El sesgo algorítmico y el enfoque del diseño sensible al valor

Revista Latinoamericana de Economía y Sociedad Digital

Issue Número Especial 1, Julio 2022

Autores: Judith Simon, Pak-Hang Wong, Gernot Rieder

DOI: 10.53857/TZVN9229

Publicado: 29 julio, 2022

Cita sugerida: Simon, J., Wong, P.-H., & Rieder, G. (2022). El sesgo algorítmico y el enfoque del diseño sensible al valor. Revista Latinoamericana de Economía Y Sociedad Digital. https://doi.org/10.53857/tzvn9229

Tipo: Ensayo, Traducción

Palabras clave: Diseño Sensible al Valor, Sesgo Algorítmico, Valores Humanos

---

Nota sobre la traducción

Este artículo fue originalmente publicado en la sección Concepts of the Digital Society de la revista Internet Policy Review, editada por Christian Katzenbach y Thomas Christian Bächle. Es publicado en la Revista Latinoamericana de Economia y Sociedad Digital según los lineamientos de acceso abierto del artículo original, con licencia Creative Commons Attribution 3.0 Germany.
Título Original: Algorithmic Bias and the Value Sensitive Design Approach
Fecha de publicación original: 18 de Diciembre de 2020
Link al artículo original: https://policyreview.info/concepts/algorithmic-bias
DOI original: https://doi.org/10.14763/2020.4.1534
Traducido del idioma inglés por The Pillow Books para la Revista Latinoamericana de Economía y Sociedad Digital en Mayo 2022.

Resumen

En los últimos tiempos, ante el aumento de la conciencia de que los algoritmos informáticos no son herramientas neutrales, sino que al reproducir y amplificar los sesgos pueden causar daño, los intentos para detectar y prevenir dichos sesgos se han intensificado. Un enfoque que ha recibido gran atención en este sentido es la metodología del diseño sensible al valor (VSD, por sus siglas en inglés), el cual busca contribuir tanto al análisis crítico de los (des)valores en las tecnologías existentes como a la construcción de nuevas tecnologías que tomen en cuenta valores deseados específicos. Este artículo ofrece un breve panorama de las principales características del enfoque del diseño sensible al valor, analiza sus contribuciones a la comprensión y el abordaje de problemas relacionados con el sesgo en los sistemas informáticos, presenta un esbozo de los debates actuales en torno al sesgo algorítmico y la equidad en el aprendizaje automático y analiza de qué manera estos debates podrían beneficiarse de las ideas y recomendaciones derivadas del VSD. Al relacionar estos debates sobre los valores en el diseño y el sesgo algorítmico con la investigación sobre los sesgos cognitivos, concluimos destacando nuestro deber colectivo no solo de detectar y contrarrestar los sesgos en los sistemas de software, sino también de abordar y remediar sus orígenes sociales.

Abstract

Recently, amid growing awareness that computer algorithms are not neutral tools but can cause harm by reproducing and amplifying bias, attempts to detect and prevent such biases have intensified. An approach that has received considerable attention in this regard is the Value Sensitive Design (VSD) methodology, which aims to contribute to both the critical analysis of (dis)values in existing technologies and the construction of novel technologies that account for specific desired values. This article provides a brief overview of the key features of the Value Sensitive Design approach, examines its contributions to understanding and addressing issues around bias in computer systems, outlines the current debates on algorithmic bias and fairness in machine learning, and discusses how such debates could profit from VSD-derived insights and recommendations. Relating these debates on values in design and algorithmic bias to research on cognitive biases, we conclude by stressing our collective duty to not only detect and counter biases in software systems, but to also address and remedy their societal origins.

1. INTRODUCCIÓN

Cuando, en 2016, periodistas de investigación de ProPublica publicaron un informe en el que declaraban que un sistema informático utilizado en los tribunales de Estados Unidos presentaba sesgos desde el punto de vista racial, sobrevino un intenso debate. En esencia,

la investigación había descubierto que COMPAS, una herramienta de apoyo en la toma de decisiones utilizada por jueces y supervisores de libertad condicional para evaluar la probabilidad de reincidencia de una persona acusada, sobrestimaba de forma sistemática el riesgo de reincidencia de la población afrodescendiente y subestimaba el de la población blanca (véase Angwin *et al*., 2016). Northpointe, la empresa que desarrolló COMPAS, refutó las acusaciones bajo el argumento de que su herramienta de evaluación era justa porque predecía la reincidencia con aproximadamente la misma precisión, sin importar el origen étnico de las personas acusadas (véase Dietrich *et al*., 2016). Los periodistas de ProPublica, a su vez, sostuvieron que un modelo algorítmico no puede ser justo si produce errores graves, es decir, falsos positivos (falsas alarmas) y falsos negativos (detecciones fallidas), con mayor frecuencia para un grupo étnico que para otro, lo cual desencadenó un debate sobre la idea misma de programar la equidad en un algoritmo informático (véase, por ejemplo, Wong, 2019). Hasta la fecha, más de mil artículos académicos han citado el artículo de ProPublica[1] y sus conclusiones se han debatido en medios de comunicación de todo el mundo.

Pero el caso de ProPublica no fue aislado. Por el contrario, marcó el inicio de una serie de informes y estudios en los que se encontró evidencia de sesgos algorítmicos en diversas áreas de aplicación: desde los sistemas de contratación (Dastin, 2018) hasta el puntaje crediticio (O'Neil, 2016) y el *software* de reconocimiento facial (Boulamwini y Gebru, 2018). Casos como estos, que ponen de manifiesto el potencial de la discriminación automatizada con base en características como la edad, el género, la etnia o el nivel socioeconómico, han reavivado viejos debates sobre la relación entre la tecnología y la sociedad (véase, por ejemplo, Winner, 1980); así, se cuestiona la neutralidad de los algoritmos y se invita a debatir sobre su capacidad de estructurar y dar forma a la sociedad, y no solo de reflejarla. Sin embargo, si las tecnologías no son neutrales desde el punto de vista moral y si los valores y disvalores incorporados tienen consecuencias tangibles tanto para las personas como para la sociedad en su conjunto, ¿no implicaría esto que los algoritmos deberían diseñarse con cuidado y que se debería tratar no solo de detectar y analizar los problemas, sino de abordarlos de forma proactiva a través de decisiones de diseño conscientes?.[2] Estas cuestiones, que ahora se debaten en la comunidad informática, no son nuevas; su historia dentro de la propia informática es larga y a menudo desdeñada □por ejemplo, a través de la investigación en diseño participativo□, pero también en otros campos y disciplinas como la ética informática, la filosofía de la tecnología, la historia de la ciencia o los estudios de la ciencia y la tecnología (STS, por sus siglas en inglés). No obstante, el intento más riguroso de diseñar de forma responsable y con atención a los valores humanos es el enfoque del diseño sensible al valor (VSD, por sus siglas en inglés), que surgió de este panorama intelectual a mediados de la década de 1990 y se ha ampliado y perfeccionado desde entonces. En fechas más recientes, y como resultado de una mayor conciencia de que "los datos no son una panacea" y de que las técnicas algorítmicas pueden "afectar la suerte de grupos enteros de personas de forma sistemáticamente desfavorable" (Barocas y Selbst, 2016, p.673), el interés por la metodología del VSD ha ido en aumento, lo que plantea la

siguiente interrogante: ¿qué aportaciones puede hacer este enfoque a los actuales debates sobre el sesgo y la equidad en la toma de decisiones algorítmicas y el aprendizaje automático?

Este artículo ofrece un breve panorama de las principales características del diseño sensible al valor (sección 2), analiza sus contribuciones a la comprensión y el abordaje de problemas relacionados con el sesgo en los sistemas informáticos (sección 3), presenta un esbozo de los debates actuales sobre el sesgo algorítmico y la equidad en el aprendizaje automático (sección 4) y analiza de qué manera estos debates podrían beneficiarse de las ideas y recomendaciones derivadas del VSD (sección 5). Al relacionar estos debates sobre los valores en el diseño y el sesgo algorítmico con la investigación sobre los sesgos cognitivos, concluimos subrayando nuestro deber colectivo no solo de detectar y contrarrestar los sesgos en los sistemas de *software*, sino también de abordar y remediar sus orígenes sociales (sección 6).

2. EL DISEÑO SENSIBLE AL VALOR: UN BREVE PANORAMA

El diseño sensible al valor, en tanto metodología con base teórica, surgió en el contexto de la acelerada informatización de la década de 1990 y como respuesta a la necesidad percibida de un enfoque de diseño que tomara en cuenta los valores humanos y el contexto social a lo largo del proceso de diseño (véase Friedman y Hendry, 2019). De hecho, el influyente libro editado por Friedman (1997) *Human Values and the Design of Computer Technology* (Los valores humanos y el diseño de la tecnología informática) ya constituía una impresionante demostración de cómo conceptualizar y abordar las cuestiones relativas a la agencia, la privacidad y los prejuicios en los sistemas informáticos, haciendo hincapié en la necesidad de "adoptar el diseño sensible al valor como parte de la cultura de la informática" (*ibid*.: p.1). En esencia, el VSD ofrece una metodología concreta sobre cómo integrar de forma deliberada los valores deseados en las nuevas tecnologías. Consta de tres fases iterativas, a saber, investigaciones conceptuales-filosóficas, empíricas y técnicas (véase Friedman *et al*., 2006; Flanagan *et al*., 2008).[3]

Las investigaciones conceptuales-filosóficas abarcan tanto la identificación de los valores humanos relevantes como la identificación de los actores directos e indirectos. En cuanto a los primeros, los objetivos de las meticulosas conceptualizaciones de trabajo relativas a valores específicos son (a) aclarar las cuestiones fundamentales que plantea el proyecto en cuestión y (b) permitir las comparaciones entre estudios y equipos de investigación basados en el VSD. Mientras que el VSD define los *valores humanos* en términos generales como "lo que las personas consideran importante en su vida, con énfasis en la ética y la moral" (Friedman y Hendry, 2019, p. 24), Friedman *et al*. (2006, p. 364f) proporcionaron una lista heurística de valores humanos con relevancia ética[4] que a menudo están implicados en el diseño de sistemas. En relación con los actores, al considerar tanto a los directos como a los

indirectos, el VSD busca hacer contrapeso a la frecuente omisión de las personas no usuarias en el diseño de la tecnología, es decir, de los grupos que quizá no utilicen de forma directa la tecnología pero que, sin embargo, se ven afectados por ella (véase Oudshoorn y Pinch, 2005; Wyatt, 2005). Dado que los valores suelen estar interrelacionados considérese, por ejemplo, el debate actual sobre la relación entre privacidad y seguridad y que lo que es importante para un grupo de actores puede no para otro, las investigaciones conceptuales también se ocupan de la importancia relativa de los distintos valores, así como de las posibles compensaciones entre valores en conflicto.

Las *investigaciones empíricas* recurren a una amplia gama de métodos cuantitativos y cualitativos de las ciencias sociales (por ejemplo, encuestas, entrevistas, observaciones, experimentos) para brindar una mejor comprensión de la forma en que los actores conciben y priorizan los valores en contextos sociotécnicos específicos. Las afiliaciones culturales, históricas, nacionales, étnicas y religiosas pueden desempeñar un papel en este proceso y determinar cómo se manejan y resuelven los conflictos de valores (véase Flanagan *et al.,* 2008, p. 328). Además, las investigaciones empíricas pueden poner de manifiesto las diferencias entre la práctica propugnada (lo que se dice) y la práctica real (lo que la gente hace), lo cual permite un análisis más matizado de las decisiones en materia de diseño y de su impacto en el uso, y por tanto complementa las investigaciones conceptuales descritas más arriba. En última instancia, es a través de esta metodología empírica de investigación como se puede lograr una comprensión pormenorizada del sistema sociotécnico, lo cual facilita no solo la observación de los patrones de uso y apropiación de los actores, sino también el que los valores previstos en el proceso de diseño se cumplan, se modifiquen o se subviertan.

Las *investigaciones técnicas* se basan en la suposición de que cualquier diseño tecnológico ofrece "valores adecuados" (Friedman y Hendry, 2019, p. 34) en tanto que apoya ciertos valores y actividades antes que otros. De acuerdo con Friedman *etal.* (2008), las investigaciones sobre estas adecuaciones pueden adoptar una de las dos formas siguientes: en la primera, las investigaciones técnicas se centran en cómo las propiedades tecnológicas existentes pueden apoyar o dificultar valores humanos específicos. Aunque este enfoque es similar al empírico, en lugar de centrarse en las personas, los grupos y los sistemas sociales más amplios, el énfasis se pone en la propia tecnología. En la segunda forma, las investigaciones tecnológicas implican el diseño proactivo de sistemas para apoyar y realizar valores identificados en la investigación conceptual. Si, por ejemplo, la privacidad es un valor a preservar, es necesario implementar mecanismos técnicos que fomenten y promuevan la protección de la privacidad en lugar de disminuirla. Dado que determinados diseños darán prioridad a ciertos valores sobre otros, las investigaciones técnicas pueden revelar jerarquías de valores existentes (primera forma) o prospectivas (segunda forma), lo cual añade otra capa de comprensión al análisis.

A través de estas tres metodologías de investigación, el VSD pretende contribuir al análisis crítico de los sistemas sociotécnicos y de los valores que se han incorporado a ellos, ya sea

de forma intencionada o involuntaria. Por tanto, el VSD sirve como herramienta analítica para abrir los procesos de valoración en el diseño y el desarrollo de la tecnología, los cuales suelen ser ignorados o descuidados. Asimismo, proporciona una herramienta constructiva que permite y apoya la realización de valores específicos deseados en el diseño y desarrollo de nuevas tecnologías.[5]

3. EL SESGO EN LOS SISTEMAS INFORMÁTICOS

Mucho antes del debate actual sobre el sesgo algorítmico y sus consecuencias, Friedman y Nissenbaum (1996) fueron pioneros en el análisis del sesgo en los sistemas informáticos, bajo el argumento de que dichos sistemas presentan sesgos si "discriminan de forma sistemática e injusta a ciertas personas o grupos de personas en favor de otras [negando] la oportunidad de obtener un bien o [asignando] un resultado indeseable a una persona o grupo de personas por motivos irracionales o inapropiados" (*ibid*.: p. 332). Para Friedman y Nissenbaum era importante comprender mejor los sesgos en los sistemas informáticos, entre otras cosas porque consideraban que los sistemas sesgados eran "instrumentos de injusticia" y subrayaban que "la ausencia de sesgos debe formar parte del selecto conjunto de criterios según los cuales debe juzgarse la calidad de los sistemas vigentes en la sociedad" (*ibid*.: p. 345f). Una adecuada comprensión de los sesgos permitiría identificar los daños potenciales de un sistema y evitarlos en el proceso de diseño o corregirlos si el sistema ya está en uso. Con este fin, Friedman y Nissenbaum propusieron una taxonomía de sesgos, la cual es aún muy relevante y útil en el debate actual sobre el sesgo y la discriminación algorítmica (véase, por ejemplo, Dobbe *et al*., 2019; Cramer *et al*., 2018). Con base en el origen respectivo del sesgo, especificaron tres tipos diferentes de sesgos, a saber, el *sesgo preexistente,* el *sesgo técnico* y el *sesgo emergente.*

Según Friedman y Nissenbaum (1996), el *sesgo preexistente* tiene su origen en las instituciones, prácticas y actitudes sociales y suele anteceder a la creación del sistema. Puede provenir de personas que han contribuido de forma significativa al diseño del sistema (sesgo individual preexistente) o de prejuicios que existen en la sociedad o en la cultural en general (sesgo social preexistente). Es importante destacar que estos sesgos se introducen en el sistema de forma implícita e inconsciente, y no a través de un esfuerzo consciente.

El *sesgo técnico,* por su parte, surge de limitaciones o consideraciones técnicas. Las fuentes del sesgo técnico pueden incluir limitaciones de las herramientas informáticas (por ejemplo, en lo que a *hardware, software* o periféricos se refiere), el uso de algoritmos que se han desarrollado para un contexto diferente y la formalización injustificada de los constructos humanos, es decir, el intento de cuantificar lo cualitativo y discretizar lo continuo.

Finalmente, el *sesgo emergente* es aquel que surge en un contexto de uso, por lo general un tiempo después de la finalización de un diseño, como resultado de (a) nuevos conocimientos sociales o valores culturales cambiantes que no se incorporan o no pueden incorporarse al diseño del sistema o (b) un desajuste entre las personas usuarias su experiencia y sus

valores que se contemplaron para el diseño del sistema y la población que en realidad lo utiliza.

En suma, la taxonomía de sesgos propuesta por Friedman y Nissenbaum pretende que las personas que se dedican al diseño y la investigación puedan identificar y anticiparse a los sesgos en los sistemas informáticos al tener en cuenta las visiones del mundo, tanto individuales como sociales, las propiedades tecnológicas y los contextos de uso. Su análisis de los sesgos pone en primer plano la naturaleza impregnada de valores de los sistemas informáticos y subraya la posibilidad de mitigar o eliminar los posibles daños mediante la participación *proactiva* en el diseño y el desarrollo de los sistemas, que es uno de los principales objetivos del enfoque del diseño sensible al valor. En consecuencia, su análisis refleja la doble función del VSD como herramienta para el análisis de los (dis)valores en las tecnologías existentes y para la construcción de nuevas tecnologías que tengan en cuenta valores deseados específicos. Ambas funciones, la analítica y la constructiva, son también fundamentales en la investigación reciente sobre el sesgo y la equidad en el aprendizaje automático.

4. SESGO ALGORÍTMICO Y EQUIDAD EN EL APRENDIZAJE AUTOMÁTICO

Cuando la matemática Cathy O'Neil publicó su popular libro *Weapons of Math Destruction* (Armas de destrucción matemática) en 2016, el mensaje era claro: los modelos matemáticos, escribió, pueden "codificar los prejuicios, los malentendidos y los sesgos humanos en los sistemas de software que gestionan cada vez más nuestras vidas. […] Sus veredictos, incluso cuando son erróneos y nocivos, son inapelables. Y tienden a castigar a las personas pobres y oprimidas de nuestra sociedad, mientras enriquecen más a los ricos" (O'Neil, 2016, p. 3). Para respaldar esta afirmación, O'Neil analiza una serie de casos desde el *software* utilizado para predecir delitos y la publicidad en línea personalizada hasta los sistemas de clasificación de universidades y las herramientas de evaluación docente, pasando por los algoritmos de crédito, de seguros y de contratación y demuestra el poder punitivo que estos sistemas pueden tener sobre quienes ya sufren las desigualdades sociales. A la vez, destaca la tarea de "incorporar de forma explícita mejores valores en nuestros algoritmos para crear modelos de macrodatos (*big data*) que sigan nuestro ejemplo ético" (*ibid*.: p. 204). El libro de O'Neil, junto con algunos otros textos académicos y no académicos, fue la punta de lanza de un movimiento que pretendió oponerse a la representación de los algoritmos como justos y objetivos, al mostrar su potencial para "fomentar la injusticia, hasta que tomemos medidas para detenerlos" (*ibid*.: p. 203).

En la comunidad de las ciencias informáticas, donde la investigación sobre el sesgo y la discriminación en los procesos computacionales se llevó a cabo incluso antes del debate actual sobre los impactos del *big data* y la *inteligencia artificial* (véase, por ejemplo, Custers *et al*., 2013), se intensificaron los intentos para detectar y prevenir tales sesgos. La

organización de la reunión anual FAT/ML[6] a partir de 2014 es un ejemplo de lo anterior, que a la luz del reconocimiento cada vez mayor de que técnicas tales como el aprendizaje automático plantean "nuevos desafíos para garantizar la no discriminación, el debido proceso y la inteligibilidad en la toma de decisiones", buscó "proporcionar a los investigadores un lugar para explorar cómo caracterizar y abordar estas cuestiones con métodos rigurosos desde el punto de vista informático" (FAT/ML, 2018). Le siguieron otros eventos como el *Taller DAT (datos y transparencia algorítmica, por sus siglas en inglés)* en 2016 o el *Taller FATREC sobre Recomendación Responsable* en 2017 y la reunión *FAT/ML* fue sustituida por la *FAT** y más tarde por la Conferencia *ACM FAccT,* que busca reunir a "personas investigadoras y especialistas interesadas en la equidad, la rendición de cuentas y la transparencia en los sistemas sociotécnicos" (Conferencia ACM FaccT, 2021). Como ya se mencionó, esta investigación y el VSD comparten el doble objetivo de (a) identificar el sesgo y la discriminación en los sistemas algorítmicos (objetivo analítico) y (b) crear y diseñar sistemas algorítmicos justos (objetivo constructivo).[7]

Con respecto al inciso (a), las personas que investigan el sesgo algorítmico han propuesto diferentes marcos para comprender y localizar las fuentes de dichos sesgos, delineando así formas de mitigarlos o corregirlos (véase, por ejemplo, Baeza-Yates, 2018; Mehrabi *et al.,* 2019; Olteanu *et al*., 2019). Barocas y Selbst (2016), por ejemplo, proporcionan una descripción detallada de las diferentes formas en que los sesgos pueden introducirse en un sistema de aprendizaje automático; por ejemplo (i) a través de la especificación del problema, donde la definición de las variables objetivo se basa en elecciones subjetivas que pueden perjudicar de forma sistemática a ciertas poblaciones más que a otras; (ii) a través de los datos de capacitación, donde los conjuntos de datos sesgados pueden conducir a modelos discriminatorios y resultados nocivos; (iii) a través de la selección de características, donde la representación reduccionista de los fenómenos del mundo real puede dar lugar a determinaciones inexactas y efectos adversos; y (iv) a través de las variables indirectas, donde determinados puntos de datos están muy correlacionados con la pertenencia a una clase, lo que facilita el trato desigual y puede implicar resultados menos favorables para quienes forman parte de grupos desfavorecidos. En una línea similar, Danks y London (2017) identifican diferentes formas de sesgo algorítmico en los sistemas autónomos, a saber: (i) sesgo de datos de capacitación, (ii) sesgo de enfoque algorítmico, (iii) sesgo de procesamiento algorítmico, (iv) sesgo de contexto de transferencia y (v) sesgo de interpretación. Los paralelismos entre estos enfoques recientes y el trabajo anterior de Friedman y Nissenbaum se hacen más evidentes cuando se toma en cuenta su objetivo común de hacer un "llamado a la cautela" (Barocas y Selbst, 2016, p. 732), proporcionar "una taxonomía de los diferentes tipos y fuentes de sesgo algorítmico" (Danks y London, 2017, p. 4691) y ofrecer un "marco para entenderlo y remediarlo" (Friedman y Nissenbaum, 1996, p. 330). En cualquiera de los casos, la designación y caracterización de los diferentes tipos de sesgos se considera, pues, un elemento clave del objetivo analítico común de reconocer y remediar dichos sesgos en los sistemas algorítmicos existentes.

Por lo que respecta al inciso (b), más allá de la tarea analítica de identificar y mitigar el sesgo, en la comunidad de aprendizaje automático existe también la aspiración de índole más constructiva de diseñar *algoritmos justos*. Kearns y Roth, por ejemplo, describen esta aspiración como "la ciencia del diseño de algoritmos con conciencia social" que busca la manera de que los algoritmos logren "incorporar de forma cuantitativa, medible y verificable muchos de los valores éticos que nos resultan relevantes como individuos y como sociedad" (2019, p. 18). Por otra parte, la investigación sobre la equidad algorítmica se ha caracterizado por "traducir matemáticamente las regulaciones de no discriminación en restricciones de no discriminación y desarrollar algoritmos de modelado predictivo que sean capaces de tener en cuenta esas restricciones y, al mismo tiempo, ser lo más precisos posible" (Žliobaitė, 2017, p. 1061). En otras palabras, la investigación sobre la equidad algorítmica no solo tiene como objetivo identificar y mitigar el sesgo, sino construir de manera más proactiva el valor de la equidad en los sistemas algorítmicos. Este tipo de investigación suele basarse en ciertas métricas de equidad o restricciones de equidad predefinidas y, a partir de ahí, su objetivo es desarrollar sistemas algorítmicos optimizados con base en las métricas propuestas o satisfacer las restricciones especificadas. Este proceso puede tener lugar (i) en la fase de preprocesamiento, en la que se modifican los datos de entrada para garantizar que los resultados de los cálculos algorítmicos cuando se apliquen a los nuevos datos sean justos, (ii) durante la fase de procesamiento, en la que se modifican o sustituyen los algoritmos para generar resultados más justos, o (iii) en la fase de posprocesamiento, en la que se modifica el resultado de cualquier modelo para que sea más justo[8]. Una vez más, los paralelismos entre estos enfoques informáticos y el objetivo del VSD de "influir en el diseño de la tecnología desde el principio y a lo largo del proceso de diseño" son evidentes (Friedman y Hendry, 2019, p. 4). En ambos casos, la adopción de una orientación proactiva es señal de un compromiso compartido con el progreso y la mejora a través de un diseño ético y basado en valores. Se trata de una agenda constructiva que pretende contribuir a la innovación responsable en lugar de adoptar un enfoque puramente analítico y *a posteriori*. En palabras de Friedman y Hendry (2019, p. 2), "si bien el estudio empírico y la crítica de los sistemas existentes son esenciales, [el VSD] se distingue por su postura ante el diseño: imaginar, diseñar e implementar la tecnología de manera moral y ética para mejorar nuestro futuro".

5. ANÁLISIS

A pesar de las similitudes conceptuales descritas más arriba y del hecho de que la comunidad FAT (equidad, responsabilidad y transparencia en los sistemas sociotécnicos, por sus siglas en inglés) cite con frecuencia la literatura sobre VSD, la asimilación e integración de algunas de las ideas centrales de este tipo de diseño en la informática todavía es insuficiente en diversos aspectos importantes.

En primer lugar, se ha planteado la inquietud de que la literatura actual sobre la equidad en el aprendizaje automático tiende a centrarse demasiado en la forma en que los sesgos

individuales o sociales pueden penetrar en los sistemas algorítmicos, con particular énfasis en lo que Friedman y Nissenbaum denominan "sesgo preexistente", mientras que se ignoran otras fuentes de sesgo, como el sesgo técnico o el sesgo emergente. En respuesta a esto, Dobbe *et al.* (2019) han destacado la necesidad de adoptar una visión más amplia en relación con el sesgo algorítmico que considere todas las categorías de la taxonomía de Friedman y Nissenbaum (1996), así como "los riesgos más allá de los preexistentes en los datos" (Dobbe *et al.,* 2019, p. 2). Por lo tanto, para cumplir mejor el objetivo analítico de identificar y mitigar el sesgo en los sistemas algorítmicos, es importante que la comunidad académica del aprendizaje automático no recurra al VSD desde un enfoque ecléctico y fragmentario, sino que aproveche la totalidad de los marcos propuestos.

En segundo lugar, es importante recordar que conceptos como el de *equidad* no son en absoluto autoexplicativos ni claros. Verma y Rubin (2018), por ejemplo, señalan que en los últimos años se han propuesto más de veinte nociones diferentes de equidad en la investigación relacionada con la inteligencia artificial (IA), una falta de acuerdo que pone en tela de juicio la idea misma de operacionalizar la equidad cuando se pretende diseñar algoritmos justos. Aunque tanto la idea de equidad como el concepto relacionado de "igualdad de oportunidades" han sido ampliamente discutidos en la investigación filosófica (véase, por ejemplo, Ryan, 2006; Hooker, 2014; Arneson, 2018), Binns (2018) ha argumentado que la mayoría de las medidas en materia de equidad en la investigación del aprendizaje automático por lo general no se han teorizado lo suficiente desde una perspectiva filosófica, lo que resulta en enfoques que se centran "en un conjunto limitado y estático de clases prescritas y protegidas […] desprovistas de contexto" (*ibid*.: p. 9). Por último, pero no por ello menos importante, Corbett-Davies y Goel (2018) han destacado la divergencia entre las nociones formalizadas de equidad y lo que la gente suele entender por equidad en contextos cotidianos de toma de decisiones. Lo que se desprende de estas objeciones es que los intentos de formalizar y operacionalizar la equidad por vías específicas son cuestionables por numerosos motivos.

Por desgracia, a la hora de presentar soluciones técnicas suele ignorarse o restarse importancia a este carácter polémico,[9] a pesar de que en los últimos años se observa una tendencia hacia enfoques más interdisciplinarios, sensibles a la necesidad de ampliar el alcance del análisis. Un uso adecuado del VSD podría contribuir a estos esfuerzos, ya que el método no solo requiere una investigación rigurosa de los valores en juego (véanse, en particular, las investigaciones filosóficas y técnicas sobre el método de VSD), sino que también exige la participación de equipos de investigación interdisciplinarios que incluyan, por ejemplo, a filósofos, científicos sociales o juristas. Por supuesto, estos enfoques interdisciplinarios pueden suponer un reto y requerir muchos recursos, pero el diseño ético exige, en última instancia, algo más que abordajes mecánicos basados en fórmulas para cumplir con los requisitos de FAT (véase Keyes *et al*., 2019). El esfuerzo por lograr diseños que sean verdaderamente *sensibles al valor* implica ser sensible a los múltiples significados de los valores en diferentes contextos sociales y culturales y requiere reconocer, relacionar y aplicar diferentes competencias disciplinarias.

Por último, y a propósito de lo anterior, no solo es necesario ampliar el alcance de las perspectivas disciplinarias, sino también el del propio objeto de investigación. Es decir, en lugar de limitarse a la equidad, la responsabilidad y la transparencia en el aprendizaje automático, la investigación sobre el sesgo algorítmico también debería considerar (a) el sistema sociotécnico más amplio en el que se inscriben las tecnologías y (b) las diferentes lógicas y órdenes que estas tecnologías algorítmicas producen y generan. En relación con el primer inciso, Gangadharan y Niklas (2019) han advertido que el enfoque tecnocéntrico en la incorporación de la equidad en los algoritmos, que se basa en la idea de que los ajustes técnicos serán suficientes para prevenir o evitar resultados discriminatorios, corre el riesgo de ignorar las condiciones sociales, políticas y económicas más amplias en las que se originan la injusticia y la desigualdad. Con respecto al segundo inciso, Hoffmann (2019, p. 910) nos recuerdan que el trabajo sobre el sesgo algorítmico no solo exige una atención sostenida a los fallos del sistema, sino también a "los tipos de mundos que se construyen tanto de forma explícita como implícita por y a través del diseño, el desarrollo y la implementación de sistemas intensivos de datos y mediados por algoritmos". Por tanto, sería necesario prestar más atención a los "órdenes institucionales, contextuales y sociales más amplios instanciados por los sistemas mediados por algoritmos y sus lógicas de reducción y optimización" (*ibid*.). La comunidad FAT ya ha dado pasos en este sentido, con la *Conferencia ACM FAT* 2020* que busca de manera explícita "mantener y continuar con la mejor de la alta calidad de la investigación en ciencias informáticas en este dominio, mientras que, de forma simultánea, se amplía el enfoque a la investigación en derecho y ciencias sociales y humanidades" (Conferencia ACM Facct, 2020). No obstante, creemos que una asimilación más amplia del VSD, que desde un inicio fue conceptualizado como un enfoque interdisciplinario, podría contribuir a este proceso.

6. OBSERVACIONES FINALES

Este artículo presentó una revisión concisa de la metodología del diseño sensible al valor y de la taxonomía de los sesgos propuesta por Friedman y Nissenbaum (1996). Se demostró que tanto el VSD como la taxonomía de los sesgos son aún muy pertinentes para la investigación actual en materia de sesgos y equidad en los sistemas sociotécnicos. Sin embargo, a pesar de su utilidad, el VSD se utiliza a menudo solo de forma parcial, e ideas cruciales por ejemplo, sobre los fundamentos conceptuales de los valores, la necesidad de considerar tanto a las personas usuarias como a las no usuarias de una tecnología[10] o la importancia de un enfoque interdisciplinar se pierden. En consecuencia, sería aconsejable intensificar los esfuerzos para revitalizar y profundizar la adopción del diseño sensible al valor en la investigación sobre equidad, responsabilidad y transparencia (FAT, por sus siglas en inglés) y otras relacionadas. Por suerte, se observa una tendencia a ampliar los debates y llevar la discusión más allá del ámbito técnico.

Es evidente que la revisión del VSD y de la investigación sobre el sesgo algorítmico contenida en este documento no abarca la totalidad del debate en curso. Además, es

importante señalar que la investigación sobre los sesgos va mucho más allá del ámbito del VSD y la informática. De hecho, la psicología y las ciencias cognitivas hace tiempo que estudian los *sesgos cognitivos* (Gigerenzer *et al*., 2012; Kahneman, 2011) y los *sesgos implícitos* (Holroyd *et al.,* 2017).[11] Aunque la noción de Friedman y Nissenbaum de sesgo preexistente considera, hasta cierto punto, los sesgos implícitos, la relación entre los sesgos cognitivos humanos y los sesgos en los sistemas informáticos requiere un análisis más profundo. Sobre todo en el contexto de la toma de decisiones automatizadas (ADM, por sus siglas en inglés), donde las decisiones humanas se complementan —o incluso se sustituyen— con las decisiones de las *máquinas*, los sesgos cognitivos humanos pueden tener ramificaciones interesantes para el diseño y el uso de los sistemas ADM.

En primer lugar, los sesgos cognitivos pueden estar relacionados de forma *causal* con una toma de decisiones automatizada sesgada. Las limitaciones y los sesgos cognitivos pueden, por ejemplo, contribuir a la formación de estereotipos sociales, prejuicios y preferencias injustificadas o a prácticas de toma de decisiones deficientes (por ejemplo, a través de la interpretación incorrecta de las probabilidades), que se introducen en los sistemas de ADM a través de los datos de capacitación. Con ello se ocultan y, al mismo tiempo, se reproducen y refuerzan dichos sesgos en máquinas en apariencia neutrales.

En segundo lugar, y de manera inversa, los sistemas ADM también pueden reducir y/o eliminar los sesgos cognitivos al dar cuenta de los defectos en el razonamiento humano y quizá, incluso, corregirlos (véase, por ejemplo, Savulescu y Maslen, 2015; Sunstein, 2018). En este sentido, si quienes diseñan e investigan los sistemas ADM pueden (a) identificar las fuentes de los sesgos cognitivos y (b) contrarrestarlos mediante elecciones metodológicas específicas a la hora de diseñar e implementar el sistema, dichos sistemas pueden concebirse como herramientas tanto para revelar los sesgos cognitivos en la toma de decisiones humanas como para reducir o incluso prevenir sus impactos negativos mediante una sofisticada interacción persona-máquina en la toma de decisiones.

Por último, el hecho de delegar sin justificación la toma de decisiones *humanas* en las máquinas puede ser un sesgo cognitivo en sí mismo, conocido como sesgo de automatización (Mosier *et al.,* 1996) o complacencia automatizada (Parasuraman y Manzei, 2010). El sesgo de automatización se caracteriza por la tendencia humana a confiar demasiado en las máquinas supuestamente neutrales y a seguir las "decisiones" erróneas (o cuestionables) de éstas sin buscar más información que corrobore, contradiga o incluso descarte dichas decisiones en otras fuentes existentes (Skitka *et al*., 1999). En este sentido, la complacencia automatizada describe la confianza de las personas que operan los sistemas en la fiabilidad de los mismos, lo que provoca que no presten suficiente atención a la supervisión del proceso y a la verificación de los resultados del sistema. Así pues, el reconocimiento de los peligros del sesgo de automatización y la complacencia automatizada —es decir, del exceso de confianza en la toma de decisiones automatizada— nos remite a las tempranas advertencias de Friedman y Nissenbaum en relación con los sesgos de los sistemas informáticos en apariencia precisos, neutrales y objetivos, y a su oportuna solicitud

de ponerlos en evidencia y contrarrestarlos de forma activa para mejorar el diseño y el discurso público documentado sobre los méritos y las limitaciones de dichas herramientas informáticas. Sin embargo, mejorar nuestras herramientas no es suficiente; tener en cuenta los valores y contrarrestar los prejuicios también requiere el reconocimiento y la reparación de las desigualdades e injusticias existentes en nuestras sociedades y la aceptación de que no todos los procesos de toma de decisiones deben ser gestionados a través de algoritmos.

REFERENCIAS

Conferencia ACM sobre Equidad, Responsabilidad y Transparencia (ACM FaccT) (2020). Conferencia *ACM FAT* 2020*. Conferencia ACM FAccT. https://facctconference.org/2020/

Conferencia ACM FAccT (2021). Conferencia *ACM FaccT 2021*. Conferencia ACM FaccT. https://facctconference.org/2021/index.html

Angwin, J., Larson, J., Mattu, S. y Kirchner, L. (2016). Machine Bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arneson, R. (2018). Four Conceptions of Equal Opportunity. *The Economic Journal*, *128*(612), 152–173. https://doi.org/10.1111/ecoj.12531

Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM*, *61*(6), 54–61. https://doi.org/10.1145/3209581

Barcoas, S. y Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*(3), 671–732. https://doi.org/10.15779/Z38BG31

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K. y Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 1–15. https://doi.org/10.1147/JRD.2019.2942287

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research*, *81*, 149–159. http://proceedings.mlr.press/v81/binns18a.html

Brey, P. (2010). Values in technology and disclosive computer ethics. En L. Floridi (ed.), *The Cambridge handbook of information and computer ethics* (pp. 41–58). Cambridge Univerity Press. https://doi.org/10.1017/CBO9780511845239.004

Buolamwini, J. y Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, *81*, 1–15. http://proceedings.mlr.press/v81/buolamwini18a.html

Corbett-Davies, S. y Goel, S. (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *ArXiv*. https://arxiv.org/abs/1808.00023

Cramer, H., Garcia-Gathright, J., Springer, A. y Reddy, S. (2018). Assessing and Addressing Algorithmic Bias in Practice. *Interactions*, *25*(6), 58–63. https://doi.org/10.1145/3278156

Custers, B., Calders, T., Schermer, B. y Zarsky, T. (eds.). (2013). *Discrimination and Privacy in the Information Society Data Mining and Profiling in Large Databases*. Springer. https://doi.org/10.1007/978-3-642-30487-3

Danks, D. y London, A. J. (2017). Algorithmic bias in autonomous systems. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 4691–4697. https://dl.acm.org/doi/10.5555/3171837.3171944

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Davis, J. y Nathan, L. P. (2014). Value sensitive design: Applications, adaptions, and critique. En J. Hoven, P. E. Vermaas e I. Poel (eds.), *Handbook of Ethics, Values, and Technological Design* (pp. 1–26). Springer. https://doi.org/10.1007/978-94-007-6970-0_3

Dieterich, W., Mendoza, C. y Brennan, T. (2016). *COMPAS Risk Scales: Demonstrating accuracy equity and predictive parity* [informe técnico]. Northpointe. https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html

Dobbe, R., Dean, S., Gilbert, T. y Kohli, N. (2018). *A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics*. Taller 2018 sobre Equidad, Responsabilidad y Transparencia en el Aprendizaje Automático. http://arxiv.org/abs/1807.00553

FAT/ML (2018). *Fairness, Accountability, and Transparency in Machine Learning (Equidad, Responsabilidad y Transparencia en el Aprendizaje Automático)*. https://www.fatml.org/

Flanagan, M., Howe, D. C. y Nissenbaum, H. (2008). Embodying Values in Technology: Theory and Practice. En J. Hoven y J. Weckert (eds.), *Information Technology and Moral Philosophy* (pp. 322–353). Cambridge University Press. https://doi.org/10.1017/CBO9780511498725.017

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P. y Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency – FAT**, *19*, 329–338. https://doi.org/10.1145/3287560.3287589

Friedman, B. (ed.). (1997). *Human Values and the Design of Computer Technology*. Cambridge University Press.

Friedman, B. y Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with*

Moral Imagination. MIT Press.

Friedman, B., Kahn, P. H. y Borning, A. (2006). Value Sensitive Design and Information Systems. En P. Zhang y D. Galetta (eds.), *Human-Computer Interaction in Management Information Systems: Foundations* (pp. 348–372). M.E. Sharpe.

Friedman, B. y Nissenbaum, N. (1996). Bias in Computer Systems. *ACM Transactions on Information Systems*, *14*(3), 330–347. https://doi.org/10.1145/230538.230561

Gangadharan, S. P. y Niklas, J. (2019). Decentering technology in discourse on discrimination. *Information, Communication & Society*, *22*(7), 882–899. https://doi.org/10.1080/1369118X.2019.1593484

Gigerenzer, G., Fiedler, K. y H, O. (2012). Rethinking Cognitive Biases as Environmental Consequences. En P. M. Todd y G. Gigerenzer (eds.), *Ecological Rationality. Intelligence in the World* (pp. 80–110). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195315448.001.0001

Hoffmann, A. L. (2019). Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, *22*(7), 900–915. https://doi.org/10.1080/1369118X.2019.1573912

Holroyd, J., Scaife, R. y Stafford, T. (2017). What is Implicit Bias? *Philosophy Compass*, *12*, *e12437*. https://doi.org/10.1111/phc3.12437

Hooker, B. (2014). Utilitarianism and fairness. En B. Eggleston y D. Miller (eds.), *Cambridge Companion to Utilitarianism* (pp. 280–302). Cambridge University Press. https://doi.org/10.1017/CCO9781139096737.015

Kahneman, D. (2011). *Think, Fast and Slow*. Allen Lane.

Kearns, M. y Roth, A. (2020). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Keyes, O., Hutson, J. y Durbin, M. (2019). A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–11. https://doi.org/10.1145/3290607.3310433

Kluttz, D. N., Kohli, N. y Mulligan, D. K. (2020). Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions. En K. Werbach (ed.), *After the Digital Tornado: Networks, Algorithms, Humanity* (pp. 137–152). Cambridge University Press. https://www.cambridge.org/core/books/after-the-digital-tornado/shaping-our-tools-contestability-as-a-means-to-promote-responsible-algorithmic-decision-making-in-the-professions/311281626ECA50F156A1DDAE7A02CECB

Lepri, B., Oliver, N., Letouzé, E., Pentland, A. y Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, *31*, 611–627. https://doi.org/10.1007/s13347-017-0279-x

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. y Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. *arXiv*. https://arxiv.org/abs/1908.09635.

Mosier, K. L. y Skitka, L. J. (1996). Human Decision Makers and Automated Decisions Aids: Made for Each Other? En R. Parasuraman y M. Mouloua (eds.), *Automation and Human Performance: Theory and Applications* (pp. 201–220). NJ. Lawrence Erlbaum Associates.

Olteanu, A., Castillo, C., Diaz, F. y Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, *2*. https://doi.org/10.3389/fdata.2019.00013

O'Neil, C. (2016). *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*. Crown Publishers.

Oudshoorn, N. y Pinch, T. (2005). How Users and Non-users Matter. En N. Oudshoorn y T. Pinch (eds.), *How Users Matter: The Co-Construction of Users and Technology* (pp. 1–28). MIT Press.

Parasuraman, R. y Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Ryan, A. (2006). Fairness and Philosophy. *Social Research*, *73*(2), 597–606. https://www.jstor.org/stable/40971838

Savulescu, J. y Maslen, H. (2015). Moral enhancement and artificial intelligence: Moral AI? En J. Romportl, E. Zackova y J. Kelemen (eds.), *Beyond Artificial Intelligence: The Disappearing Human-Machine Divide* (pp. 79–95). Springer. https://doi.org/10.1007/978-3-319-09668-1_6

Simon, J. (2017). Value-sensitive design and responsible research and innovation. En S. O. Hansson (ed.), *The ethics of technology methods and approaches* (pp. 219–235). Rowman & Littlefield.

Skitka, L. J., Mosier, K. L. y Burdick, M. (1999). Does Automation Bias Decision-Making? *International Journal of Human–Computer Studies*, *51*, 991–1006. https://doi.org/10.1006/ijhc.1999.0252

Sunstein, C. R. (2019). Algorithms, correcting biases. *Social Research*, *86*(2), 499–511. https://www.muse.jhu.edu/article/732187

Verma, S. y Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness*, 1–7. https://doi.org/10.1145/3194770.3194776

Winner, L. (1980). Do Artifacts Have Politics? *Daedalus*, *109*(1), 121 136. https://www.jstor.org/stable/20024652

Wong, P. H. (2019). Democratizing Algorithmic Fairness. *Philosophy & Technology*, *33*, 225–244. https://doi.org/10.1007/s13347-019-00355-w

Wong, P. H. y Simon, J. (2020). Thinking About 'Ethics' in the Ethics of AI. *IDEES*, *48*. https://revistaidees.cat/en/thinking-about-ethics-in-the-ethics-of-ai/

Wyatt, S. (2005). Non-Users Also Matter: The Construction of Users and Non-Users of the Internet. En N. Oudshoorn y T. Pinch (eds.), *How Users Matter: The Co-Construction of Users and Technology* (pp. 67–79). MIT Press.

Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, *31*(4), 1060–1089. https://doi.org/10.1007/s10618-017-0506-1

Acerca de los Autores

Judith Simon: es Profesora Titular de Ética en Tecnologías de la Información en la Universität Hamburg. Está interesada en las cuestiones éticas, epistemológicas y políticas que surgen en el contexto de las tecnologías digitales, en particular en lo que respecta a la Big Data y la Inteligencia Artificial. Judith Simon es miembro del Consejo de Ética Alemán, así como de varios otros comités de asesoramiento sobre políticas científicas y también ha sido miembro de la Comisión de Ética de Datos del Gobierno Federal Alemán (2018-2019). Su Routledge Handbook of Trust and Philosophy se publicó en junio de 2020.

Pak-Hang Wong: es Investigador Asociado en el Grupo de Investigación de Ética en TI en el Departamento de Informática de la Universität Hamburg, Alemania. Su investigación explora los problemas sociales, éticos y políticos de la Inteligencia Artificial, la robótica y otras tecnologías emergentes. Wong recibió su doctorado en Filosofía de la Universidad de Twente, Países Bajos, en 2012, y luego ocupó cargos académicos en Oxford y Hong Kong, antes de ocupar su puesto actual. Es coeditor de Well-Being in Contemporary Society (2015, Springer) y ha publicado en Philosophy & Technology, Zygon, Science and Engineering Ethics, y otras revistas académicas.

Gernot Rieder: es investigador asociado posdoctoral en el Departamento de Informática de la Universität Hamburg y miembro del grupo de investigación Ética en Tecnologías de la Información (EIT). En su disertación (Universidad de TI de Copenhague), investigó el surgimiento de Big Data en las políticas públicas y las implicaciones sociales y éticas de la toma de decisiones algorítmica. Su investigación actual se centra en la economía política de la IA del bienestar, las raíces de la confianza en los datos y la TI, y la historia de la informatización.

Notas

↑1 Véase el recuento de citas del artículo en Google Académico en https://scholar.google.com/scholar?cites=9718961392046448783&as_sdt=2005&sciodt=0,5&hl=en.

↑2 En este trabajo utilizamos el término *valor* para referirnos a "las cosas que la gente considera valiosas y que son a la vez ideales y generales" y el término *disvalor* para referirnos a "aquellas cualidades generales que se consideran malas o perversas" (Brey, 2010, p. 46).

↑3 Los siguientes párrafos son una versión corregida y aumentada de la primera sección de "Value-Sensitive Design as a Methodology" (el diseño sensible al valor como metodología) (Simon, 2017).

↑4 Algunos ejemplos de estos "valores con importancia ética" son la *privacidad*, que significa "el derecho de una persona a determinar la información sobre sí misma que se puede comunicar a los demás"; la *autonomía*, que significa "la capacidad de las personas para decidir, planificar y actuar de la forma que creen que les ayudará a alcanzar sus objetivos"; o el *consentimiento informado,* que se refiere a "obtener el acuerdo de las personas, que abarca criterios de divulgación y comprensión (en lo que respecta a 'informado') y voluntariedad, competencia y acuerdo (en lo que respecta a 'consentimiento')" (Friedman *et al.,* 2006, p. 364).

↑5 Por supuesto, como cualquier metodología madura, el enfoque del diseño sensible al valor también ha sido objeto de numerosas críticas (véase Friedman y Hendry, 2019, p. 172f). Para una revisión exhaustiva de estas críticas, véase Davis y Nathan (2014).

↑6 FAT/ML son las siglas en inglés de equidad, responsabilidad y transparencia en el aprendizaje automático (Fairness, Accountability and Transparency in Machine Learning).

↑7 Desde el punto de vista del VSD, el desarrollo de un sistema algorítmico "justo" implicaría la incorporación de valores específicos como la equidad, la responsabilidad o la transparencia en el sistema.

↑8 Véase Lepri *et al*. (2018), Bellamy *et al*. (2019) y Friedler *et al*. (2019) para un panorama de estas técnicas.

↑9 Para un análisis detallado del concepto de contestabilidad y de la importancia del diseño contestable, véase Kluttz *et al.,* 2020.

↑10 Para un análisis más detallado de la necesidad de también tomar en cuenta a las personas no usuarias, véase Wong (2019) y Wong y Simon (2020).

↑11 Hay que tener en cuenta que el sesgo cognitivo y el sesgo implícito no necesariamente tienen una connotación moral negativa como en el caso del sesgo en el VSD.