Apuntes sobre los aspectos de valor prescriptivo del razonamiento abductivo

Alger Sans Pinillos[153]

Resumen

En este trabajo presento el papel de la abducción a la hora de analizar la interacción entre seres humanos y máquinas. La propuesta parte del modelo eco-cognitivo de la abducción, el cual plantea una relación de intercambio de información, la cual implica que nuestro entorno es cada vez más sofisticado, en la medida que lo usamos. Este hecho nos obliga a plantearnos qué tipo de interacción deseamos con estos dispositivos y analizar si estas transacciones son neutras o si, por el contrario, contienen elementos valorativos que, sino tenemos en cuenta, podemos acabar cometiendo grandes injusticias. La idea central que intento argumentar es que la abducción contiene de base un elemento valorativo de tipo prescriptivo el cual, por un lado, nos obliga a considerar ciertos aspectos de nuestra concepción del conocimiento y, por el otro, nos permite analizar el problema de las máquinas destinadas a ayudarnos en cuestiones éticas. En estos contextos, es necesario preguntarnos de qué manera consideraremos la armonización entre los valores humanos y los que computemos, hecho que nos ha de obligar a revisar los modelos de representación humanos que luego usamos para diseñar nuestras máquinas. Investigaciones recientes permiten plantear la abducción como un elemento clave para la investigación, pues es el punto de unión entre todos estos temas. Ofrece la posibilidad de implementar cuestiones de valor, además permite desencajar problemas técnicos, como lo es el debate entre sistemas miméticos o de anclaje en la alineación de valores y, finalmente, es una herramienta útil para denunciar las injusticias generadas por la falacia naturalista que se comete al confundir valores con preferencias, a través, precisamente, de darnos herramientas para disolver la dicotomía clásica de la que parte tal error lógico.

Palabras clave: mimesis, falacia naturalista, abducción, interacción.

[153]Universidad de Pavía, Pavía, Italia: alger.sanspinillos@unipv.it

7.1. Introducción

El hecho de que cada vez más nuestras herramientas sean dispositivos tecnológicos con un determinado grado de inteligencia (máquinas) en vez de artefactos, nos obliga a replantearnos, no solamente esta circunstancia, sino todo un seguido de cuestiones que han ido adquiriendo importancia en la medida que nuestra dependencia hacía ellos también ha argumentado mediante la continua integración en cada uno de los aspectos de nuestro día a día.

La diferencia más remarcable es que nosotros podemos interactuar con estos dispositivos, mientras que este grado de relación es imposible con los simples artefactos[154]. No obstante, no se ha de confundir el tipo de relación compleja que estoy exponiendo con la creación y uso de los artefactos tecnológicos, los cuales nos han servido de herramientas para realizar un tipo de tareas complejas. Aquí me refiero exclusivamente a esas entidades que atienden a nuestras necesidades mediante la asimilación de datos para, así, ofrecer experiencias personalizadas al usuario. Esto es, el paradigma computacional, en el que la diferencia crucial es que estos dispositivos nos ofrecen un entorno al que ya no podemos simplemente investigar, sino que lo podemos interrogar y, éste, a su vez, puede hacer lo mismo con nosotros.

No obstante, este fenómeno también ha abierto las puertas a nuevas problemáticas. Un ejemplo es la cuestión sobre la estandarización del interés y la construcción de un conocimiento general, causado a través de los algoritmos que filtran las búsquedas particulares para optimizar resultados generales (*Big data*). Este hecho nos obliga a plantearnos si estamos viviendo una hegemonía global del conocimiento o si, por el contrario, realmente tenemos en nuestras manos el camino hacía una emancipación epistémica.

Un recurso habitual es el de la *concienciación* para el *uso responsable* de la tecnología. Palabras que guardan poca o ninguna relación con la realidad a la que me refiero pues, una cosa es el asunto praxeológico dirigido al *uso* de cualquier objeto (sea dispositivo, artefacto, así como con el grado que fuere, tanto de materialidad como de simbolismo) y otra muy diferente es cuando se está tratando el tema de la interacción. Ciertamente, hay una diferencia entre el tipo de relación que tenemos cuando analizamos la eficacia de nuestras herramientas con las cuestiones prescriptivas que versan sobre los límites éticos de la forma en que incidimos en el mundo y que contabilizan y determinas de manera diferente en el total del cómputo que hacemos al decidir qué hacer; a saber, la mirada ética. No obstante, es una mirada ética en que los instrumentos se pueden hacer equivaler a nuestras acciones, a las maneras como actuamos y esto, en definitiva, dirige el problema ético a la esfera del ser humano. Ejemplos son las éticas hacía el medio ambiente, en la gestión de datos. . .

Por el contrario, en el momento en que involucramos la interacción conside-

[154]Imposibilidad en el sentido de que el sentimiento de interacción con estas entidades no es más que una ficción, por el simple hecho de que no hay la correspondencia que este tipo de relación demanda para darse.

ramos que la relación que establecemos con esa entidad es más o menos correcta y, por ende, que podemos exigirle cierta eticidad. Un ejemplo son los animales los cuales, una vez que no los consideramos como meros objetos y/o propiedades, entendemos que nuestra interacción con ellos ha de contemplar sus características y que, en definitiva, no debemos tratarlos como simples cosas que sirven para nuestros fines. Este paso implica aceptar a estas entidades dentro de nuestro constructo ético, del cual ahora me refiero a la máxima comúnmente aceptada: "Handle so, dass du die Menschheit sowohl in deiner Person, als in der Person eines jeden andern jederzeit zugleich als Zweck, niemals bloß als Mittel brauchst" (Kant, 1960, AA IV). Evidentemente, el paso no es el sinsentido de considerar estas entidades como personas, sino identificarlas dentro de nuestra mirada ética del mundo. De hecho, el "como si" es una petición de principio cruel que compara y nivela a partir de una vara de medida escogida arbitrariamente.

La idea que pretendo defender en este escrito es que el paradigma en el que nos encontramos actualmente, no solamente permite, sino que demanda una ampliación de nuestra mirada ética y contemplar algunos dispositivos tecnológicos como entidades con las que interactuamos. El motivo es doble. El primero es que investigamos y disponemos de dispositivos tecnológicos que usamos para solucionar problemas éticos, mientras que el segundo es que tenemos otros dispositivos que, aunque no están directamente diseñados para que incidan directamente en esta faceta de nuestras vidas, la manera como interactúan con nosotros no obliga a plantearnos si nos ofrece un trato ético. Los dos temas convergen en que todas estas entidades son hechas por nosotros, hecho que ayuda a concentrar nuestros esfuerzos en analizar los modelos que se usan para generar los sistemas con los que después tendremos que convivir.

Estos modelos han sido creados con la intención de integrar sistemas complejos de acumulación de datos para dar respuestas precisas. Esto es, como herramientas sofisticadas para informar de la manera más personalizada y concreta posible. Como he dicho al principio del texto, la gestión de estos datos se optimiza mediante la estandarización del interés, el cual se genera mediante la accesibilidad de los mismos datos recompilados. No obstante, la cuestión ética implica un problema más pues, por un lado, no puede responderse mediante una acumulación de datos y, segundo, porque implica una interacción real, esto es, una adaptación a casos concretos en que los cambios mínimos en las demandas es lo que realmente importa, la cual requiere, no solamente de modelos computacionales que permitan que los dispositivos devengan morales, sino también el paso de que nosotros aceptemos este grado de eticidad dentro de nuestra mirada ética del mundo. Esto es, que reconozcamos estos dispositivos como entidades éticas.

Como se puede ver, el segundo problema se desprende del primero, pero solamente en parte. Como bien nos muestra nuestra historia, nada garantiza que, llegados al punto que consigamos una máquina ética de verdad, esta sea aceptada como entidad con la que interactuar moralmente. No obstante, esta es una espe-

culación que no toca ser planteada ahora[155]. El primer punto es el que realmente compete a este escrito y se abordará del siguiente modo. En el segundo apartado presentaré una pincelada a la confusión entre valores y preferencias que se da en las teorías relativas a la Moral Machine (MM). En la tercera parte abordo una nueva manera de plantear el problema des del modelo eco-cognitivo de la abducción. En la cuarta sección mostraré el peligro de cometer la falacia naturalista con este modelo y, finalmente, ofrezco mis reflexiones alrededor de éste, así como la necesidad de considerar el aspecto circunstancialmente ético de la abducción en los casos que se teorice sobre la mirada ética del mundo.

7.2. Problemas éticos en la MM

Las investigaciones que actualmente han intentado dar cuenta de lo expuesto en la introducción se han centrado en la manera como han de ser los algoritmos para poder ser éticos. Desde el consuensalismo propuesto por GENETH (Anderson y Anderson, 2015) a la formalización del lenguaje ético de SIROCCO (McLaren, 2003), la intención es superar el modelo que podemos representar con el proyecto de la MM del MIT, el cual intenta capturar las intuiciones éticas básicas de los participantes de un estudio basado en el experimento mental "*Trolley Problem*" de Philippa Foot (1967). Este enfoque ya ha proyectado una visión generalizada de lo que sería una ética computacional, la cual parte de una confusión entre valores y preferencias, la cual puede generar todo un seguido de injusticias cuando, por un lado, se considera que dichos resultados son éticos y, también, cuando nuestro conocimiento parte de los sesgos generados por los algoritmos que recopilan la información que nos sirve para hacer nuestras inferencias éticas (Sans y Casacuberta, 2019, 320-321).

Una definición estándar de "valor" sería *una propiedad que hace que un objeto o hecho sea mejor que otro a partir de un criterio no cuantificativo*, en la que "criterio cuantificativo" es la manera como decidimos algo en una escala definida ordinariamente. Un hecho importante de remarcar es que un criterio cuantificativo anula *de facto* el cualitativo por mor de la generalización. En el caso de la ética en relación con la computación, esta anulación se da en la transformación de una norma (moral) a una regla (algoritmo). Este hecho anula la parte intrínseca de los juicios éticos, los cuales implican el aspecto psicológico de los valores éticos, esto es, preferir algo no por su beneficio cuantitativo, sino por su *corrección*. Este elemento es el que permite buscar la repetición por una vía diferente a la generalización, esto es, buscándolo en cada una de nuestras acciones. Como hemos introducido, cuando la acción moral a la que nos referimos es hacía otra persona, se da una interacción, la cual contiene implícitamente la reciprocidad, en el sentido de que, si queremos establecer este tipo de relación entre seres humanos y dispositivos/máquinas, será necesario buscar la manera de que se de una comunicación

[155] Aunque existen debates sobre la forma, fisiognomía, voz, etc., que debería tener un dispositivo para que un ser humano confiase de tal manera que generase un vínculo emocional/moral con él.

bidireccional entre las normas y las reglas. Como hemos argumentado en otro sitio, en computación, esta traducción puede recibir el nombre de *regulación* (Sans y Casacuberta, 2019, 322). No obstante, en este paso, la regla pierde factores fundamentales para que se de una norma, a saber, que los juicios éticos no versan sobre lo que es, sino sobre lo que debe ser, esto es, una *posibilidad*. El motivo es porque no hay una función determinada para los juicios éticos, sino la simple interacción con los demás y, esta, se da en casos particulares, los cuales han de afrontar a cada momento problemas/situaciones, nuevas. En un sentido figurado, se puede entender que los valores éticos estructuran las acciones que hacemos, confeccionando así un sistema ético en el que se puede regular de manera no cuantificativa, incluso aquello que es cuantificable.

Por lo tanto, un primer paso sería preguntarnos cómo conseguir que un dispositivo *entre* en nuestra forma de vida, en el sentido en el que Wittgenstein mostró que es necesario saber cómo se usa el *Sprachspiel* de una comunidad para poder formar parte de ella (para compartir y estar en su forma de vida). Este punto de vista puede ayudarnos a entender que el problema ético al que nos enfrentamos no es el de si la percepción de una situación es ética, sino de que se llegue a reconocer a través de procesos que sean éticos para, así, ofrecer el tipo de soluciones que se exigen en cada caso (como, por ejemplo, el de la individualidad y vulnerabilidad de la persona que está sufriendo una injusticia).

En este sentido, la MM es insuficiente porque genera resultados a partir de cuantificaciones, las cuales se hacen a partir de una respuesta que no contempla los motivos por los cuales cada una de las personas se ha decantado por ella. Dicho de otro modo, no cuantifica valores, sino las preferencias de las personas en casos concretos. Estos motivos para preferir algo pueden estar construidos a partir de infinitud de sensaciones, emociones y otros aspectos psicológicos, los cuales muchas veces no son conscientes; elementos que actualmente son imposibles de capturar con precisión. Es extremadamente difícil hacer una lista de emociones que se puedan inferir, etc., principalmente porqué el ser humano no dispone de capacidad ni lenguaje pera matizar cada uno de los aspectos que lo llevan a decantarse (esto es, mostrar una preferencia en una escala cuantificadora) (Sans y Casacuberta, 2019, 326). Por lo tanto, parece que la investigación sobre máquinas éticas necesita buscar otro camino para poder dar resultados que den cuenta de las críticas arriba expuestas. Como veremos a continuación, un posible camino es el de buscar la *mimesis* a través del modelo eco-cognitivo de la abducción.

7.3. El razonamiento abductivo

En este apartado presento una aproximación teórica del mecanismo que podría servir para abordar el problema de entender y programar los aspectos morales. Esta propuesta parte de las teorías sobre el razonamiento abductivo para capturar la parte psicológica de los razonamientos, las cuales empezaron a diseñarse en el seno de la discusión de la dicotomía entre el contexto de justificación y el de des-

cubrimiento. Esta propuesta, si bien no soluciona el problema de la comprensión de la ética, sí que permite romper algunas fronteras que sirvieron en su momento para definir la lógica que nos ha servido hasta día de hoy para investigar en computación e AI[156] y, así, avanzar hacía la comprensión de la eticidad interna del ser humano a través de la intencionalidad hacía los objetos del mundo, la cual se pueden entender como *mediadores morales* (Magnani y Bardone, 2007). Cuando esta intencionalidad es percibida por otro agente humano, la comprensión ética se da *de facto*, siempre y cuando los sujetos formen parte de la misma forma de vida. No obstante, en caso contrario, se da la comprensión ética de la situación de todos modos, sin que eso implique una comprensión concreta de lo que se estaba proyectando con la acción.

Estos dos factores son, a mi entender, la clave para el propósito computacional, pues proponen un nuevo camino de comprensión de los algoritmos de aprendizaje automático a través de procesos abductivos dirigidos a reproducir una mimesis (Magnani, 2018). Esta propuesta empieza situándonos dentro de un *juego de imitación* (Turing, 1950), en que los dispositivos electrónicos y las máquinas son potencialmente capaces de modificar a los seres humanos y a la AI a través de la interacción. En este sentido, siguiendo la propuesta de Magnani y Bardeone, entiendo que los dispositivos tecnológicos y las máquinas pueden ser entendidas como entidades con una *moral pasiva* (Magnani y Bardone, 2007, 70), capaces de distribuir el aspecto moral de las acciones de los humanos (Magnani, 2018, 68). Esto es, que puedan gestionar un aspecto que es tácito en la interacción e implícito en el comportamiento (Magnani y Bardone, 2007, 100). Como puede verse, esta propuesta nos remite a la postulación de la dicotomía entre aspectos psicológicos y lógicos a la que me he referido más arriba[157], la cual es imposible tratar aquí (Niiniluoto, 2014, 378).

No obstante, aunque no explique con detalle esta problemática, me permite resaltar el valor de la abducción en este debate, pues precisamente ha cobrado importancia porqué, tal y como dice Thagard, puede ser "both a component in the discovery of hypotheses and a key ingredient in their justification" (Thagard, 1988, 52). Dicho de otro modo y remitiendo a la dicotomía de la que se desprende la arriba mencionada, se ha entendido a la aducción como el puente entre los hechos y los valores y esto es, en definitiva, la clave para disolver la problemática. Evidentemente, no estoy argumentando una explosión en cadena en que, si se anula la dicotomía de la que se desprenden las demás, entonces estás también se disolverán. En absoluto. Lo que intento introducir es que la sentencia que afirma que no se pueden hacer prescripciones de las descripciones o la que afirma que no se pueden declarar juicios verificables a partir de juicios psicológicos, sino solamente a partir de argumentos reductibles a la lógica, etc., parten de una definición concreta y compartida por todos de lo que significan cada uno de estos conceptos que contraponen, así como el papel que juegan dentro de nuestro entramado con-

[156] Me refiero a Frege y a la dicotomía entre psicología/lógica que empezó (Frege, 1918/1919, 58).
[157] V. Supra, 3-3n.

ceptual; y es ahora que vemos que solamente podemos avanzar derrumbando estos significados, pues la realidad se nos impone al no poder dar cuenta de ella.

Así, siendo el de la ética el escollo contra el que ha chocado el sistema clásico, al menos debemos proceder hipotetizando dos cosas, a saber, la primera, que esta redefinición conceptual debe permitir introducir los factores morales sin que eso sea tachable de error metodológico y, la segunda, que la misma moral debe ser redefinida para que encaje en todo este pulido de la arquitectónica conceptual con la que conocemos a la vez que diseñamos la realidad[158]; y un buen punto de partida sería entenderla distribuida, donde "distribución" tendrá una connotación en la que la *interacción* sea posible con todas las entidades con las que interactuamos de manera recíproca.

En este sentido, es fácil entender la abducción como un mecanismo para colapsar las dicotomías que he mencionado pues, como he dicho, aunque no declaren lo mismo en un sentido, sí que parten del mismo principio por otro, siendo posible hacerlas converger, como es el caso de la que separa la justificación del descubrimiento y la de hecho y valor (Putnam, 2002). Este colapso viene de la mano de que, evidentemente, una dicotomía es una *metanorma* que se impone a la posibilidad de influencia entre dos conceptos y, por lo tanto, es asumible que nuestra caja de herramientas de razonamiento es, por un lado, tanto descriptiva como basada en valores y, además, capaz de sacar conclusiones a partir de la mezcla que después se ha prohibido *ad hoc*, *por mor* de un método determinado que opera —bien— en un campo concreto (Feyerabend, 1993)[159].

Por lo tanto, partiendo de esta idea de que nuestro aparato de razonamiento opera conjuntamente de manera descriptiva y valorativa, la división entre capacidades lógicas y psicológicas se difumina y permite entender a las primeras como una forma más o menos fiable de representar a las segundas (Sans y Casacuberta, 2019, 327). Y digo más o menos fiable por el hecho de que, dependiendo de la finalidad, el nivel de simplificación de los procesos del concepto, puede variar tanto que, al final, la abducción definida en un área de ciencia cognitiva sea diferente a la de otro de sus campos; siendo el motivo el hecho de que la representación de la característica esencial de la abducción, esto es, plantear posibilidades, sea redefinida para que se pueda dar en otro campo. No obstante, este paso, de momento, ha implicado que el concepto redefinido no sea ya una abducción, sino un proceso que se le puede llamar como tal, en contraste con los otros. Un ejemplo lo encontramos en Kakas (2017), quien muy atinadamente llama a su abducción como "inducción a la inversa". Hay que tener en cuenta que "abducción" es una de las diversas maneras de cómo se ha traducido la $\grave{\alpha}\pi\alpha\gamma\omega\gamma\acute{\eta}$ aristotélica y que, dependiendo del contendido teórico y la carga que se pone de éste al traducirla, se implica aquello que queremos de dicho concepto. Así, con la definición de Kakas no se consigue lo que otros queremos capturar con la "abducción" que acuñó Peirce, inspirándose en Aristóteles.

[158] Como dijo Hintikka (2007,11): "we are both producers and consumers of knowledge".

[159] Es importante tener presente que él escribió contra un método, para que otros pudieran (co)existir.

Los intentos para conseguir esto son las aproximaciones teóricas que, de momento, han servido sobre todo para mostrar los límites de las teorías actuales en computación y AI, así como los de sus aplicaciones. Un ejemplo de esto es el de Thagard, uno de los pioneros en la investigación sobre la abducción, el cual mostró la importancia de aceptar el reto de conseguir implementar el aspecto psicológico en le aparato descriptivo de los algoritmos. En el planteamiento de su programa IP, nos muestra que nosotros podemos hacer un tipo de inferencias, las cuales se nos presentan imposibles de computar hoy en día. Esto es lo que nos introduce Thagard con el ejemplo de "suppose you are wondering why a young man, Michael, is dressed outrageously, so that you set yourself the problem of explaining (*dresses-outrageously (michael) true*)", el cual *puede* desencadenar la regla "if x is a rock musician, then x dresses outrageously" (Thagard, 1988, 54. Itálicas mías). Primero de todo, la regla se desencadenará a partir de factores que no controlamos y, seguido de esto, inferiremos de ellos una cosa u otra. Dicho de otro modo, ante un hecho sorprendente, la viabilidad de la opción generada recae más en su capacidad de plantear una línea de acción, que en la veracidad de ésta.

La *sorpresa* es el disparador psicológico que se da, a veces, circunstancial y contingentemente. Otro ejemplo es el descubrimiento del planeta Neptuno, hecho en que se mezclan diferentes maneras de comparar y sopesar la información disponible, ya que ninguno de estos factores era realmente determinante para la conclusión final de la inferencia que, precisamente por todo lo que he dicho, se puede identificar como abductiva (Sans, 2017). Estos dos casos muestran las dificultades de conseguir computar el proceso abductivo. Por un lado, nuestro profundo desconocimiento de cómo funciona nuestro aparato cognitivo y, más concretamente, por las extrañas características epistémicas de la abducción. Me refiero al hecho de que su papel en la ampliación de nuestro conocimiento parece ser el de permanecer en la ignorancia.

Esta aparente paradoja fue, a mi entender, la que configuró la primera etapa de recuperación de la abducción des de que Peirce la planteó como la lógica que subyace en el pragmatismo. El debate generado entre los esquemas GW (Gabbay y Wood, 2005) y el AKM (Aliseda, 2006)[160] ayudaron para definir los temas implícitos en este razonamiento, esto es, que se ocupa de la generación de conocimiento (*fill-up problem*) y la selección de una línea de acción de entre todas las opciones generadas (*cutdown problem*). No obstante, éste debate ha generado un subproblema que ha mostrado las carencias de nuestros sistemas representacionales a la hora de capturar la virtud epistémica que he mencionado al parágrafo de arriba, a saber, que el papel de la abducción en nuestros procedimientos gnoseológicos se basa en que permanece en la ignorancia, dónde "ignorancia" significa aquí el desconocimiento concreto de algo que intentamos resolver.

El problema al que me refiero es que, si se quiere dar peso a la permanencia del

[160]Esta teoría ha sido diseñada y defendida por bastantes investigadores. Usaré la obra de Aliseda como ejemplo, por dos razones, a saber, la primera, es una de las mejores reconstrucciones de esta aproximación a la abducción y, la segunda, que puede que sea de las pocas concepciones de este razonamiento que se puede usar localmente para dar cuenta de algunos de estos hechos sorprendentes.

estado de ignorancia del proceso abductivo (GW), entonces la abducción pierde la razón de ser en la ampliación de nuestro conocimiento, pues siempre hará falta una activación de la hipótesis conjeturada, hecho que ya no será abductivo: [teniendo en cuenta que $C(H)$ es la conjetura de la hipótesis de un agente y que Hc es la activación de dicha conjetura:]

10. Therefore, $C(H)$

11. Therefore, H^c (Woods, 2013, 370)

Estos dos procesos conclusivos son el resultado de intentar dar cuenta de un problema de ignorancia, el cual se desarrolla a través de un seguido de proposiciones que intentan capturar la condiciones y pasos necesarios para que se dé una abducción. La misma explicación que acabo de dar se puede aplicar al siguiente caso.

La otra alternativa es darle a la abducción un valor epistémico que permita que proceso y resultado impliquen una regla que ofrezca un tipo de conocimiento, esto es, una —mejor— explicación del hecho sorprendente. No obstante, ser una explicación mejor o peor de algo implica una comparación y evaluación, hechos que no se contienen en lo que la abducción quiere capturar, sino en una inducción del tipo que definió Harman (1965): [teniendo en cuenta que K es el conocimiento del agente, H es la hipótesis que ha de armonizar nuestro conocimiento con el hecho sorprendente, $\curvearrowright$ representa la inferencia no-monotónica de consecuencia presuntiva y que E es el hecho sorprendente al que queremos explicar:]

6. $K(H) \curvearrowright E$

7. Therefore, H (Gabbay y Wood, 2005, 48-49)

Una solución es la que ha ofrecido Magnani con su modelo eco-cognitivo, el cual ofrece una teoría de la abducción que se puede interpretar des de la perspectiva enactivista en que este razonamiento es la clave para dar soluciones tentativas, partiendo de un contexto en el que nosotros somos un elemento distribuido más, dentro de la totalidad de la operación cognitiva que se da al intentar dar cuenta del caso (Magnani, 2017)[161]. Esta aproximación es interesante porque realza elementos que no se habían tenido en cuenta porque se había teorizado desde los modelos lógicos que, a la postre, contienen una carga teórica e histórica sobre cómo se da el razonamiento humano. Des de aquí, se puede tener en cuenta el trabajo hecho por lógicos como Łukasiewicz que, bastante tiempo antes de que se reafirmara la dicotomía entre los hechos psicológicos y los lógicos, trabajó sobre los aspectos creativos que operaban a la par con los formales (Łukasiewicz, 1970) los cuales, en definitiva, son la clave para explicar el tipo de heurísticas que trabajan en la medida que conceptualizamos como, por ejemplo, generalizar o concretar ante la multiplicidad de hechos que se nos pueden aparecer.

[161] Este es uno de los trabajos más completos para entrar en esta teoría.

Esta teoría es altamente efectiva por dos razones, la primera, porque ha mostrado que el subproblema generado a causa del debate entre los esquemas GW y AKM es puramente contextual, en el sentido de que, dependiendo de cual sea el caso y nuestros intereses, una abducción puede dar un resultado de tipo explicativo en un lugar, mientras que las exigencias de otro contexto implicaran una abducción, el resultado de la cual no será una explicación. No obstante, el problema del que ha adolecido este modelo es que, en muchos sentidos, había dado la espalda al problema computacional original, *por mor* de dar cuenta del hecho cognitivo de los seres vivos. Actualmente, esta propuesta ha hecho un viraje muy interesante en que se retoma este tema de manera secundaría, a través de intentar gestionar el papel de la ignorancia que contiene la abducción (Bertolotti et al., 2016) y que se manifiesta en la manera como interactuamos con el entorno, los seres vivos y, como he introducido al primer apartado, con los dispositivos tecnológicos. Esta idea parte de los esquemas antiguos, en que se da por supuesto que nadie es ignorante totalmente y que, por lo tanto, siempre tenemos un punto de anclaje del que partir, sea voluntaria o involuntariamente. Des de esta posición, Magnani trata la ignorancia como un elemento epistémico relevante (Magnani, 2020).

El hecho de que la interacción con estos dispositivos tecnológicos sea cada vez más inevitable e incluso igual o más habitual que la que se da entre el resto de personas nos abre la pregunta sobre la *manera* como se ha de dar esta interacción. El proceso abductivo como elemento cognoscitivo de ampliación epistémica es un campo de investigación extremadamente importante, el cual puede dar por supuesto el aspecto ético, precisamente porque simplemente se da contextualmente. No obstante, en el momento en que tenemos en cuenta las herramientas que diseñamos para que, en este proceso epistémico, interactúen, más o menos de la manera como lo hacemos entre nosotros, entonces es más que necesario que abordemos el aspecto ético. Por un lado, esto es importante porque puede asentar las bases de un mundo tecnológico justo, pero, además, porque ayudará a denunciar aquellos sistemas éticos, los errores de los cuales quedan ocultos por el entramado cultural en el que vivimos pero que, cuando se desgranan y se presentan en sus formas sintéticas para ser más o menos traducidos en algoritmos, nos ponen sobre la mesa una realidad que no deseamos.

7.4. La falacia naturalista en AI

El problema ético se da de la mano de las propuestas más actuales en computación y AI, las cuales intentan dar cuenta de las carencias de la MM. El caso que quiero abordar en este escrito es el de los procesos miméticos, los cuales han de permitir una interacción mucho más similar a la que tenemos entre los seres humanos. No obstante, abrir estas puertas implica que este proceso de mimesis debe ser capaz de captar la manera como se generan y organizan nuestros razonamientos, inclusive los más lógicos. Esto nos obliga a redefinir la concepción de nuestros procesos cognitivos y, a la vez, nos permite poder debatir con sentido el tipo de

modelo ético que queremos para la interacción con la tecnología. Esta tarea es todavía una entelequia, pero es preciso comenzar a tratarlo para que, así, se pongan sobre la mesa los problemas que existirán realmente en un futuro.

A este tema le compete introducir la relación entre la ética y la abducción pues, si se obvia, se corre el riesgo de caer en un modelo que, al no estar interactuando a partir de lo que reconocemos como ético, corre el riesgo de, por el lado formal, ser falaz y, des del punto de vista de la vida práctica, ser un sistema injusto. Además, como he dicho al parágrafo anterior, estos aspectos valorativos también son parte componente del conocimiento explícito de tipo formal. Esta es una de las partes más importantes para tener en mente el pragmatismo, del cual la abducción es su piedra de toque, porque nos permite declarar una de sus máximas, a saber, que en la medida que conocemos, hacemos mundo. Un ejemplo que ya he dado es el de la generalización, otro más directo es el criterio de belleza, el cual nos muestra que la dicotomía entre hecho y valor es un supuesto más, destinado a solucionar toda una montaña de asunciones filosóficas que, a su vez, también están organizadas a partir de otras divisiones teóricas (Putnam, 2002).

Del mismo modo que el criterio de belleza, el ético ofrece el tipo de heurística que empleamos cuando conocemos y que nos ayuda a plantear los posibles mundos deseables, así como los que no. Lo que me gustaría plantear aquí es que este papel epistemológico, el cual difumina la dicotomía entre hecho y valor, puede servir para la cuestión ética a la que nos hemos de encarar en la medida en que los sistemas sean más afines a nosotros. Mi propuesta es tener en cuenta que la abducción contiene este elemento valorativo que se da en todo tipo de razonamiento y que nos abre las puertas a una forma diferente de entender nuestra aproximación al mundo que nos rodea, esto es, la pragmatista arriba mencionada.

La idea de relacionar la mimesis con la abducción que he introducido al principio del texto se desarrolla en la base del modelo eco-cognitivo de Magnani, actualizándolo y permitiendo que se adapte mejor a la realidad en la que vivimos. Como hemos visto, esta propuesta intenta nivelar la manera como los dispositivos tecnológicos con cierto grado de inteligencia adquieren "experiencia" a la manera como lo hacemos nosotros para, así, dar cuenta de que nuestro aparato cognoscitivo es altamente distributivo y se moldea en la medida que, interactuando, modificamos el entorno. La idea central de esta propuesta es que los dispositivos copien los elementos simples que conocemos de nuestros procesos cognitivos para ver cómo se adaptan a nuestra interacción cuando, por ejemplo, hacemos intervenir elementos no computables como nuestra memoria, sentimientos, etc. (Magnani, 2020).

Así, también es un paso más en la teoría epistemológica que subyace en toda teoría sobre la abducción. Por lo tanto, el problema de no tener en cuenta el factor ético de este razonamiento se inscribe dentro del problema del papel de los valores en la epistemología, pero también abre las puertas a considerar una teoría de corte pragmatista, en la que el problema del valor ético modifica de base la manera como trabajamos en epistemología.

No obstante, al no contener esta propuesta este análisis ético, nos encontramos

con el problema de que los valores que inevitablemente se involucran en la nivelación a través de la mimesis implican que, cuando se quiere usar un sistema para asuntos que convergen dentro de nuestra vida moral, entonces se comete la falacia naturalista.

Esta es la propuesta de Wam Kim, Donaldson y Hooker (2018), quienes detectan que, si se quiere hacer compaginar nuestros valores con los automatismos de una AI, debemos plantearnos cómo se dará el mecanismo de nivelación entre los dos factores. Esta investigación busca un sistema que pueda inferir las preferencias de un agente particular racional que se relaciona con el entorno, el cual se inspire en la manera cómo los seres humanos adquirimos los valores.

Esta propuesta es que la alineación hecha a partir de procesos miméticos nos hace caer en la falacia naturalista y ofrecen la alineación hecha a partir de valores anclados. La propuesta mimética se basa en procesos de imitación de los valores relevantes de nuestra actividad en relación a conseguir dicha alineación. El conjunto de lo valores contiene preferencias, análisis *big data* del comportamiento humano, expresiones lingüísticas, etc. Por otro lado, la propuesta alternativa parte de procesos que anclan el entorno de la máquina con valores normativos. Otra alternativa es la versión híbrida, la cual incorpora elementos miméticos, pero que no cae en la falacia naturalista, gracias a la propuesta de anclaje.

La falacia naturalista se da porque el peso de la investigación recae en la manera en cómo se debe dar la alineación, en vez de intentar dar cuenta de lo que es un valor. Este hecho implica que se da más peso a datos normativos obtenidos por análisis de preferencias, etc. Por lo tanto, se deriva una prescripción de una descripción. Como es sabido, la falacia naturalista es el tipo de error lógico que se da cuando se quiere sustituir "bueno" por alguna propiedad de un objeto natural, en el sentido de que se considera un elemento descriptivo; hecho que implicaría poder sustituir la Ética por algún sistema de la ciencias naturales o metafísico (Moore, 2002, 92).

Así, como hemos visto, el punto al que ahora llegamos se deriva de la confusión entre valores y preferencias del tercer apartado de este texto, la cual está asumida culturalmente, mucho antes de empezar a plantear la posibilidad del sistema del que se trata en la investigación que estoy presentando. El punto más interesante de esta propuesta parte de que se asume la totalidad de la crítica de Moore y entiende "valor" en el sentido intrínseco que él propuso, el cual, entre muchas cosas, trata de eliminar la asunción mereológica (en su momento, hegeliana) de que la totalidad de un sistema tiene más valor que la suma de sus partes (Moore, 2002, 78). Sin entrar en detalles de qué significa "valor intrínseco", una de las ideas a las que apunta Moore es que la bondad tiene un valor propio que puede hacer variar todo un conjunto descrito.

Es en este sentido que Kim, Donaldson y Hooker entienden "valor", a saber, que no se puede capturar con el conjunto de listas de preferencias. Por el contrario, mediante la abducción, proponen los valores éticos en el sentido prescriptivo de que son guías para nuestras acciones. El motivo de usar la abducción como herramienta básica para este propósito está precisamente en el factor ampliativo

no-clásico que este concepto tiene, pues permite entender que muchas de las aproximaciones básicas que hacemos los humanos con el entorno contienen ya una proyección, la cual se manifiesta en lo que escogemos, cómo nos comportamos con los objetos que nos rodean, etc. Para su propósito, uno de los factores lógicos más interesantes es que la relevancia de la abducción en este proceso no guarda relación con las premisas y conclusiones que se puedan conseguir, sino que solamente anclan patrones de acción.

7.5. Conclusiones

En este artículo he querido mostrar la relevancia de la investigación sobre el razonamiento abductivo para dar cuenta de los problemas que están surgiendo en el ámbito de la computación y la AI. A mi entender, no se pueden desligar estas problemáticas de las puramente epistemológicas y cognitivas pues, por un lado, son la base de nuestros modelos de representación que, después, usamos para nuestras máquinas y, además, porque estos dispositivos que creamos son el reflejo de la manera como nos entendemos. Por lo tanto, corremos el riesgo de cometer errores que van más allá de una investigación aislada.

De hecho, en la medida que la convivencia con dispositivos electrónicos con un grado de inteligencia determinado es ya una realidad, es menester investigar para dar cuenta del problema ético que se nos presenta, el cual, como he mostrado, está estrechamente relacionado con la manera como nos relacionamos con el mundo. El razonamiento abductivo es una de las claves de este problema filosófico, el cual es actualmente una de las piedras de toque de diversas teorías. Mi propuesta de implicar los elementos valorativos de tipo prescriptivo en este razonamiento se presenta interesante porque ayuda a armonizar los problemas que he presentado de las propuestas más frutíferas sobre este razonamiento, las cuales permiten esta incorporación sin perder el valor del trabajo que ya han conseguido.

7.6. Agradecimientos

Quiero agradecer a los miembros del grupo de investigación TecnoCog i al GEHUCT (Grupo de Estudios Humanísticos sobre la Ciencia y la Tecnología) la ayuda que me han dado durante todos estos años de investigación y camaradería. Además, este artículo ha sido posible gracias al proyecto de investigación "Research group Epistemic Innovation: the case of the biomedical sciences" (FFI2017-85711-P) y al programa de contractos predoctorales en formación de profesorado universitario (FPU).

Bibliografía

Aliseda, A. (2006). *Abductive reasoning: logical investigations into discovery and explanation*. Springer, The Netherlands.

Anderson, M. y Anderson, S. L. (2015). Toward ensuring ethical behavior from autonomous systems: a case-supported principle based paradigm. In Walsh, T., editor, *Artificial Intelligence and Ethics: Papers from the 2015 AAAI Workshop*. The AAAI Press.

Bertolotti, T., Arfini, S., y Magnani, L. (2016). Abduction: From the ignorance problem to the ignorance virtue. *FLAP*, (3):153–173.

Feyerabend, P. (1993). *Against Method*. Verso, London & New York.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Rev.*, (5):5–15.

Frege, G. (1918/1919). Der gedanke. eine logische untersuchung. *Beiträge zur Philosophie des deutschen Idealismus*, (2):58–77.

Gabbay, M. y Wood, J. (2005). *A practical logic of cognitive systems: The reach of abduction: Insight and trial*. Elsevier, Amsterdam.

Harman, G. (1965). The inference to the best explanation. *Philosophical Review*, 74(1):88–95.

Hintikka, J. (2007). *Socratic epistemology*. Cambridge University Press, Cambridge.

Kakas, A. C. (2017). Abduction. In Sammut, C.and Webb, G. I., editor, *Encyclopedia of machine learning and data mining*. Springer, New York.

Kant, I. (1960). *Grundlegung zur Metaphysik der Sitten*. Felix Meiner, Hamburg.

Łukasiewicz, J. (1970). *Creative elements in science*. North-Holland Publishing Company, Amsterdam.

Magnani, L. (2017). *The Abductive Structure of Scientific Creativity. An Essay on the Ecology of Cognition*, volume 37 of *Studies in Applied Philosophy, Epistemology and Rational Ethics*. Springer, Cham.

Magnani, L. (2018). Eco-cognitive computationalism: from mimetic minds to morphologybased enchancement of mimetic bodies. *Etropy*, (20):430–446.

Magnani, L. (2020). Computational domestication of ignorant entities. *Synthese*, pages 1–30.

Magnani, L. y Bardone, E. (2007). Distributed morality: exteralizing ethical knowledge in technological artifacts. *Sci*, (13):99–108.

McLaren, B. M. (2003). Extensionally defining principles and cases in ethics: an ai model. *Artif. Intell. J.*, (150):145–181.

Moore, G. E. (2002). *Principia Ethica*. University of Cambridge Press, Cambridge.

Niiniluoto, I. (2014). Representation and truthlikeness. *Sci*, 19(4):375–379.

Putnam, H. (2002). *The collapse of fact/value dichotomy and other essays*. Harvard University Press, Cambridge.

Sans, A. (2017). El lado epistemológico de las abducciones: La creatividad en las verdades proyectadas. *Revista iberoamericana de argumentación*, (15):77–91.

Sans, A. y Casacuberta, D. (2019). *Remarks on the Possibility of Ethical Reasoning in an Artificial Intelligence System by Means of Abductive Models*, volume 49 of *Studies in Applied Philosophy, Epistemology and Rational Ethics*, pages 318–333. Springer, Cham.

Thagard, P. (1988). *Computational philosophy of science*. MIT Press, Massachusetts.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, (49):433–460.

Wan Kim, T., Donaldson, T., y Hooker, J. (2018). Mimetic vs. anchored value alignment in artificial intelligence. *arXiv*.

Woods, J. (2013). *Errors of Reasoning. Naturalizing the Logic of Inference*. College Publications, London.