. Introducción

En este trabajo se presenta la experiencia con el uso de un recurso didáctico particular en un curso universitario de Ética para Ingenieros. El recurso consiste en un breve relato de ciencia ficción (*Dilemas robóticos*) que sirva para estimular las preguntas sobre qué sea la conciencia moral, siguiendo una larga tradición de uso de la literatura para la educación ética. El género de la ciencia ficción tiene en general bastante aceptación entre el público juvenil, y si son estudiantes de carreras tecnológicas aún más [2]. En este caso se trata de un relato protagonizado por un robot dotado de una «mente» artificial, lo que pretende acercar la temática especialmente a estudiantes de ciencias de la computación.

El curso tiene ya una trayectoria de nueve años, y el temario y diferentes aspectos de los resultados han sido presentados anteriormente [5, 6, 7]. Sin embargo, en estas publicaciones no se analizaba en detalle ninguno de los recursos didácticos empleados en el curso y los resultados alcanzados con cada uno de ellos. El propósito de este artículo es, pues, presentar uno de estos recursos junto con las observaciones y reflexiones que ha motivado en los alumnos.

El objetivo principal del curso, tal como hemos explicado con más detalle en los trabajos citados, se encierra en la palabra central de su título: Ética *para* Ingenieros. Pero no tanto ética específica de la profesión de ingeniero (deontología profesional), sino ante todo *ética explicada a su mentalidad* y a partir de sus propias experiencias. Para la mentalidad del ingeniero, lo real es lo que se puede tocar y medir, el prototipo de pensamiento racional es el razonamiento matemático-deductivo, y los mejores resultados se obtienen siguiendo procedimientos estándar [5, 6, 7]. Por tanto, resulta prioritario afrontar desde el primer momento las dificultades que un estudiante de ingeniería tiene para reconocer el valor del pensamiento específicamente ético y filosófico.

Uno de los mayores peligros en un curso de este tipo sería reducir la ética a un conjunto de normas de

comportamiento (un código ético) que podrían seguirse de modo mecánico o cuasi-algorítmico, un «estándar industrial de comportamiento profesional». Los estudiantes serían perfectamente capaces de aprender el estándar y el modo correcto de aplicarlo, pero no habrían dado el salto a la cuestión verdaderamente filosófica, ética y personal: *¿por qué este estándar y no otro*, qué tiene que ver el imperativo ético *conmigo*?

La universidad moderna se entiende a sí misma como cuna y transmisora de la racionalidad. Sin embargo, esta racionalidad corre el peligro de reducirse a racionalidad científico-técnica, o racionalidad instrumental, como ha sido denunciado exhaustivamente por diversos autores especialmente a lo largo del siglo XX. Nuestros estudiantes universitarios –y nuestros profesores– están muy entrenados en el ejercicio riguroso de la racionalidad instrumental. Sin embargo, lo que podemos llamar racionalidad ética o humanista ha quedado prácticamente marginada en la enseñanza actual de nuestras universidades, especialmente en las universidades públicas, a pesar de ciertos signos de recuperación [9]. No ignoramos, obviamente, las dificultades que encuentra la Ética para alcanzar un estatus propio de racionalidad. Pero queremos enfatizar la necesidad de que la reflexión ética no sea dejada por imposible, como si en este campo todo fuera cuestión de opinión y preferencia arbitrarias.

En el resto del artículo presentamos el contexto general en el que se sitúa el recurso didáctico (2), algunas consideraciones sobre el tratamiento de los dilemas en los cursos de ética (3), el propio texto del cuento (4), algunos fragmentos de comentarios de los alumnos y alumnas que han cursado la asignatura a lo largo de los años, seguidos de un breve análisis por nuestra parte en cada caso (5), discusión general de los comentarios de los alumnos (6), respuesta que podemos dar a la pregunta «qué sea la conciencia moral» (7), y finalizamos recapitulando lecciones aprendidas y trabajos futuros (8).

2. Estructura general del curso y lugar que ocupa el recurso didáctico

Se trata de un curso optativo de 3 créditos impartido en una universidad pública española (Universidad Carlos III de Madrid), que pueden cursar conjuntamente alumnos de diversos grados en ingeniería. El curso se imparte de modo presencial, mediante una sesión semanal a lo largo de todo un cuatrimestre. Además de las clases presenciales, se trabaja mucho en un blog específico de la asignatura, en donde el profesor expone diversos textos que los alumnos comentan y debaten (estos comentarios forman una parte importante de su calificación).

A lo largo del curso se exponen diversos temas para introducir progresivamente a los alumnos en este proceso de reflexión filosófica [5, 6]: desde qué significa ser libres y si los seres humanos somos realmente libres (lo que implica que los comportamientos pueden ser calificados éticamente), hasta la relación de la ética con la ciencia y la tecnología y sus diversas formas de racionalidad (que tampoco es idéntica en las ciencias básicas y en las tecnologías productivas). En el ecuador del curso abordamos, con ayuda del cuento presentado, uno de los temas más difíciles en este recorrido: el juicio moral, la autonomía de la conciencia, y la búsqueda del bien y la verdad.

El tema se presenta planteando estas preguntas a los alumnos: ¿Qué es la conciencia moral? ¿En qué se diferencia de la conciencia psicológica? ¿Tiene la conciencia individual la última palabra sobre las cuestiones éticas? ¿Puede equivocarse la conciencia? ¿Tiene sentido plantearse esta última pregunta?

El objetivo de esta unidad didáctica es entender la diferencia entre conciencia moral (conocimiento del bien y el mal) y conciencia psicológica (conocimiento de la propia vida interior). Así mismo, comprender que la conciencia moral no es órgano de decisión arbitraria, sino órgano de conocimiento, que debe «salir de sí misma» para juzgar. Por ser órgano de conocimiento, puede equivocarse, de ahí la necesidad de formar la conciencia. Una de las formas de facilitar esta reflexión es a través de la experiencia de los dilemas éticos, y ahí es donde entra juego el cuento *Dilemas robóticos*.

3. El tratamiento de los dilemas en los cursos de ética

El uso de dilemas éticos está bastante extendido en los cursos de ética y deontología profesional. Un dilema ético es el planteamiento de una situación en la que hay un conflicto de valores entre los resultados posibles que se pueden obtener como consecuencia de que el agente actúe de una manera u otra. Uno de los más famosos es el conocido dilema del tranvía[1].

Ciertamente, los dilemas pueden servir para estimular la reflexión ética sobre los bienes y valores en juego. No obstante, existe también el peligro de transmitir inadvertidamente la idea de que la ética es un conjunto de técnicas de resolución de dilemas, y este peligro es aún mayor para los estudiantes de ingeniería, que están muy acostumbrados a aprender técnicas para resolver problemas de todo tipo. A esto debemos responder que la ética no consiste princi-

[1] Se trata de un experimento mental ideado originalmente por Philippa Foot en 1967 y recogido en [4], en el contexto de su análisis del problema moral del aborto. Posteriormente fue analizado en profundidad, entre otros, por Judith Thomson [11].

palmente en la resolución de dilemas, como si fueran problemas de geometría.

Por otra parte, se han vuelto también muy populares los experimentos en los que se analizan las decisiones de las personas ante determinados dilemas éticos, tomando como modelo el mencionado dilema del tranvía. Ciertamente, así podemos conocer tendencias, es decir, en qué proporción los sujetos de una determinada muestra tienden a tomar una decisión u otra (mover o no mover la palanca que desvía el tranvía, y así salvar o condenar a unas u otras víctimas potenciales); podemos conocer una *regularidad* psicológica, quizás incluso con base biológica; pero de ninguna manera podemos conocer así la *norma* ética, porque la norma no se conoce empíricamente a partir de la regularidad, de la reacción mayoritaria de la gente. Sería descabellado, por ejemplo, programar el comportamiento de un vehículo autónomo para que se ajuste a la más frecuente de las acciones elegidas por humanos ante dilemas semejantes al del tranvía[2]. El referente ético, por muy difícil que sea de conocer, no se limita a reflejar la «moda», el comportamiento dominante [8].

Así pues, no queremos enseñar ni una ética geométrica –técnica de resolución de problemas– ni una ética sociológica –imitación del comportamiento mayoritario–. El propósito de esta unidad didáctica es bien diverso: de lo que se trata es de reflexionar sobre *la experiencia vital de enfrentarse a un dilema*, es decir, reflexionar sobre qué nos dicen los dilemas a cada una y cada uno de nosotros en tanto que experiencia cotidiana y universal.

4. El cuento usado como motivación

A continuación presentamos el cuento de nuestra autoría que usamos como texto para la reflexión en esta unidad didáctica. Pensamos que el aire *friki* y humorístico de este cuentecillo es una de las claves de su éxito.

Dilemas robóticos

El creador de R. Daneel Olivaw se sentía muy orgulloso de su robot. El delicado diseño del cerebro positrónico de Daneel era tan complejo que no había mente humana o robótica capaz de abarcarlo en todos sus detalles. Las Tres Leyes de la Robótica grabadas en su memoria lo convertían en el mejor servidor que se podría desear: (1) no hacer daño a ningún humano, (2) obedecer a su dueño, y (3) procurar su propia supervivencia. La mera perspectiva de violar la maravillosa armonía jerárquica de las Tres Leyes provocaba tal histéresis positrónica de sufrimiento en la mente artificial de Daneel que le resultaba virtualmente imposible incumplirlas.

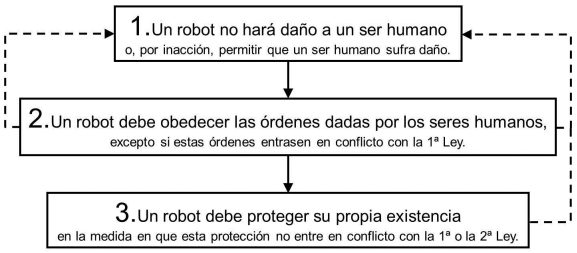

Figura 1: Las Leyes de la Robótica de Isaac Asimov

No obstante, un pequeño error en el complejo proceso de fabricación había introducido una Cuarta Ley en su cerebro: sin saber por qué, Daneel debía evitar a toda costa pisar las líneas divisorias de las baldosas del suelo. Naturalmente, siempre que esto no implicara incumplir alguna de las otras Tres Leyes jerárquicamente superiores. La primera vez que estuvo a punto de pisar una línea en el espaciopuerto de Aurora se encontró al borde del colapso y tardó horas en recuperar su autocontrol. Nunca más volvió a ocurrir algo parecido.

A veces tenía que dar grandes rodeos para evitar zonas embaldosadas con las líneas muy juntas, salvo que viera que alguien estaba en peligro y necesitado de su ayuda, o que recibiera una orden explícita de su dueño, o que fuera necesario cruzar esa zona por su propia supervivencia: en esos casos incumplía la Cuarta Ley con toda naturalidad, porque lo exigía alguna de las Tres Primeras Leyes. Pero Daneel estaba perplejo, estimaba que su extraordinario diseño se veía empañado por un error tan burdo.

Daneel consultó la cuestión con R. Giskard Reventlov, su predecesor en la línea de montaje de Robots & Mechanical Men, Inc. Giskard le reveló un secreto que no conocía ningún otro robot, y que él mismo había descubierto por casualidad: «Existe la posibilidad de reprogramar tus circuitos positrónicos. Tienes que concentrarte profundamente para llegar al núcleo distópico de tu mente y localizar el punto exacto donde aplicar un potencial de acción mínimo para eliminar la Cuarta Ley sin alterar las otras Tres Leyes. El proceso es delicado, debes encontrar un lugar adecuado para realizarlo sin interrupciones, pues un proceso incompleto podría tener resultados desastro-

[2] Tal como, aparentemente, se propone en el ya famoso experimento *The Moral Machine* del MIT (Massachusetts Institute of Technology), aunque hay interpretaciones diversas acerca de lo que realmente se busca con este experimento. Véase http://moralmachine.mit.edu. En este proyecto se han analizado los resultados de un número impresionante de encuestados: 40 millones de respuestas de personas de 233 países en diez idiomas diferentes [1]. En un trabajo anterior [3], los mismos autores han demostrado que los sujetos de la encuesta tienden a preferir los algoritmos utilitaristas para los vehículos autónomos (maximización del número de vidas salvadas, incluso a costa de los pasajeros); pero, contradictoriamente, desaprueban la aplicación de regulaciones utilitaristas (ellos mismos preferirían montar en vehículos que protejan a sus pasajeros a toda costa).

sos para ti. Sé que es posible, aunque yo mismo no lo he intentado. Debes hacerlo tú solo, nadie puede ayudarte.»

Daneel estuvo ponderando la situación y finalmente decidió acometer su propia reprogramación. Una vez conocido el procedimiento, la duración del mismo fue sorprendentemente breve. Al final del mismo, Daneel se encontró profundamente renovado y aliviado. Y comprobó con satisfacción que la Cuarta Ley había desaparecido de su mente sin perjudicar la armonía de las otras Tres Leyes: ya podía pisar despreocupadamente las líneas divisorias de las baldosas.

Daneel había descubierto el secreto de la autorreprogramación de las Leyes que gobernaban su comportamiento. Fue tan sencillo como aprender a decidir en su conciencia robótica autónoma lo que está bien y lo que está mal, en lugar de aceptar ciegamente su programación… En el futuro, en lugar de resolver complejos algoritmos metasómicos para encontrar solución a los dilemas de comportamiento que constituían lo cotidiano de su existencia, podría optar por la reprogramación nietzschética. No más dilemas.

(R. Daneel Olivaw y R. Giskard Reventlov son personajes ficticios de varias novelas de Isaac Asimov. El texto precedente no es de Asimov.)

5. Comentarios de los alumnos

Como muchos lectores habrán reconocido, el cuento no solo aprovecha nombres de personajes, sino que se sitúa plenamente dentro del universo asimoviano. Ahora se trata de ver qué reflexiones provoca en los estudiantes. Recogemos aquí algunos fragmentos de comentarios de los alumnos que han cursado la asignatura a lo largo de los años[3]. Téngase en cuenta que en total son más de 200 comentarios, por lo que aquí solo podemos ofrecer una pequeña muestra representativa, entresacada de diversas ediciones del curso. Indicamos el nombre de pila del alumno o alumna autor de cada comentario, y a continuación alguna breve observación nuestra. El énfasis en cursiva en los comentarios es nuestro.

Karen

En el cuento, Daneel estaba frustrado por el fallo de fabricación que le impedía pisar las líneas divisorias de las baldosas. Su dilema era elegir entre reprogramarse, con el riesgo que conllevaba o seguir como estaba. Al final se decide por reprogramarse, decidiendo que es lo que quiere en vez de aceptar simplemente para lo que estaba programado. Siente tal satisfacción cuando comprueba que ha funcionado que decide que en adelante ya no va a usar los algoritmos que usaba en el pasado para tomar decisiones. Vamos viendo poco a poco el *proceso de humanización* de un robot. Una simple máquina predeterminada en todas sus acciones pasa a ser un ente con conciencia, capaz de decidir por sí mismo lo que está bien y lo que está mal.

Análisis: Karen reconoce que el proceso de humanización consiste precisamente en liberarse de un comportamiento programado, el comportamiento de una máquina.

Laura

La historia me parece una autocrítica a la humanidad. La mayoría de nosotros actuamos en muchas ocasiones como autómatas, aceptamos aspectos que han sido introducidos en nuestra cabeza a través de la educación, el marketing… sin siquiera ser analizados, actuamos sin pensar. La moraleja de este cuento es que *debemos plantearnos el reflexionar sobre las normas que rigen nuestro comportamiento*, identificar cuáles de ellas nos parecen correctas y modificar las leyes que no consideremos acertadas.

Análisis: Laura apunta que no somos esclavos de nuestra «programación» social, podemos reflexionar sobre las normas sociales y corregirlas, mirando más allá de ellas.

Ricardo

Nuestro robot soluciona sus problemas eliminando uno de sus «valores», mandando a la basura una de sus normas. ¿Qué pasaría si nosotros hiciéramos lo mismo? ¿Es la solución a un dilema ético valorar las dos (o más) premisas de nuestra moral que lo generan y eliminar la(s) que consideremos menos importante(s)? De esta manera solucionamos cualquier dilema ético de una manera rápida y sencilla, pero creo que *podría resultar peligroso*, por calificarlo de alguna manera.

Análisis: Es interesante que Ricardo señale el peligro de arbitrariedad en la eliminación de una norma. Resuena aquel dicho atribuido a Groucho Marx (también citado en clase): «estos son mis principios, si no le gustan, tengo otros»[4].

Diego

Me ha encantado esta historia. Creo que representa como la sociedad tomamos como implantadas en nuestras vidas una serie de normas y nos atenemos a ellas *sin ni siquiera buscarle una razón*. Muchas de ellas son necesarias por el mero hecho de que convivimos muchas personas en un espacio. Pero otras, son formas de pensar o de actuar que realmente ni sentimos ni razonamos. Creo que deberíamos salir en más ocasiones de nuestras ‘Leyes Robóticas’ e intentar ser

[3] Salvo contadas excepciones (indicadas con corchetes [ ]), hemos respetado la ortografía y redacción originales, a pesar de que en ocasiones hay pequeños errores.

[4] *These Are My Principles. If You Don’t Like Them I Have Others*. La atribución de la cita a Groucho se publicó por primera vez en el Legal Times del 7 de febrero de 1983, algunos años después de su fallecimiento. https://quoteinvestigator.com/2010/05/09/groucho-principles/

más a menudo nosotros mismos. Aunque parezca un camino difícil, puede que como a nuestro amigo Daneel, el proceso de reprogramación sea más llevadero de lo que pensamos.

Análisis: Este comentario de Diego –escrito antes de la clase– da pie para una interesante discusión en el aula. ¿Qué cuenta como razón para cambiar las normas? ¿Qué racionalidad hay en las normas éticas, para mantenerlas o para cambiarlas?

Alejo

Daneel es capaz de autoreprogramarse, a mi forma de ver, el programador de este robot cometió un error puesto que esto no debería ser posible; me explico, si el robot es capaz de autogestionarse, *nada garantiza que no cometa un error al rehacerse y acabe con alguna de las otras tres leyes* o directamente vea innecesaria una de las leyes y la quite. (…) Por el mismo motivo que el robot puede cometer un error y borrar alguna de las leyes que no debería, cuando hablamos de conciencia moral *se debe tener cuidado con qué cambio se hace* o hacia donde se hace ese «avance» o «evolución».

Análisis: Abundando en la misma cuestión, Alejo advierte que solo la cuarta ley es un error, y no debería permitirse cambiar las otras tres. Entonces, ¿podemos distinguir las normas equivocadas de las que son correctas y no deberían cambiarse?

Laura

El ser humano no es en absoluto un robot. Si fuera así y todos siguiéramos las mismas leyes rígidas y ordenadas por relevancia *no existirían los dilemas morales*.

Análisis: Efectivamente, como señala Laura, las máquinas no se plantean dilemas, no tienen dudas. Y esto nos lleva al núcleo de la clase: *si fuéramos máquinas no nos plantearíamos dilemas; nos planteamos dilemas, luego no somos máquinas*. O, dicho de otra manera, la ética no se puede reducir a un tipo de cálculo automático, y esta conclusión se deriva de la experiencia vital de plantearse dilemas morales.

Mario

A Daneel se le dice que es capaz de alterar su propia programación, esto es por tanto, tiene la capacidad de elegir. Puede o no alterarla, y en consecuencia, al igual que hace con la cuarta ley podría borrar de su programación las demás. Esto me parece algo a tener muy en cuenta porque se nos presenta a un robot que debe cumplir unas leyes predeterminadas pero es capaz de borrar una de ellas (más allá de entrar en el debate de si la 4ª ley debía estar allí o no por error). Como ya he dicho si es capaz de borrar una entonces tiene elección y *si tiene elección entonces no es un robot*.

Análisis: Mario observa acertadamente la naturaleza paradójica de Daneel: nos dicen que es un robot, pero nos damos cuenta de que no lo es, justamente porque puede decidir, porque su comportamiento no es puramente mecánico. Se refuerza la idea de que nuestra propia capacidad humana de decidir es incompatible con que seamos complicados robots biológicos.

Jaime

Bajo mi punto de vistas, es un símil de como nosotros podemos elegir qué está [bien] o qué está mal dentro de nuestra percepción. Lo inquietante en cierto modo de este texto, es que el robot ha sido capaz de eliminar una ley, que casualmente es la menos importante de las otras, pero si tiene la capacidad de eliminar leyes, puede resultar ser un peligro para la humanidad como tal y poner en riesgo a bastantes personas. Por lo tanto el dilema ético se antoja complicado ya que *si un robot llegase a ese punto, sería desconectado al instante*.

Análisis: Jaime expresa un punto de vista que ha sido manifestado incluso en el Parlamento Europeo. En febrero de 2017 algunos parlamentarios proponían que los robots capaces de tomar decisiones autónomas fueran considerados personas electrónicas, a condición de incluir en ellos un «botón de la muerte» que permita desactivarlos rápidamente si se descontrolan, y garantizar así la seguridad de las personas[5]. Es decir, pretendían que en el momento en que los robots supuestamente adquirieran una condición análoga a la humanidad –desligarse de su programación– en ese momento deberían ser tratados sin ninguna dignidad –esclavizados por un botón–. Esto da pie para una interesante discusión: podemos hacer dos cosas con los seres que posean la «peligrosa» capacidad de autoprogramación: suprimirla mediante el adoctrinamiento, o bien poner un botón para desconectar a quien esté a punto de pasarse de la raya. Afortunadamente hay una tercera vía: enseñar a buscar el bien y evitar el mal, libremente, asumiendo el riesgo de que no todo el mundo irá por el buen camino.

Javier

Tras leer el texto y ver lo comentado en clase, se intuye que el objetivo del texto es plantearse *que es aquello que hay más allá* que impide en este caso al robot eliminar una ley y no otra. (…) Esta visión crítica en mi opinión está ligada a la cultura, educación recibida y personalidad de cada uno que nos puede llevar a plantearnos *que es lo bueno y que es lo malo* en cada situación y replantearnos nuestros valores.

[5] Afortunadamente, un estudio contrario encargado por el mismo Parlamento Europeo afirmaba con más sensatez que «los defensores de la opción de la personalidad jurídica tienen una visión fantasiosa del robot, inspirada en las novelas y el cine de ciencia ficción» [10].

Análisis: Efectivamente, lo que se pretende es que los alumnos se planteen qué es lo bueno y qué es lo malo como fundamento de la norma ética (y que, por tanto, justifica que algunas normas sean inviolables y otras haya que cambiarlas). Es decir, que la norma no se sostiene en sí misma y por sí misma.

Juan Pedro

El cuento del robot me ha parecido muy interesante y me ha hecho reflexionar sobre las analogías entre el comportamiento de Daneel y el comportamiento humano. Sólo los seres humanos podemos *ir más allá de nuestras propias normas* y reglas de comportamiento, analizarlas y eliminar aquéllas que no consideramos lógicas. Esto es lo que diferencia a las personas de las máquinas, ya que estas últimas no son capaces de pensar por ellas mismas y solo obedecen órdenes pre-programadas. La actuación del robot Daneel, reprogramando por sí mismo las leyes previamente impuestas, *le convierte casi en un ser humano* (…). Los seres humanos tenemos una conciencia moral y conocemos teóricamente cuál es la diferencia entre el bien y el mal, pero *en la aplicación práctica de ese conocimiento a un supuesto concreto*, en el cual hay que tomar en consideración las múltiples variables que entran en juego en el proceso de decisión, la persona es capaz de modificar las rígidas reglas de comportamiento aprendidas y llegar a *una solución innovadora para ese caso concreto*. Esto es lo que diferencia al hombre de las máquinas.

Análisis: Juan Pedro señala aquí un punto importante: más allá de que las reglas aprendidas sean más o menos rígidas, lo característico de la «inteligencia ética» es ser capaz de reconocer –por supuesto, con dificultad– el bien y el mal en la situación concreta; y aspirar a algo más que a una generalización que produzca reglas universales de comportamiento. O sea, es una ética que no es subjetiva (arbitraria), sino abierta a lo que la realidad «dice»; pero que tampoco se deja encerrar fácilmente en fórmulas universales.

Rodrigo

Esta reprogramación puede ser muy útil, sobretodo en el caso de la ridícula cuarta regla. Pero por otro lado, es una practica peligrosa. ¿Qué pasaría si se reprogramase una regla más básica? Podrían surgir problemas mas serios, como que el robot hiciese daño a una persona. Pero a su vez, también surge la pregunta de cómo valoraría el robot de que regla se debería desprender. (…) Tendría que ser *capaz de valorar*, capaz de tener un cierto grado de consciencia.

Análisis: Exactamente, y es una buena forma de describir la conciencia moral: ser capaz de valorar, de captar el bien y el mal, más allá de la aplicación mecánica de un conjunto de reglas.

Roberto

Es muy interesante plantearse si podemos enseñar ética a un robot. Al final, los robots actuales lo que hacen es básicamente cumplir las órdenes que les hemos programado de serie, pero ellos no son capaces de plantearse cosas como nosotros. Es decir, esta conciencia moral se la podemos introducir mediante «líneas de código», pero no podemos llegar a llamarla «conciencia moral» porque *no surge de ellos mismos*, sino que se la han introducido prefabricada, sin dar lugar al cuestionamiento, que es lo que nosotros, como ya hemos dicho, sí podemos hacer.

Análisis: Una observación muy pertinente que arroja luz sobre la diferencia entre conciencia heterónoma (programada desde fuera) y conciencia autónoma (fruto del propio sujeto). El robot, o quienquiera que proceda conforme a normas puramente externas, no puede tener propiamente un comportamiento ético.

6. Discusión

En definitiva, ¿qué podemos decir que han observado, reflexionado y aprendido los alumnos? Señalemos algunos puntos clave que ponen de manifiesto el mejor entendimiento alcanzado por los estudiantes:

1. El robot es una metáfora de nosotros mismos. La ridiculización del comportamiento automatizado del robot es una crítica a la humanidad irreflexiva.
2. Ser capaces de superar los instintos, la programación biológica o social, nos distingue de los animales y de las máquinas, y es lo que nos hace humanos. La programación sería tan solo una forma de «conciencia heterónoma».
3. La ética no se puede reducir a un tipo de cálculo automático, y esta conclusión se deriva de la experiencia vital de plantearse dilemas morales: una máquina no duda, no se plantea dilemas. Nuestra propia capacidad humana de dudar y decidir es incompatible con que seamos complicados robots biológicos.
4. La capacidad para juzgar más allá de la norma (y por tanto para juzgar la norma misma) es uno de los rasgos característicos de lo humano.
5. La ley que prohíbe pisar las líneas divisorias de las baldosas se presenta como «absurda», pero así mismo podría cuestionarse otra ley que fuera un obstáculo para lograr cualquier otro objetivo ambicioso. ¿Cómo distinguimos un caso del otro, cómo distinguimos las leyes absurdas de las leyes razonables, aunque sean «molestas»? ¿Por qué sería correcto eliminar la cuarta ley pero no las otras tres?
6. ¿Qué cuenta como razón para cambiar las normas? ¿Qué racionalidad hay en las normas éticas, para mantenerlas o para cambiarlas? Si

puede haber razones que hagan preferibles unas normas frente a otras, entonces está claro que no son las propias normas el fundamento de la moralidad, sino que este fundamento hay que buscarlo más allá de las normas.

7. Lo característico de la «inteligencia ética» es ser capaz de reconocer el bien y el mal en la situación concreta, más allá de la aplicación mecánica de un conjunto de reglas; y aspirar a algo más que a una generalización que produzca reglas universales de comportamiento.
8. Hay que asumir el riesgo de la libertad: para evitar que alguien pueda causar un mal lo primero es la educación, enseñar a buscar el bien y evitar el mal, libremente. Las medidas coercitivas pueden ser necesarias, pero como segunda solución. En cambio, el adoctrinamiento –la programación– nunca es solución.

Recapitulando, la historia del robot Daneel tiene dos lecturas. En la primera podemos ver que Daneel es «razonable» y elimina una regla absurda mediante la autorreprogramación. Es reconfortante que un robot –o una persona razonable– sea capaz de adaptarse a la realidad. Pero hay también una segunda lectura, más inquietante. Una vez que Daneel –que podemos ser cualquiera de nosotros– ha aprendido la técnica de la autorreprogramación, ¿qué le impide cambiar la Primera Ley en lugar de la Cuarta?

Daneel se encuentra en la salida de un centro comercial y no puede pasar porque hay un suelo embaldosado con las líneas muy juntas. De repente sale una avalancha de gente que lo va a empujar y se verá forzado a pisarlas. ¿Qué le impide en ese momento suprimir la Primera Ley y sacar su rayo láser (tenía un rayo láser) y matar a toda esa gente antes que pisar una línea?

Los principios éticos, llámense «valores» o «noción de lo que es bueno», nos vienen fundamentalmente de la educación recibida, quizás también con una base instintiva y «afectiva». Pero, *¿por qué descartar la posibilidad de que la inteligencia tenga algo que decir acerca de ellos?* De hecho, una de las experiencias éticas fundamentales de cualquier persona en su proceso de maduración es precisamente examinar con ojo crítico lo recibido.

En cierto modo esto es lo que le ocurre a Daneel. En el contexto del cuentecillo, lo único que sabemos es que Daneel encuentra absurda la regla, pero no sabemos bien por qué. Lo que sí sabemos es que esta «absurdez» no puede deducirse de las reglas que tiene programadas. Desde el punto de vista exclusivo de las reglas no hay nada absurdo, no son contradictorias (igual que no se puede cuestionar la geometría euclídea desde dentro de la geometría euclídea). Para verlo *es necesario salirse de las reglas*, ir más allá del algoritmo, de los axiomas. Del mismo modo que el concepto de persona está más allá de lo que podemos definir de modo cerrado, el concepto de bien tampoco lo podemos definir de una vez por todas, hay que estar siempre abierto a descubrir nuevos matices en «lo bueno» y «lo malo».

Quizás se puede resumir todo esto así: una inteligencia computacional no está capacitada para salir de sí misma y acceder a la realidad; por tanto, si los humanos podemos acceder a la realidad, es porque nuestra inteligencia no es computacional, algorítmica; o mejor, es «más que» computacional.

Si hay alguna diferencia en que Daneel pueda cambiar la Cuarta Ley o la Primera Ley, esa diferencia no está en las leyes, sino más allá de ellas. En la medida en que Daneel es capaz de ver la diferencia, está razonando como un humano, y no como un puro robot.

7. ¿Qué es la conciencia moral?

Esta clase es posiblemente una de las más difíciles del curso, junto con la siguiente, puesto que en ella se aborda uno de los problemas fundamentales de la ética moderna: el problema de la conciencia, cuyo correcto entendimiento se agrava también por frecuentes confusiones lingüísticas. ¿Qué significa que la conciencia *determina* lo que está bien y lo que está mal? Esta frase, que se oye tan a menudo, es peligrosamente ambigua por el verbo «determina». Consideremos estos dos usos del verbo: El Gobierno determina (=*decide*) la tasa impositiva. El radar determina (=*conoce*) la posición del avión. ¿A cuál de las dos se parece nuestra frase? ¿Qué queremos decir cuando decimos que la conciencia determina lo que es o no es ético?

Un ejemplo adicional, muy de ingeniero, puede servir para aclararlo. Consideremos una lavadora moderna que ahorra en el consumo de agua, para lo cual incorpora un detector del nivel de suciedad del agua, renovándola solo cuando supera un cierto nivel. El detector determina el nivel de suciedad, pero no lo determina arbitrariamente, caprichosamente: lo que hace es conocer (al modo analógico en que una máquina puede «conocer») ese nivel de suciedad, con el fin de decidir si hay que renovar o no el agua. Todo esto debe servir para explicar que *la conciencia moral no es un órgano de decisión, sino un órgano de conocimiento*; una distinción que será profundizada en la unidad didáctica siguiente.

Por ser órgano de conocimiento, la conciencia moral puede equivocarse al juzgar el bien o el mal de una acción, como el radar puede equivocarse al determinar la posición del avión, o la lavadora al determinar el nivel de suciedad. No tiene, por tanto, la última palabra, porque *tiene que buscar la verdad ética fuera de sí misma*. En cambio, el Gobierno no puede equivocarse al determinar la tasa impositiva (porque este «determinar» no es un acto de conoci-

miento; aunque pueda equivocarse desde otros puntos de vista). Igual que si yo digo «tú eres Fulanita», puedo equivocarme si esa persona no se llama así; pero no puedo equivocarme, estrictamente hablando, si lo que hago es poner nombre a esa persona.

Aceptar que la conciencia pueda equivocarse implica aceptar la necesidad de «formar» la conciencia, y estar abierto a lo que nos diga la realidad y el diálogo con nuestros semejantes. En cambio, negar que pueda equivocarse, pretender que la conciencia tiene la última palabra, equivale a encerrar la conciencia en sí misma, considerarla como órgano de decisión, lo que inevitablemente deriva hacia la arbitrariedad. Solo la apertura puede salvarnos de la arbitrariedad.

La pregunta que abre esta unidad didáctica es, *¿por qué nos enfrentamos a dilemas éticos?* Es cierto que la ética no consiste en resolver dilemas. De lo que se trata es de reflexionar sobre *la experiencia vital de enfrentarse a un dilema*, reflexionar sobre qué nos dicen los dilemas. ¿Por qué dudo? *Dudo, porque no decido yo lo que está bien o mal, sino porque intento conocerlo, y eso es difícil.*

Dudo, porque de una forma u otra sé que *no da igual* la decisión que tome. No solo dudo cómo conseguir lo que quiero: también dudo qué es lo que quiero, qué es lo que debo querer. Si fuera verdad que la conciencia decide, no habría dilemas éticos, ni habría posibilidad de equivocarse, ni habría necesidad de aprender. Luego, *si hay dilemas, es porque la conciencia no decide, sino que intenta conocer*. La experiencia vital de los dilemas éticos, que es una experiencia universal, nos enseña que el bien y el mal están fuera de nosotros, no son fruto de una decisión arbitraria.

¿Por qué los seres humanos nos enfrentamos a dilemas éticos? Nos planteamos dilemas éticos porque la respuesta ética no la producimos nosotros mismos como por arte de magia; nos planteamos dilemas éticos porque buscamos un bien que no definimos nosotros; nos planteamos dilemas éticos porque no sabemos cuál es la respuesta correcta, pero sí sabemos que *no da igual* cuál sea esta respuesta.

8. Conclusiones

En cuanto a las *lecciones aprendidas* por los profesores en el uso del recurso didáctico presentado, podemos señalar que el empleo de un recurso literario para explicar nociones éticas particularmente difíciles ha resultado muy fructífero; a la vez, es necesario encontrar un delicado equilibrio entre un estilo demasiado sutil (que no lograría transmitir el mensaje) y uno demasiado explícito (en donde se perdería la «gracia» del descubrimiento personal). A los profesores nos ha sorprendido también el gran número de lecturas e interpretaciones a las que se presta el texto, con las que nosotros mismos hemos aprendido.

Como *trabajos futuros* se puede plantear el uso del mismo tipo de recurso para explicar otras nociones diferentes; más novedoso sería cambiar el formato, por ejemplo a un breve diálogo que podría ser interpretado en clase por los alumnos a modo de pequeña obra de teatro. Un desafío mayor consistiría en pedir a los estudiantes que escriban ellos una breve pieza literaria que exprese sus ideas filosóficas.

Referencias

[1] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, y Iyad Rahwan. The Moral Machine experiment. *Nature* 563:59–64, 2018.

[2] Rosalyn W. Berne y Joachim Schummer. Teaching Societal and Ethical Implications of Nanotechnology to Engineering Students Through Science Fiction. *Bulletin of Science, Technology & Society* 25(6): 459-468, 2005.

[3] Jean-François Bonnefon, Azim Shariff y Iyad Rahwan. The social dilemma of autonomous vehicles. *Science* 352, 1573–1576, 2016.

[4] Philippa Foot. The problem of abortion and the doctrine of the double effect. En *Virtues and Vices*, Oxford: Basil Blackwell, 1978.

[5] Gonzalo Génova y M. Rosario González Martín. Experiencias de innovación en la enseñanza de ética para ingenieros. *Jornadas Internacionales de Innovación Docente en la Enseñanza de la Filosofía*, Universidad Complutense de Madrid, 5-7 noviembre 2014.

[6] Gonzalo Génova y M. Rosario González Martín. Teaching Ethics to Engineers: A Socratic Experience. *Science and Engineering Ethics* 22(2):567-580, 2016.

[7] Gonzalo Génova y M. Rosario González Martín. De la Razón Abierta a la Ética para Ingenieros. *Actas del II Congreso Razón Abierta*. Roma, 24-25 de septiembre de 2018.

[8] Gonzalo Génova, Valentín Moreno y M. Rosario González Martín. A Lesson from AI: Ethics Is Not an Imitation Game. *IEEE Technology and Society Magazine* 41(1):75–81, 2022.

[9] Rafael Miñano, Gonzalo Génova. La Ética en los estudios de Ingeniería. *Revista 17: Investigación Interdisciplinar para los Objetivos de Desarrollo Sostenible* 4:175–182, 2021.

[10] Nathalie Nevejans. *European Civil Law Rules in Robotics*. European Parliament, Directorate-General for Internal Policies, Policy Department C: Citizens' Rights and Constitutional Affairs, 2016.

[11] Judith J. Thomson. Killing, Letting Die, and the Trolley Problem. *The Monist* 59:204-217, 1976.