TRADUCCIÓN

Lassi
p-ISSN 0252-8681 | año 2022 | núm. 44 | pp. 145-177

La ética de los algoritmos: un mapeo del debate [1]

Recibido: 04/11/2022 Aceptado: 12/12/2022

Agustina Lassi (traductora)
Universidad Nacional de La Matanza, Universidad Nacional Arturo Jauretche (Argentina)
ORCID: https://orcid.org/0000-0003-3171-6258
DOI: https://doi.org/10.29166/csociales.v1i44.4213

Resumen

En las sociedades de la información, las operaciones, decisiones y elecciones que previamente se otorgaban a los humanos están crecientemente siendo delegadas a algoritmos, los cuales podrían recomendar, o más aun, decidir cómo se deben interpretar los datos y qué acciones deben ser tomadas como resultado. Con mayor asiduidad, los algoritmos median procesos sociales, transacciones bursátiles, decisiones de gobierno y la manera en que percibimos, entendemos e interactuamos entre nosotros y con el ambiente. Las brechas entre el diseño y operación de los algoritmos y nuestra comprensión de sus implicancias éticas podrían tener consecuencias severas afectando individuos, así como también grupos y sociedades completas. Este artículo realiza tres contribuciones para clarificar la importancia ética de la mediación algorítmica: provee un mapa prescriptivo para organizar el debate, revisa la discusión actual acerca de los aspectos éticos de los algoritmos, y evalúa la literatura disponible a fin de identificar áreas que requieran de mayor trabajo para desarrollar la ética de los algoritmos.

Palabras claves: *Algoritmos, automatización, big data, análisis de datos, minería de datos, ética, machine learning.*

1 Traducción al español de: Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: mapping the debate. *Big Data & Society*, *3*(2). https://doi.org/10.1177/2053951716679679. Además de la autorización expresa del autor para realizar la traducción, se cuenta con el permiso respectivo otorgado por parte de SAGE Publications, quienes manejan los derechos del artículo, el número de licencia es el 5411511302779, emitido el 17 de octubre de 2022.

Introducción

En las sociedades de la información, las operaciones, decisiones y elecciones previamente tomadas por humanos están siendo crecientemente delegadas a algoritmos, los cuales podrían aconsejar, o decidir, acerca de cómo los datos deben ser interpretados y qué acciones deberían ser tomadas como resultado.[2]

Los ejemplos abundan. Los algoritmos de perfilado y clasificación determinan cómo se les da forma y se gestiona a individuos y grupos (Floridi, 2012). Los sistemas de recomendación dan a los usuarios directivas acerca de cómo y cuándo ejercitar, qué comprar, qué ruta tomar y a quién contactar (Vries, 2010, p. 81). Se plantea que los algoritmos de minería de datos prometen ayudar a comprender las corrientes emergentes de datos comportamentales generadas por la «internet de las cosas» (Portmess y Tower, 2014, p. 1). Los proveedores de servicios *online* continúan midiendo la forma en que se accede a la información a través de la personalización y el filtrado algorítmico (Newell y Marabelli, 2015; Taddeo y Floridi, 2015). Los algoritmos de *machine learning* algorítmico identifican automáticamente el carácter sesgado, inexacto o engañoso del conocimiento en el punto de su creación (por ejemplo, el servicio de revisión y evaluación objetiva de *Wikipedia*). Como sugieren estos ejemplos, la forma en que percibimos y comprendemos nuestro ambiente e interactuamos con ellos y entre nosotros está crecientemente mediada por algoritmos.

Los algoritmos están cargados de valores inexorablemente (Brey y Soraker, 2009; Wiener, 1988). Los parámetros operacionales son especificados por desarrolladores y configurados por usuarios teniendo en cuenta resultados deseados que privilegian algunos valores e intereses por sobre otros (Friedman y Nissenbaum, 1996; Johnson, 2006; Kraemer et al., 2011; Nakamura, 2013). Al mismo tiempo, operar entre parámetros aceptables no garantiza comportamientos éticamente aceptables. Esto se observa, por ejemplo, con los algoritmos de perfilado que discriminan inadvertidamente a poblaciones marginalizadas (Barocas y Selbst, 2015; Birrer, 2005), como se observó en el envío de publicidades *online* de acuerdo a la etnicidad percibida (Sweeney, 2013).

Determinar el potencial y actual impacto ético de un algoritmo es difícil por muchas razones. Identificar la influencia de la subjetividad humana en el diseño y la configuración algorítmicos requiere usualmente investigación a largo plazo y procesos de desarrollo multiusuario. Aun con los recursos suficientes, los problemas y valores subyacentes no aparecerán hasta que se produzca un caso de uso problemático. Los algoritmos de aprendizaje, usualmente citados como el «futuro» de los algoritmos y la analítica (Tutt, 2016), introducen incertidumbre acerca de cómo y por qué se toman decisiones debido a su capacidad de torcer parámetros operacionales y reglas de toma de decisión «en la naturaleza» (Burrell, 2016). Determinar si una decisión problemática es meramente un «error» único o evidencia de un fallo o sesgo sistémico podría ser imposible (o al menos altamente difícil) con algoritmos de aprendizaje poco interpretables y predecibles. Semejantes desafíos solo crecerán mientras los algoritmos aumentan su complejidad e interactúan entre ellos y sus resultados para tomar decisiones (Tutt, 2016). La brecha resultante entre el diseño y la operación de algoritmos y nuestra comprensión de sus implicancias éticas pueden tener severas consecuencias que afecten individuos, grupos y segmentos completos de una sociedad.

En este artículo, mapearemos los problemas éticos generados por la toma de decisión

2 Nos gustaría reconocer los valiosos comentarios y retroalimentación de los revisores de *Big Data & Society*.

algorítmica [*algorithmic decision-making*]. El artículo responde dos preguntas: ¿qué tipos de cuestiones éticas generan los algoritmos?, y ¿cómo aplican estas cuestiones a los algoritmos directamente, o si, por el contrario, deberían aplicar a las tecnologías construidas sobre algoritmos? Proponemos primero un mapa conceptual basado en seis tipos de preocupaciones que, en conjunto, son suficientes para una organización del campo con base en principios. Argumentamos que el mapa nos permite un diagnóstico más riguroso de los desafíos éticos relacionados con el uso de algoritmos. Luego revisamos la literatura científica que discute aspectos éticos de los algoritmos, evaluando su utilidad y exactitud en el mapa propuesto. Emergieron siete temas en la literatura que demuestran cómo las preocupaciones definidas en el mapa se observan en la práctica. En conjunto, el mapa y la revisión proveen una estructura común para la discusión futura de la ética de los algoritmos. En la sección final del artículo evaluamos cómo se ajustarán el mapa propuesto y los temas referidos en la literatura revisada para identificar áreas de la ética de los algoritmos que requieran mayor investigación. El marco conceptual, revisión y análisis crítico ofrecido en este artículo apunta a informar sobre interrogantes éticos, el desarrollo y la gobernanza de algoritmos.

Antecedentes

Para realizar un mapa de la ética de los algoritmos, debemos primero definir algunos términos clave. «Algoritmo» cuenta con una colección de significados que atraviesa la ciencia computacional, las matemáticas y el discurso público. Como explica Hill, «vemos evidencia que cualquier procedimiento o proceso de decisión, aunque mal definido, puede ser llamado un "algoritmo" en la prensa y el discurso público. Escuchamos hablar, en las noticias, de algoritmos que sugieren potenciales compañeros para personas solteras y algoritmos que detectan tendencias de beneficios financieros a comerciantes, con la implicancia de que esos algoritmos puedan estar en lo correcto o no...» (Hill, 2015, p. 36). Muchos críticos desde la academia también fallan al especificar categorías técnicas o una definición formal de algoritmo (Burrell, 2016; Kitchin, 2016). En ambos casos, el término se usa no en referencia al algoritmo como constructo matemático, sino más bien como la implementación e interacción de uno o más algoritmos en un programa en particular, *software* o sistema de información. Cualquier intento por mapear una ética de algoritmos debe atender a esta fusión entre las definiciones formales y el uso popular de «algoritmo».

Aquí seguimos la definición formal de Hill (2015, p. 47) de un algoritmo como un *constructo matemático* con «una estructura de control finita, abstracta, efectiva y compuesta, dada imperativamente, para cumplir un propósito dado bajo disposiciones dadas». De todas formas, nuestra investigación no se verá limitada a los algoritmos como constructos matemáticos. Como se sugirió en la inclusión de «propósito» y «disposición» en la definición de Hill, los algoritmos deben ser implementados y ejecutados para accionar y surtir un efecto. El uso popular del término se vuelve relevante aquí. Las referencias hacia los algoritmos en el discurso público no refieren normalmente a los algoritmos como constructos matemáticos, sino como implementaciones particulares de ellos. El uso establecido de algoritmo también incluye la *implementación* de un constructo matemático al interior de una tecnología, y la aplicación de esa tecnología *configurada* para una tarea en particular.[3] Un algoritmo completamente configurado incorpora la estructura matemática abstracta

3 Compárese con Turner (2016) sobre la ontología de los programas.

que ha sido implementada en un sistema para el análisis de tareas en un ámbito analítico en particular. Dada esta aclaración, la configuración de un algoritmo a una tarea específica, o a un *dataset* no cambia su representación matemática subyacente o su implementación en un sistema; es, de hecho, un ajuste más de la operación del algoritmo en relación con un caso o problema específico.

En este sentido, no resulta válido considerar la ética de los algoritmos independientemente de cómo son implementados y ejecutados en programas de computadora, *software* y sistemas informacionales. Nuestro fin aquí es hacer un mapa de la ética de los algoritmos, cuando se pretende interpretar «algoritmo» en el marco del discurso público. Nuestro mapa incluirá asuntos éticos que surgen de los algoritmos como constructos matemáticos, implementaciones (tecnologías, programas) y configuraciones (aplicaciones).[4] Allí donde la discusión se focalice en implementaciones y configuraciones (por ejemplo un artefacto con un algoritmo integrado), limitaremos nuestro foco a asuntos relacionados con el trabajo del algoritmo, más que en todos los asuntos relacionados con el artefacto.

De todas formas, como notó Hill anteriormente, un problema con el uso popular de algoritmos es que puede describir «cualquier procedimiento o proceso de decisión», resultando en un amplio rango casi prohibitivo de artefactos a considerar para el ejercicio que nos proponemos. El discurso público está dominado actualmente por preocupaciones con una clase en particular de algoritmo que toma decisiones, por ejemplo, la mejor acción a tomar en una situación dada, la mejor interpretación de datos y demás. Estos algoritmos aumentan o reemplazan el análisis y el proceso de decisión humano, la mayoría de las veces debido al alcance o escala de datos y reglas involucradas. Sin ofrecer una definición de la clase precisa, los algoritmos que nos interesan aquí son aquellos que toman generalmente decisiones confiables (pero subjetivas y no necesariamente correctas) basadas en reglas complejas que desafían o desconciertan las capacidades humanas para la acción y comprensión.[5] En otras palabras, nos interesan los algoritmos cuyas acciones son difíciles de predecir para los humanos o cuya lógica de toma de decisión es difícil de explicar luego de ocurrido el hecho. No nos interesan aquí los algoritmos que automatizan tareas mundanas, por ejemplo, en la manufactura.

Los algoritmos para la toma de decisiones son utilizados en una variedad de ámbitos, desde modelos de toma de decisión simplistas (Levenson y Pettreym, 1994) hasta complejos algoritmos de perfilado (Hildebrandt, 2008). Ejemplos contemporáneos notables incluyen agentes en línea en *softwares* utilizados por proveedores de servicios *online* para realizar operaciones en favor de los usuarios (Kim et al., 2014); algoritmos de resolución de disputas en línea que han remplazado a los seres humanos como tomadores de decisiones en mediaciones (Raymond, 2014; Shackelford y Raymond, 2014); sistemas de filtrado y recomendación que comparan grupos de usuarios para proveerles contenido personalizado (Barnet, 2009); sistemas de apoyo a decisiones clínicas (CDSS) que recomiendan diagnósticos y tratamientos a médicos (Diamond et al., 1987; Mazoué, 1990); y sistemas policíacos predictivos que predicen lugares clave de actividad criminal.

La disciplina de la analítica de datos es un ejemplo sobresaliente, definida como la práctica

4 En busca de la simplicidad, en el resto del artículo nos referiremos genéricamente a algoritmos en lugar de constructos, implementaciones y configuraciones.

5 Tufekci pareciera tener una clase similar de algoritmos en mente en su exploración de detección de daños. Ella describe como *gatekeeping algorithms* a los algoritmos que no resultan en respuestas simples y «correctas». Apuntamos en realidad a aquellas respuestas que son utilizadas en procesos de decisión subjetiva (Tufekci, 2015, p. 206).

del uso de algoritmos para volver comprensibles los flujos de datos. La analítica informa respuestas inmediatas a las necesidades y preferencias de los usuarios de un sistema, así como también brinda planificación estratégica y desarrollo a largo plazo en plataformas o en sus proveedores de servicio (Grindrod, 2014). Identifica relaciones y pequeños patrones a través de un vasto y ampliamente distribuido *dataset* (Floridi, 2012). Se permiten nuevos tipos de búsquedas, incluyendo investigación de comportamientos en datos que han sido «escrapeados» (Lomborg y Bechmann, 2014: 256); búsqueda de comportamientos y preferencias específicas (por ejemplo, orientación sexual u opiniones políticas) (Mahajan et al., 2012); y la predicción de comportamiento a futuro (como se utiliza policialmente, o para dar créditos, seguros y empleo) (Zarsky, 2016). Las *percepciones sobre las que accionar* [*actionable insights*] (más sobre esto luego) son más buscadas que las relaciones causales (Grindrod, 2014; Hildebryt, 2011; Johnson, 2013).

Las analíticas demuestran cómo los algoritmos pueden desafiar la comprensión y los procesos de toma de decisión humanos aun en tareas que previamente eran llevadas adelante por humanos. Para la toma de decisiones (por ejemplo, a qué clase de riesgo pertenece un potencial cliente de un seguro), los algoritmos analíticos trabajan con datos de alta dimensión para determinar qué características son relevantes en determinada decisión. El número de características consideradas en tareas asociadas a cualquier tipo de clasificación como esta puede ser cercana a las decenas de miles. Este tipo de tareas es, por ende, una replicación del trabajo anteriormente llevado a cabo por trabajadores humanos (estratificación de riesgo, por ejemplo) pero involucra una lógica cualitativa de toma de decisiones diferente, cuyo resultado se aplica a grandes cantidades de *inputs*.

Los algoritmos no son éticamente desafiantes solo por la escala de análisis y complejidad de la toma de decisiones. La falta de certeza y opacidad del trabajo realizado por los algoritmos y su impacto es también crecientemente problemática. Los algoritmos requirieron tradicionalmente de reglas y pesos para ser individualmente definidos y programados «a mano». Aunque algunos sigan siendo de este tipo (el *PageRank* de *Google* es un ejemplo sobresaliente de ello), los algoritmos se apoyan cada vez más en sus capacidades de aprendizaje (Tutt, 2016).

Machine learning es «cualquier metodología y conjunto de técnicas que puede emplear datos para generar saberes y patrones nuevos, y generar modelos utilizados para predicciones efectivas sobre esos datos» (Van Otterlo, 2013). El *machine learning* se define por la capacidad de definir o modificar reglas de toma de decisiones de manera autónoma. Un algoritmo de *machine learning* aplicado a tareas de clasificación, por ejemplo, consiste típicamente en dos componentes, un *aprendiz* que produce un *clasificador* con la intención de desarrollar clases que puedan generalizar más allá de los datos de entrenamiento (Domingos, 2012). El trabajo del algoritmo involucra establecer nuevos *inputs* en un modelo o estructura de clasificación. Las tecnologías de reconocimiento de imágenes, por ejemplo, pueden decidir qué tipo de objetos aparecen en una imagen. El algoritmo «aprende» definiendo reglas para determinar cómo serán clasificados los nuevos *inputs*. El modelo puede ser enseñado al algoritmo por *inputs* etiquetados manualmente (aprendizaje supervisado); en otros casos el algoritmo en sí mismo define modelos acordes que hagan tener sentido a un set de *inputs* (aprendizaje sin supervisión)[6] (Schermer, 2011; Van Otterlo, 2013). En ambos casos, el algoritmo define

6 La distinción entre aprendizaje supervisado y sin supervisión puede ser mapeado en las analíticas para revelar las distintas maneras en las que los seres humanos «encuentran sentido» a través de los datos. Las analíticas descriptivas basadas en aprendizaje sin supervisión buscan

las reglas de la toma de decisión para manejar nuevos *inputs*. Críticamente, el operador humano no necesita entender la racionalidad de las reglas de toma de decisión producidas por el algoritmo (Matthias, 2004, p. 179).

Como sugiere esta explicación, las capacidades de aprendizaje garantizan a los algoritmos algún grado de autonomía. El impacto de esta autonomía debe permanecer incierto hasta cierto nivel. Como resultado, las tareas llevadas a cabo por el aprendizaje automatizado son difíciles de predecir de antemano (cómo será manejado un nuevo *input*) o de explicar posteriormente (como fue tomada una decisión en particular). La incertidumbre puede, empero, inhibir la identificación y compensación de los desafíos éticos en el diseño y operación de algoritmos.

Mapa de la ética de los algoritmos

Utilizando los términos clave definidos en la sección previa, proponemos un mapa conceptual (ver Figura 1) basada en seis tipos de preocupaciones que son en su conjunto suficientes para una organización de principios del campo, y conjeturamos que permite un diagnóstico riguroso de los desafíos éticos relacionados con el uso de los algoritmos. El mapa no se propone desde un abordaje particular de la ética, teórico o metodológico, sino que pretende funcionar como un marco prescriptivo de los tipos de cuestiones que surgen de los algoritmos debido a tres aspectos de su funcionamiento. El mapa toma en consideración que los algoritmos sobre los que trata este artículo son utilizados para (1) tornar datos en evidencia para un *outcome específico* (llamado de aquí en más resultado), y que este resultado luego es utilizado para (2) disparar *y* motivar una acción que (por sí misma, o combinada con otras acciones) podría no ser éticamente neutral. Estos trabajos se realizan de maneras complejas y (semi)autónomas, las cuales (3) complican la distribución de la responsabilidad por los efectos o acciones conducidas por el algoritmo. El mapa en sí no está pensado con la intención de ser una herramienta para resolver dilemas éticos surgidos de acciones problemáticas conducidas por algoritmos, sino que se propone como una estructura organizativa basada en la forma en la que operan los algoritmos y que podría estructurar la discusión futura de asuntos éticos. Esto nos lleva a postular tres tipos de preocupaciones éticas epistémicas y dos normativas, basadas en el modo en que los algoritmos procesan datos para producir evidencia y motivar acciones. Estas preocupaciones se asocian con potenciales fracasos que podrían involucrar múltiples actores y, de esa manera, complejizar la pregunta acerca de quién debería ser considerado responsable y/o obligado a rendir cuentas o explicar la situación por esas fallas. Semejantes dificultades motivan la adición de la *trazabilidad (traceability)* como una preocupación última y general.

Evidencia poco concluyente

Cuando un algoritmo saca conclusiones de los datos que procesa utilizando estadística inferencial y/o técnicas de *machine learning*, produce un conocimiento probable[7] y, por eso, inevitablemente incierto. La teoría del aprendizaje estadístico (James et al., 2013) y la teoría del aprendizaje computacional (Valiant, 1984) están interesadas en la caracterización y cuantificación de esta incertidumbre. Sumado a esto, y como ya fue indicado con frecuencia, los métodos estadísticos pueden ayudar a identificar

identificar correlaciones no previstas entre casos para aprender sobre la entidad o fenómeno. Aquí, el análisis es exploratorio, lo cual significa que no posee un objetivo o hipótesis. En ese sentido, se pueden definir nuevos modelos y clasificaciones. En contraste, el análisis predictivo basado en aprendizaje supervisado busca generar casos coincidentes para clases preexistentes que permitan inferir conocimiento sobre el caso. El conocimiento sobre las clases asignadas suele utilizarse para hacer predicciones (Van Otterlo, 2013).

7 El término «conocimiento probable» es utilizado aquí con el sentido de Hacking (2006), quien lo asocia con el surgimiento de la probabilística y el pensamiento estadístico (por ejemplo, en el contexto de los seguros) que comenzó en el siglo XVII.

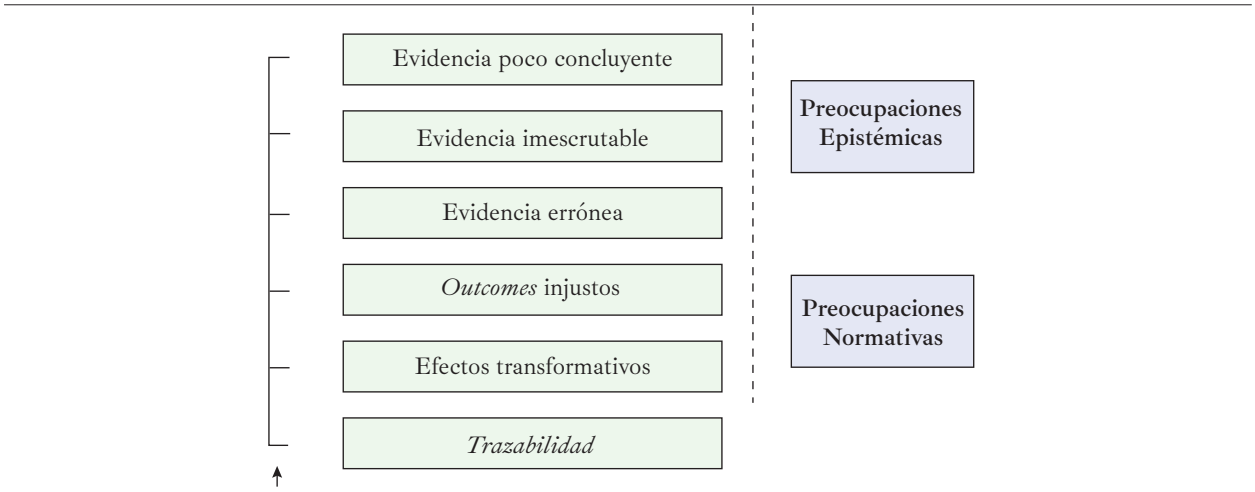

Figura 1. Seis tipos de preocupaciones éticas generadas por algoritmos

correlaciones significativas, pero estas son raramente consideradas como suficientes para postular la existencia de una conexión casual (Illari y Russo, 2014, capítulo 8), podrían ser insuficientes para motivar acciones basadas en el conocimiento de tal conexión. El término *actionable insight* que mencionamos previamente puede ser visto como un reconocimiento explícito de estas limitaciones epistémicas.

Los algoritmos son desplegados típicamente en contextos en los que técnicas más confiables no están disponibles o son muy costosas de implementar y, por ende, raramente se espera que sean infalibles. Reconocer estas limitaciones es importante, pero debería complementarse con evaluaciones acerca de cómo el riesgo de estar equivocado afecta la responsabilidad epistémica individual (Miller y Record, 2013) por ejemplo, volviendo más débil la justificación que uno posee para un resultado, más allá de lo que sería considerable como aceptable, que sustente una acción en un contexto dado.

Evidencia inescrutable

Cuando los datos son utilizados como (o procesados para generar) evidencia para un resultado, es razonable esperar que la conexión entre los datos y el resultado deba ser accesible (por ejemplo, inteligible, abierta al escrutinio y quizá a la crítica).[8] Allí donde la conexión no sea obvia, las expectativas pueden satisfacerse con un mejor acceso, así como también con explicaciones adicionales. Dada la manera en la que opera un algoritmo, estos requerimientos no se satisfacen automáticamente. Una falta de conocimiento en relación con los datos que están siendo utilizados (por ejemplo, relacionados con su alcance, proveniencia y calidad), pero de modo más importante, también la dificultad inherente en la interpretación de cómo cada uno de los muchos *data-points* utilizados por el algoritmo de *machine learning* contribuyen al resultado que genera, causa limitaciones prácticas y también de principios (Miller y Record, 2013).

8 En la epistemología analítica convencional, este asunto se conecta con la naturaleza de la justificación y la importancia de tener acceso a su propia justificación para una creencia específica (Kornblith, 2001). En el contexto actual nos interesa un tipo de justificación más interactivo: los agentes humanos necesitan ser capaces de comprender cómo se justifica una conclusión realizada por un algoritmo a la luz de los datos.

Evidencia errónea

Los algoritmos procesan datos y están por ello sujetos a una limitación compartida por todos los tipos de procesamientos de datos, a saber, que el *output* nunca puede exceder al *input*. Mientras que la teoría matemática de la comunicación de Shannon (Shannon y Weaver, 1998), y especialmente algunas de sus inequidades informacionales, brindan un preciso detalle formal de este hecho, el principio informal que sostiene que «si ingresa basura, sale basura» ilustra con claridad qué es lo que está en juego aquí; específicamente, que los resultados solo pueden ser tan confiables (pero también tan neutrales) como los datos en los que se basan. Las evaluaciones acerca de la neutralidad del proceso, y por ende acerca de si la evidencia producida es errónea o no, dependen por supuesto de un observador.

Outcomes injustos

Las tres preocupaciones epistémicas detalladas hasta aquí apuntan a la calidad de la *evidencia* producida por un algoritmo que motiva una acción particular. Sin embargo, la evaluación ética de los algoritmos también puede hacer foco solamente en la *acción* en sí misma. Las acciones llevadas a cabo por los algoritmos pueden ser evaluadas de acuerdo a numerosos criterios y principios éticos, a los cuales nos referimos genéricamente como la «justicia» dependiente del observador de esas acciones y sus efectos. Una acción puede ser considerada discriminatoria, por ejemplo, únicamente por sus efectos en una clase de personas protegidas, aun cuando esté realizada sobre la base de evidencia concluyente, escrutable y bien fundada.

Efectos transformativos

Los desafíos éticos postulados por el uso extensivo de los algoritmos no pueden siempre ser remontado a casos claros de fracasos éticos o epistémicos, ya que algunos de los efectos de la confianza en el procesamiento de datos algorítmicos y tomas de decisión (semi)autónomas pueden ser cuestionables y aun así parecer éticamente neutrales porque no parecen causar ningún daño a simple vista. Esto se debe a que los algoritmos pueden afectar el modo en que conceptualizamos el mundo y modificar su organización social y política (Floridi, 2014). Las actividades algorítmicas, como el perfilado, re-ontologizan al mundo comprendiendo y conceptualizándolo en nuevas e inesperadas formas, desencadenando y motivando acciones basadas en los *insights* que generan.

Trazabilidad

Los algoritmos son artefactos de *software* utilizados en el procesamiento de datos, y como tales heredan desafíos éticos asociados con el diseño y disponibilidad de nuevas tecnologías y aquellos asociados a la manipulación de grandes volúmenes de datos personales y de otros tipos. Esto implica que el daño causado por la actividad algorítmica es difícil de depurar (por ejemplo, detectar el daño y encontrar su causa), pero también que raramente es fácil identificar quién debería ser considerado responsable por el daño causado.[9] Cuando se identifica un problema vinculado a todos o alguno de los cinco tipos precedentes, las evaluaciones éticas requieren rastrear tanto la causa como la responsabilidad por el daño causado.

9 La opacidad de los algoritmos puede explicar solo parcialmente esta cuestión. Otro aspecto está más cercanamente vinculado al rol de la reutilización en el desarrollo de algoritmos y artefactos basados en *software*; desde el habitual uso de bibliotecas existentes hasta la reutilización de herramientas preexistentes y métodos para diferentes propósitos (por ejemplo, el uso de modelos sismológicos en vigilancia predictiva (Mohler et al., 2011), pasando por el desarrollo de herramientas generales para métodos específicos. Más allá de la inevitable distribución de responsabilidades, esto resalta la compleja relación entre el buen diseño (la filosofía de la reutilización promovida en la *programación estructurada*) y la ausencia de mal funcionamiento, y revela que aun los diseñadores de artefactos basados en *software* tratan regularmente gran parte de su trabajo como cajas negras (Sametinger, 1997).

Gracias a este mapa (Figura 1), ahora somos capaces de distinguir los tipos epistemológicos, los estrictamente éticos y los de trazabilidad en las descripciones de los problemas éticos con los algoritmos. El mapa está destinado a servir como una herramienta para organizar una gran variedad de discursos académicos que se refieren a la diversidad de tecnologías unificadas por su confianza en los algoritmos. Para evaluar la utilidad del mapa, y observar cómo cada uno de estos tipos de preocupaciones se manifiestan en problemas éticos ya observados en algoritmos, se llevó a cabo una revisión sistematizada de literatura académica.[10] Las siguientes secciones (4 a 10) describen cómo son tratados los asuntos y conceptos éticos en la literatura que discute explícitamente los aspectos éticos de los algoritmos.

Evidencia poco concluyente que conduce a acciones injustificadas

Mucho del proceso de decisión algorítmica y minería de datos descansa en el conocimiento inductivo y en correlaciones identificadas dentro de un *dataset*. La causalidad no se establece previamente a la toma de acción sobre la evidencia producida por el algoritmo. La búsqueda de vínculos causales es difícil en la medida en que las correlaciones establecidas a partir de *datasets* de gran volumen y propietarios no son, con frecuencia, reproducibles ni falsificables (Ioannidis, 2005; Lazer et al., 2014). A pesar de esto, las correlaciones basadas en un volumen suficiente de datos son vistas cada vez más como suficientemente creíbles para dirigir acciones sin establecer, primero, una causalidad (Hilderbryt, 2011; Hildebryt y Kroops, 2010; Mayer-Schónberg y Cukier, 2013; Zarsky, 2016). En este sentido, la minería de datos y los algoritmos de perfilización solo necesitan establecer una base de evidencia confiable suficiente para dirigir la acción, referida como *actionable insights*.

Actuar sobre correlaciones puede ser doblemente problemático.[11] Se pueden descubrir más correlaciones espurias que conocimiento causal genuino. En análisis predictivo, las correlaciones son doblemente inciertas (Ananny, 2016). Aun si se encuentran correlaciones sólidas o conocimiento causal, estas son direccionadas hacia individuos (Illari y Russo, 2014). Como explica Aanny (2016, p. 103), «las categorías algorítmicas [...] indican certeza, desalientan exploraciones alternativas y crean coherencia entre objetos dispares», todo lo cual contribuyen a que los individuos sean descritos (posiblemente de forma inexacta) a través de modelos o clases simplificadas (Barocas, 2014). Finalmente, si tanto las acciones como el conocimiento se ubican en el nivel de la población, nuestras acciones podrían volcarse sobre el nivel individual. Esto ocurre, por ejemplo, cuando se establece una prima de seguros para una cierta subpoblación, y por ende debe ser pagada por cada miembro. La acción tomada sobre la base de correlaciones inductivas tiene un impacto real en intereses humanos independientemente de su validez.

Evidencia inescrutable que conduce a la opacidad

La escrutabilidad de evidencia, evaluada en términos de la transparencia u opacidad de los algoritmos, probó ser una preocupación mayor en la literatura revisada. La transparencia es algo generalmente deseado porque los algoritmos que son poco predecibles o

10 Ver el Apéndice 1 para información acerca de la metodología, búsqueda de términos y resultados de búsqueda de la revisión.

11 De todas formas, debe realizarse una distinción entre la justificación ética de actuar basándose en meras correlaciones y una ética más amplia de razonamiento inductivo, la cual se superpone con las críticas existentes de métodos estadísticos y cuantitativos en investigación. Las primeras se vinculan con los umbrales de la evidencia requerida para justificar acciones con impacto ético. Las segundas aluden a la falta de reproducibilidad en la analítica que la distinga en la práctica de la ciencia (Feynman, 1974; Ioannidis, 2005; Vasilevsky et al., 2013), y es mejor comprendida como un asunto epistemológico.

explicables son difíciles de controlar, monitorear y corregir (Tutt, 2016). Como han observado muchos críticos (Crawford, 2016; Neyly, 2016; Raymond, 2014), la transparencia es usualmente tratada de manera *naïf* como una panacea para los asuntos éticos que surgen de las nuevas tecnologías. La transparencia es generalmente definida en relación con «la disponibilidad de información, las condiciones de accesibilidad y la forma en que la información [...] podría, pragmática o epistémicamente, apoyar el proceso de toma de decisión del usuario» (Turilli y Floridi, 2009, p. 106). El debate acerca de este tópico no es nuevo. La literatura acerca de la información y la ética computacional, por ejemplo, empezó a enfocarse en él a comienzos del siglo XXI, cuando crecieron los temas relacionados con la información algorítmica y el filtrado por motores de búsqueda.[12] Los componentes primarios de la transparencia son la *accesibilidad* y *comprensibilidad* de la información. La información acerca de la funcionalidad de los algoritmos es en general poco accesible. Los algoritmos de propiedad privada son mantenidos en secreto en nombre de la ventaja competitiva (Glenn y Monteith, 2014; Kitchin, 2016; Stark y Fins, 2013), la seguridad nacional (Leese, 2014) o la privacidad. Sin embargo, la transparencia puede ir en contra de otros ideales éticos, en particular, la privacidad de los datos de los sujetos y la autonomía de las organizaciones.

Granka (2010) observa una lucha de poder entre los intereses de los sujetos y la viabilidad comercial de quienes procesan esos datos. Dar a conocer la estructura de estos algoritmos podría facilitar manipulaciones con intenciones negativas de resultados de búsqueda (o «jugar con el sistema»), sin tener por ello ventaja alguna para el usuario neófito en tecnología (Granka, 2010; Zarsky, 2016). La viabilidad comercial de los procesadores de datos en muchas industrias (como los informes crediticios y el comercio de alta frecuencia) podrían verse amenazados por la transparencia. De todas maneras, los sujetos implicados están interesados en comprender cómo se crea la información acerca de ellos y de qué modo influye en las decisiones tomadas en prácticas direccionadas por los datos [*data-driven practices*]. Esta lucha está marcada por la asimetría de información y por un «desbalance en el conocimiento y poder del proceso de toma de decisiones», favoreciendo a quienes procesan datos (Tene y Polonetsky, 2013a, p. 252).

Además de ser accesible, la información debe de ser comprensible para ser considerada transparente (Turilli y Floridi, 2009). Los esfuerzos por volver transparentes a los algoritmos enfrentan un desafío significativo al hacer accesibles y comprensibles complejos procesos de toma de decisión. El problema de larga data de la capacidad de interpretación en los algoritmos de *machine learning* indica el desafío de la opacidad en los algoritmos (Burrel, 2016; Hildebryt, 2011; Leese, 2014; Tutt, 2016). El *machine learning* es experto en crear y modificar las reglas para clasificar y empaquetar [*cluster*] vastos *datasets.* El algoritmo modifica su estructura comportamental durante las operaciones (Markowetz et al., 2014). Esta alteración de cómo el algoritmo clasifica nuevos *inputs* es el modo en que aprende (Burrell, 2016, p. 5). El entrenamiento produce una estructura (clases, conglomerados, rankings, pesos, etc.) para clasificar nuevos *inputs* o predecir variables

12 El artículo de Introna & Nissenbaum (2000) se encuentra entre las primeras publicaciones sobre este tópico. El artículo compara motores de búsqueda con editores y sugiere que, como los editores, los motores de búsqueda filtran la información de acuerdo con las condiciones de mercado, por ejemplo, de acuerdo a los gustos y preferencias de los consumidores, y favorecen a los actores poderosos. Dos mecanismos correctivos son sugeridos: embeber el «valor de lo justo, así como un conjunto de valores representados por la ideología de la Web como un bien público» (Introna y Nissenbaum, 2000, p. 182) en el diseño de la indexación y rankeo algorítmico, y la transparencia de los algoritmos utilizados por los motores de búsqueda. Recientemente, Zarsky (2013) ha aportado un marco teórico y una revisión legal en profundidad de la transparencia en la analítica predictiva.

desconocidas. Una vez entrenado, nuevos datos pueden ser procesados y categorizados automáticamente sin intervención de un operador (Leese, 2014). La lógica del algoritmo es oscurecida, brindando motivos para retratar a los algoritmos de *machine learning* como «cajas negras».

Burrell (2016) y Schermer (2011) argumentan que la opacidad de los algoritmos de *machine learning* inhibe la vigilancia. Los algoritmos «son opacos en el sentido que uno es el destinatario del *output* del algoritmo (la decisión clasificatoria), pero raramente uno posee un sentido completo de cómo o por qué se ha llegado a esta particular clasificación a partir de un cierto *input*» (Burrell, 2016, p. 1). Tanto el *input* (datos sobre humanos) como los *outputs* (clasificaciones) pueden ser desconocidos o imposibles de conocer. La opacidad en los algoritmos de *machine learning* es producto de la alta dimensionalidad de los datos, la complejidad del código y la lógica variable del proceso de toma de decisiones (Burrell, 2016). Matthias (2004, p. 179) sugiere que el *machine learning* puede generar *outputs* para los cuales «el entrenador humano mismo es incapaz de proveer una representación algorítmica». Los algoritmos pueden considerarse explicables hasta el grado en que un humano pueda articular al modelo entrenado o dar una razón lógica para una decisión particular, por ejemplo, explicar la influencia (cuantificada) de *inputs* particulares o atributos (Datta et al., 2016). La vigilancia significativa y la intervención humana en la decisión algorítmica «es imposible cuando la máquina tiene una ventaja informacional sobre el operador... [o] cuando la máquina no puede ser controlada por un humano en tiempo real debido a su velocidad de procesamiento y la cantidad de variables operacionalizables» (Matthias, 2004, pp. 182-183). Este es, nuevamente, el problema de la caja negra. De todas maneras, se debe plantear una distinción entre la inviabilidad técnica de la vigilancia y las barreras prácticas causadas, por ejemplo, por falta de experiencia, acceso o recursos.

Más allá del *machine learning*, los algoritmos con reglas de toma de decisiones «escritas a mano» pueden ser también altamente complejos y prácticamente inescrutables desde la perspectiva del sujeto cuyos datos estén implicados (Kitchin, 2016). Las estructuras de toma de decisión algorítmicas que contienen «cientos de reglas son muy difíciles de inspeccionar visualmente, especialmente cuando sus predicciones se combinan probabilísticamente en formas complejas» (Van Otterlo, 2013). Más aún, los algoritmos son a menudo desarrollados por grandes equipos de ingenieros a lo largo del tiempo, con una comprensión holística del proceso de desarrollo enmarcado en sus valores, sesgos e interdependencias inevitablemente reproducidas (Syvig et al., 2014). En ambos sentidos, el procesamiento algorítmico contrasta con la toma de decisiones tradicional, donde los decisores humanos pueden, en principio, articular su base lógica cuando les sea requerido, limitados solo por su deseo y capacidad de brindar una explicación y por la capacidad de comprenderla por parte de quien consulta. Por el contrario, la base lógica de un algoritmo puede ser incomprensible para los humanos, haciendo que la legitimidad de la decisión sea difícil de cuestionar.

Bajo estas condiciones, la toma de decisión es poco transparente. Rubel y Jones (2014) argumentan que el fracaso para volver comprensible a los sujetos la lógica procesual ofende la agencia de estos (volveremos a este punto en la sección 8). El consentimiento informado al procesamiento de datos no es posible si la opacidad excluye la evaluación de riesgos (Schermer, 2011). Brindar información sobre la lógica algorítmica de toma de decisiones en un formato simplificado puede ayudar (Datta et al., 2016; Tene y Polonets-

ky, 2013a). De todas formas, las estructuras complejas de toma de decisión pueden exceder rápidamente los recursos humanos y organizacionales disponibles para la vigilancia (Kitchin, 2016). Como resultado, los sujetos pueden perder la confianza tanto en los algoritmos como en los procesadores de datos (Cohen et al., 2014; Rubel y Jones, 2014; Shackelford y Raymond, 2014)[13].

Aun cuando los procesadores y controladores de datos den a conocer información operacional, el beneficio neto para la sociedad es incierto. La falta de interés público respecto de los mecanismos de transparencia existentes refleja esta incertidumbre, que se ve por ejemplo en los *scorings* crediticios (Zarsky, 2016). La divulgación de la transparencia podría tener más impacto si se efectuaran desde terceros entrenados o reguladores que representen el interés público como algo opuesto a los sujetos implicados (Tutt, 2016; Zarsky, 2013).

El proceso de transparentar por parte de procesadores y controladores de datos será crucial en el futuro para mantener una relación de confianza con los sujetos (Cohen et al., 2014; Rubel y Jones, 2014; Shackelford y Raymond, 2014). La confianza implica las expectativas de quien confía (el agente que confía) hacia el depositario de la misma (el agente en quien se confía) para efectuar una tarea (Taddeo, 2010), y la aceptación del riesgo de que el depositario de esa confianza vaya a traicionar esas expectativas (Wiegel y Berg, 2009). La confianza en los procesadores de datos puede, por ejemplo, aliviar las preocupaciones acerca del procesamiento opaco de datos personales (Mazoué, 1990). De todas maneras, puede existir confianza entre agentes artificiales exclusivamente, como, por ejemplo, en los agentes de un sistema distribuido trabajando cooperativamente para lograr un cierto objetivo (Grodzinsky et al., 2010; Simon 2010; Taddeo, 2010). Más aun, los algoritmos pueden ser percibidos como confiables independientemente de (o quizá a pesar de) cualquier confianza puesta en quien procesa los datos; incluso queda abierta la pregunta acerca de cuándo esto podría ser apropiado.[14]

Evidencia errónea que conduce a sesgos

La automatización de la toma de decisión humana es a menudo justificada por una supuesta falta de sesgos en los algoritmos (Bozdag, 2013; Naik y Bhide, 2014). Esta creencia es insostenible, como ha sido demostrado por trabajos anteriores que probaron la normatividad de las tecnologías de la información en general y en el desarrollo de algoritmos en particular[15] (Bozdag, 2013; Friedman y Nissenbaum, 1996; Kraemer et al., 2011; Macnisj, 2012; Newell y Marabelli, 2015, p. 6; Tene y Polonetsky, 2013b). Mucha de la literatura revisada apunta a la forma en la que los sesgos se manifiestan en los algoritmos y la evidencia que producen.

Los algoritmos toman decisiones sesgadas inevitablemente. El diseño y funcionalidad algorítmica refleja los valores de su dise-

13 Esta es una afirmación controvertida. Bozdag (2013) sugiere que la comprensión de los humanos no tuvo un crecimiento paralelo al exponencial de los datos sociales en los años recientes debido a las limitaciones biológicas sobre las capacidades de procesamiento de información. De todos modos, esto parecería descartar los avances en la visualización de datos y técnicas de clasificación para asistir a los humanos en la comprensión de amplios *datasets* y flujos de información (Turilli y Floridi, 2009). Las capacidades biológicas podrían no haberse incrementado, pero no se puede decir lo mismo de la comprensión asistida. Nuestra posición al respecto gira en torno a definir si lo *asistido tecnológicamente* y la *comprensión humana* son categóricamente diferentes.

14 El contexto de sistemas de armamento autónomo es particularmente relevante en este sentido, ver Swiatek (2012).

15 El argumento que sugiere que el diseño tecnológico está cargado inevitablemente de valores no es universalmente aceptado. Kraemer et al. (2011) proveen un contraargumento desde la literatura revisada. Para ellos, los algoritmos están cargados de valores solo «si uno no puede racionalmente elegir entre ellos sin tener en cuenta preocupaciones éticas explícita o implícitamente». En otras palabras, los diseñadores realizan juicios de valor que expresan puntos de vista «acerca de cómo deberían ser las cosas o no, o qué es bueno o malo, deseable e indeseable» (Kraemer et al., 2011, p. 252). Para Kraemer et al. (2011), los algoritmos que producen juicios de valor hipotéticos o recomiendan cursos de acción, como los sistemas de apoyo a decisiones clínicas, pueden ser neutrales en sus valores porque los juicios producidos son hipotéticos.

ñador y sus usos intencionados, aun desde el momento en que se elige un diseño particular como el mejor o el más eficiente. El desarrollo no es un camino neutral, lineal; no hay una opción objetivamente correcta en ninguna etapa dada, sino muchas opciones posibles (Johnson, 2006). Como resultado, «los valores del autor (de un algoritmo), de forma voluntaria o no, son congelados al interior del código, institucionalizando efectivamente esos valores» (Macnish, 2012, p. 158). Es difícil detectar sesgos latentes en los algoritmos y los modelos que producen al enfrentarse con la historia de desarrollo del algoritmo de manera aislada (Friedman y Nissenbaum, 1996; Hildebryt, 2011; Morek, 2006).

Friedman y Nissenbaum (1996) argumentan que los sesgos pueden surgir desde (1) valores sociales preexistentes hallados en las «instituciones sociales, prácticas y actitudes» desde las cuales surge la tecnología, (2) obstáculos técnicos, y (3) aspectos emergentes de un contexto de utilización. Los sesgos sociales pueden estar imbricados en el diseño del sistema con un propósito por diseñadores individuales, como se ha visto, por ejemplo, en ajustes manuales a índices de motores de búsqueda y criterios de ranking (Goldman, 2006). Los sesgos sociales pueden ser también no intencionados, un mero reflejo de valores culturales u organizacionales. Por ejemplo, los algoritmos de *machine learning* entrenados con datos etiquetados por humanos aprenden inadvertidamente a reflejar sesgos de quienes realizan esas etiquetas (Diakopoulos, 2015).

Los sesgos técnicos surgen de limitaciones tecnológicas, errores o decisiones de diseño que favorecen a un grupo particular sin un valor direccionado subyacente (Friedman y Nissenbaum, 1996). Hay ejemplos como el del listado alfabético de compañías aéreas que lleva a incrementar los negocios para aquellos que están más arriba en el alfabeto, o un error en el diseño de un generador de números aleatorio que causa que ciertos números sean favorecidos. Los errores también pueden manifestarse en los *datasets* procesados por los algoritmos. Los fallos en los datos son adoptados sin advertencia por los algoritmos y escondidos en los *outputs* y los modelos que producen (Barocas y Selbst, 2015; Romei y Ruggeri, 2014).

Los sesgos emergentes se vinculan con los avances en el conocimiento o cambios introducidos intencionalmente al sistema por los usuarios y los involucrados (Friedman y Nissenbaum, 1996). Por ejemplo, los *Clinical Decision Support Systems* (CDSS) están inevitablemente sesgados hacia tratamientos incluidos en su arquitectura de decisiones. Aunque los sesgos emergentes se vinculan con los usuarios, otros pueden surgir inesperadamente desde reglas decisionales desarrolladas por el algoritmo, más que por cualquier estructura de toma de decisión «escrita a mano» (Hajian y Domingo-Ferrer, 2013; Kamiran y Calders, 2010). El monitoreo humano podría prevenir que algunos sesgos ingresen en la toma de decisión del algoritmo en estos casos (Raymond, 2014).

Los *outputs* de los algoritmos también requieren interpretación (por ejemplo, lo que uno debería hacer basado en lo que indica el algoritmo); para los datos comportamentales, las correlaciones «objetivas» pueden ayudar a reflejar «las motivaciones inconscientes, emociones particulares, decisiones deliberadas, determinaciones socio-económicas, y las influencias geográficas o demográficas» de quien interpreta esos datos (Hildelbryt, 2011, p. 376). Explicar las correlaciones en cualquiera de estos términos requiere de justificación

Este abordaje sugeriría que los algoritmos autónomos poseen valores por definición, pero solo porque los juicios producidos se ponen en acción por cuenta del algoritmo. Esta concepción de la neutralidad como valor pareciera sugerir que los algoritmos son diseñados en espacios neutrales en lo relativo a valores, con el diseñador desconectado de un contexto social y moral y una historia que inevitablemente influencia sus percepciones y decisiones. Es difícil observar cómo sería este el caso (Friedman y Nissenbaum, 1996).

adicional, lo cual significa que no es autoevidente en los modelos estadísticos. Distintas métricas «vuelven visibles aspectos de individuos o grupos que no serían perceptibles de otra forma» (Lupton, 2014, p. 859). Aun así, no puede asumirse que la interpretación de quien observa va a reflejar correctamente la percepción del actor más que los sesgos del interpretante.

Resultados injustos que conducen a la discriminación

Mucha de la literatura revisada también se refiere a resultados discriminatorios a partir de evidencia sesgada y procesos de toma de decisión.[16] La perfilización por algoritmos, ampliamente definida «como la construcción o inferencia de patrones por medio de la minería de datos y [...] la aplicación de los consiguientes perfiles a personas cuyos datos coinciden con ellos» (Hildelbryt y Koops, 2010, p. 431), es citada frecuentemente como una fuente de discriminación. Los algoritmos de perfilización identifican correlaciones y hacen predicciones sobre comportamientos a nivel grupal, aunque se trate de grupos (o perfiles) que están en constante cambio y redefinición por parte del algoritmo (Zarsky, 2013). Ya sea dinámico o estático, el individuo es comprendido basándose en conexiones con otros identificados por el algoritmo, más que por su comportamiento real (Danna y Marabelli, 2015, p. 5). Las elecciones de los individuos se estructuran de acuerdo con la información sobre el grupo (Danna y Gyy, 2002, p. 382). La perfilización puede crear inadvertidamente una evidencia de base que conduzca a la discriminación (Vries, 2010).

Para las partes afectadas, el tratamiento discriminatorio *data-driven* es menos probable de ser aceptable que la discriminación generada por prejuicios o evidencia anecdótica.

Esto es lo que está implícito en el argumento de Schermer (2011), quien sugiere que el tratamiento discriminatorio no es éticamente problemático en sí mismo, sino que son los efectos del tratamiento los que determinan su aceptabilidad en términos éticos. De todas formas, Schermer confunde sesgos y discriminación en un solo concepto. Lo que él llama discriminación puede ser descrito como simples sesgos, o la expresión consistente y repetida de una preferencia en particular, creencia o valoración en el proceso de toma de decisión (Friedman y Nissenbaum, 1996). En contraste, lo que describe como efectos problemáticos del tratamiento discriminativo puede definirse como discriminación *tout court*.[17] Entonces, los sesgos son una dimensión del proceso de toma de decisión en sí mismos, mientras que la discriminación describe los efectos de una decisión, en términos de impacto adverso y desproporcionado resultante de una toma de decisión algorítmica. Barocas y Selbst (2015) muestran que precisamente esta definición guía «la detección del impacto dispar», un mecanismo de refuerzo para la ley antidiscriminación estadounidense en áreas como desarrollo social y empleo. Ellos sugieren que la detección del impacto dispar provee un modelo para la detección de sesgos y discriminación en la toma de decisión mediada por algoritmos que es sensible para una privacidad diferencial.

Podría ser posible direccionar algoritmos para que no consideren atributos sensibles que contribuyan a discriminar (Barocas y Selbst, 2015), tales como género o etnicidad (Calders et al., 2009; Kamiran y Calders, 2010; Schermer, 2011), basándose en el surgimiento de discriminación en un contexto

16 No se identificaron fuentes claras de discriminación en la literatura revisada. Barocas (2014) aporta con claridad cinco posibles fuentes de discriminación vinculadas a analíticas sesgadas: (1) pertenencia inferida a una clase protegida; (2) sesgos estadísticos; (3) inferencias defectuosas; (4) inferencias demasiado precisas; (5) cambios en el encuadre de la muestra.

17 En francés en el original. Se sugiere interpretación como «a secas».

en particular. De todas formas, la protección a través de variables sustitutivas o cuidado de ciertos atributos no es fácil de predecir ni de detectar (Romei y Ruggieri, 2014; Zarsky, 2016), particularmente cuando los algoritmos acceden a *datasets* vinculados (Barocas y Selbst, 2015). Los perfiles construidos a partir de características neutrales como puede ser un código postal pueden, de manera inadvertida, superponerse con otros perfiles vinculados a etnicidad, género, preferencias sexuales, entre muchos otros (Macnish, 2012; Schermer, 2011).

Hay esfuerzos puestos en marcha que buscan evitar tal exclusión a través de los atributos sensibles o variables sustitutivas. Romei y Ruggieri (2014) observan cuatro estrategias coincidentes para la prevención de la discriminación en analíticas: (1) distorsión controlada del entrenamiento de datos; (2) integración de criterios antidiscriminación en el algoritmo clasificatorio; (3) procesamiento posterior de los modelos de clasificación; (4) modificación de las predicciones y decisiones para mantener una justa proporción de efectos entre grupos protegidos y desprotegidos. Estas estrategias son tenidas en cuenta en el desarrollo de la minería de datos en pos de la preservación de la privacidad, el trato equitativo y la conciencia de la discriminación (Dwork et al., 2011; Kamishima et al., 2012). La minería de datos con conciencia de equidad toma el abordaje más amplio y presta atención no solo a la discriminación, sino también a la equidad, la neutralidad y la independencia (Kamishima et al., 2012). Variadas métricas de equidad posiblemente estén basadas en paridad estadística, privacidad diferencial y otras relaciones entre los sujetos implicados en las tareas de clasificación (Dwork et al., 2011; Romei y Ruggieri, 2014).

La práctica de la personalización puede segmentar una población de manera que solo algunos segmentos sean dignos de recibir ciertas oportunidades o información, reforzando las (des)ventajas sociales existentes. Son comunes las preocupaciones vinculadas a la equidad y al tratamiento igualitario de estas prácticas (Cohen et al., 2014; Danna y Gandy, 2002; Rubel y Jones, 2014). La asignación de precios personalizada, por ejemplo, puede ser «una invitación a irse silenciosamente» hacia ciertos sujetos implicados debido a su falta de valor o capacidad de pago.[18]

Las razones para considerar los efectos discriminatorios como *adversos* y éticamente problemáticos son diversas. Las analíticas discriminatorias pueden contribuir a profecías autocumplidas y a la estigmatización en grupos específicos, disminuyendo su autonomía y participación en la sociedad (Barocas, 2014; Leese, 2014; Macnish, 2012). La personalización a través de la perfilización no distributiva, presente, por ejemplo, en los precios a medida de la personalización en los seguros (Hildebryt y Koops, 2010; Van Wel y Royakkers, 2004), puede ser discriminatoria violando los principios, tanto éticos como legales, de tratamiento igualitario y equitativo de los individuos (Newell y Marabelli, 2015). Más aun, como ya ha sido descrito, la capacidad de esos individuos de investigar la relevancia de factores personales utilizados en el proceso de toma de decisión está inhibida por la opacidad y la automatización (Zarsky, 2016).

18 Danna y Gandy (2002) proveen un ejemplo demostrativo en el Banco Real de Canadá el cual impulsaba a los clientes con pago de tarifa por servicio hacia un servicio con tarifa plana luego de descubrir (a través de la minería de datos propia) que los segundos le agregan mayor valor al banco. Aquellos clientes que se negaran a moverse hacia la tarifa plana por el servicio se enfrentaban con desincentivos, incluyendo precios más altos. A través de la discriminación por el precio del servicio, los clientes fueron empujados hacia opciones que reflejaban exclusivamente los intereses del banco. Aquellos que se negaron a pasarse fueron puestos en una posición de negociación inferior en la cual eran «invitados a irse». El hecho de que se perdieran algunos clientes en el proceso del paso de la mayoría a paquetes de tarifa plana más redituables significó que el banco no mostró incentivo en dar lugar a los intereses de la minoría a pesar del riesgo de perder esos clientes a manos de sus competidores.

Efectos transformadores que generan desafíos con la autonomía

Las decisiones cargadas de valores tomadas por algoritmos pueden también plantear una amenaza a la autonomía de los sujetos. La literatura revisada conecta particularmente a los algoritmos de personalización con estas amenazas. La personalización puede ser definida como la construcción de arquitecturas de elección que no son las mismas aplicadas en una misma muestra (Tene y Polonetsky, 2013a). Similares a las tecnologías persuasivas explícitas, los algoritmos pueden modificar el comportamiento de los sujetos y de los humanos que toman las decisiones a través del filtrado de información (Ananny, 2016). Distintos contenidos, información, precios, etc., se ofrecen a grupos o clases de personas dentro de una población de acuerdo a atributos particulares, por ejemplo, la capacidad de pago.

Los algoritmos de personalización trazan una fina línea entre apoyar y controlar decisiones a través del filtrado por el cual la información es presentada al usuario basado en el entendimiento en profundidad de preferencias, comportamientos y quizás vulnerabilidades a influir (Bozdagh, 2013; Goldman, 2006; Newell y Marabelli, 2015; Zarsky, 2016). Corrientes de datos comportamentales y clasificatorios son utilizados para vincular la información con los intereses y atributos de los sujetos. La autonomía de los sujetos en la toma de decisión no se respeta cuando la elección deseada refleja los intereses de un tercero por sobre los del individuo (Applin y Fischer, 2015; Stark y Fins, 2013).

Esta situación resulta paradójica. En principio, la personalización debería mejorar la toma de decisión, proveyendo al sujeto de información únicamente relevante al verse confrontado con una sobrecarga de información potencial; de todas maneras, decidir qué información es relevante es inherentemente subjetivo. El sujeto puede ser empujado a realizar «la acción institucionalmente predilecta por encima de su propia preferencia» (Johnson, 2013); los consumidores en línea, por ejemplo, pueden ser impulsados a adecuarse a las necesidades del mercado a través del filtrado sobre cómo se posicionan los productos (Coll, 2013). Lewis y Westlund (2015, p. 14) sugieren que los algoritmos de personalización necesitan ser entrenados para «actuar éticamente» y así lograr un equilibrio entre la coerción y el apoyo a la autonomía decisoria de los usuarios.

Los algoritmos de personalización reducen la diversidad de información que encuentran los usuarios a través de la exclusión del contenido considerado irrelevante o contradictorio de acuerdo con las creencias del usuario (Barnett, 2009; Pariser, 2011). La diversidad en la información puede entonces ser considerada como condición habilitante para la autonomía (van den Hoven y Rooksby, 2008). Los algoritmos de filtrado que crean «cámaras de eco» desprovistas de información contradictoria pueden impedir la autonomía en términos de decisiones (Newell y Marabelli, 2015). Los algoritmos podrían no ser capaces de replicar el «descubrimiento espontáneo de nuevas cosas ideas y opciones» que se consideren como anomalías contrarias a los intereses del sujeto perfilizado (Raymond, 2014). Con el acceso casi omnipresente a la información que ahora es posible en la era de internet, las cuestiones del acceso se refieren al hecho de que se pueda llegar a la información «correcta», más que a cualquier información. El control sobre la personalización y los mecanismos de filtrado pueden mejorar la autonomía del usuario, pero al costo potencial de la pérdida de diversidad de información (Bozdag, 2013). Así, los algoritmos de personalización, y la práctica subyacente de las analíticas, podrían mejorar tanto como socavar la agencia de los sujetos.

Efectos transformadores que generan desafíos con la privacidad informacional

Los algoritmos también están transformando las nociones de privacidad. Las respuestas a la discriminación, la desindividualización y las amenazas de las tomas de decisión opacas para la agencia de los sujetos implicados apelan a menudo a la privacidad informacional (Schermer, 2011), o al derecho de los sujetos de «proteger los datos personales de terceros». La privacidad informacional se vincula con la capacidad de un individuo de controlar la información sobre sí mismo (Van Wel y Royakkers, 2004), y con el esfuerzo requerido por esos terceros para obtener información.

El derecho a la identidad derivado de los intereses por la privacidad informacional sugiere que el perfilado opaco o secreto es problemático.[19] La toma de decisiones a través del uso de algoritmos es opaca (ver la sección «Evidencia poco concluyente que conduce a acciones injustificadas») e inhibe la supervisión y las tomas de decisión informadas en relación con la forma en que se comparten los datos (Kim et al., 2014). Los sujetos no pueden definir las normas de privacidad para gobernar todo tipo de datos de manera genérica porque su valor solo está establecido a través del procesamiento (Hildelbryt, 2011; Van Wel y Royakkers, 2004).

Más allá de la opacidad, las protecciones de privacidad basadas en la identificabilidad son poco adecuadas para limitar el manejo externo de la identidad vía analíticas. La identidad es influida cada vez más por el conocimiento producido a través de las analíticas que generan nociones sobre corrientes de datos comportamentales. El «individuo identificable» no es necesariamente una parte del proceso. Schermer (2011) argumenta que la privacidad informacional es un marco conceptual inadecuado porque la perfilización vuelve irrelevante la identificabilidad de los sujetos implicados.

La perfilización busca ensamblar individuos en grupos dotados de sentido, para los cuales la identidad es irrelevante (Floridi, 2012; Hildebryt, 2011; Leese, 2014). Van Wel y Royakkers (2004, p. 133) sostienen que la construcción de identidad externa por algoritmos es un tipo de desindividualización, o una «tendencia a juzgar y tratar a las personas sobre las bases de las características grupales más que en las características individuales de cada uno de ellos y sus méritos». Nunca es necesario identificar a los individuos cuando el perfil es construido para ser afectado por el saber y las acciones que deriven de ello (Louch et al., 2010, p. 4). La identidad informacional del individuo (Floridi, 2011) se viola por el sentido que generan los algoritmos que vinculan al sujeto con otros dentro de un mismo *dataset* (Vries, 2010).

Las protecciones regulatorias actuales luchan por responder a los riesgos de la privacidad informacional de las analíticas. Los datos personales se definen en la ley europea de protección de datos, como datos que describen a una persona identificable; los datos anonimizados y agregados no son considerados datos personales (Comisión Europea, 2012). El cuidado de la privacidad de las técnicas de minería de datos, las cuales no requieren acceso a registros individuales e identificables, podrían mitigar estos riesgos (Agrawal y Srikant, 2000; Fule y Roddick, 2004). Otros sugieren mecanismos para optar por no ser parte de la perfilización para

19 Los sujetos cuyos datos están implicados pueden ser considerados como poseedores de derecho a la identidad. Este derecho puede tomar muchas formas, pero la existencia de algunos derechos de identidad es difícil de disputar. Floridi (2011) concibe la identidad personal como aquella constituida por información. Tomado como tal, cualquier derecho a privacidad informacional se traslada a un derecho a la identidad por *default*, entendido como el derecho a manejar la información sobre uno mismo que constituye su propia identidad. De la misma manera, Hildebryt y Koops (2010) reconocen el derecho a formarse una identidad sin la influencia irracional externa. Ambos abordajes pueden ser conectados con el derecho a la personalidad derivado de la Convención Europea de Derechos Humanos.

un propósito en particular o contexto que ayude a proteger los intereses de privacidad de los sujetos y sus datos (Hildebryt, 2011; Rubel y Jones, 2014). La falta de recursos y mecanismos para que los sujetos cuestionen la validez de las decisiones algorítmicas exacerba aún más los desafíos de controlar la identidad y los datos sobre uno mismo (Schermer, 2011). Como respuesta, Hildebryt y Koops (2010) llaman a una «transparencia inteligente» a través del diseño de infraestructuras socio-técnicas responsables de la perfilización de manera que permitan a los individuos anticiparse y responder a la manera en que son perfilizados.

Trazabilidad que conduce a responsabilidad moral

Cuando una tecnología falla, la culpa y las sanciones deben ser repartidas. Uno o más de un diseñador (o desarrollador) de la tecnología, productor o usuario son típicamente responsabilizados. Típicamente se culpa a los diseñadores y usuarios de algoritmos de los problemas que surgen (Kraemer et al., 2011, p. 251). La culpa puede ser atribuible a que el actor tiene algún grado de control (Matthias, 2014) e intencionalidad en llevar adelante esa acción.

Tradicionalmente, los programadores han tenido «control sobre el comportamiento de la máquina en cada detalle» siempre y cuando puedan explicar su diseño y función a un tercer actor (Matthias, 2004). Esta concepción tradicional de responsabilidad en el diseño de *software* asume que el programador puede reflexionar sobre los probables efectos y el potencial mal funcionamiento de la tecnología (Floridi et al., 2014) y hacer elecciones de diseño para elegir el resultado más deseable de acuerdo con la funcionalidad específica (Matthias, 2004). Aun así, los programadores podrían retener el control solo en principio, debido a la complejidad y volumen del código (Syvig et al., 2014), y el uso de bibliotecas externas usualmente definidas por los programadores como «cajas negras» (ver nota 7).

En la superficie, la tradicional concepción de responsabilidad encaja para los algoritmos que no conllevan *machine-learning*. Cuando las reglas sobre procesos de decisión son escritas «a mano», sus autores conservan responsabilidad (Bozdag, 2013). Las reglas de los procesos de decisión determinan el peso relativo otorgado a las variables o dimensiones de los datos considerados por el algoritmo. Un ejemplo popular es el algoritmo de personalización *EdgeRank* de *Facebook*, que prioriza el contenido basado en la fecha de publicación, la frecuencia de interacción entre el autor y el lector, el tipo de contenido y otras dimensiones. Alterar la importancia relativa de cada factor cambia la relación que se estimula a tener a los usuarios. El grupo que programa los intervalos de confianza para un algoritmo con estructura de proceso decisional comparte responsabilidad por los efectos resultantes en falsos positivos, falsos negativos y correlaciones espurias (Birrer, 2005; Johnson, 2013, Kraemer et al., 2011). Fule y Roddick (2004, p. 159) sugieren que los operadores también tienen responsabilidad en monitorear el impacto ético de los procesos de toma de decisión algorítmicos porque «la sensibilidad de una regla podría no ser aparente para el minero [...] la capacidad de dañar o de causar ofensa puede pasar inadvertida».

De manera similar, Schremer (2011) sugiere que los procesadores de datos deberían buscar activamente errores y sesgos en sus algoritmos y modelos. De todas maneras, la supervisión humana de sistemas complejos como mecanismos de medición de responsabilidad podría ser imposible debido a los desafíos relacionados con la transparencia mencionada anteriormente (ver sección «evidencia inescrutable que conduce a la opacidad»). Más aun, los humanos mantenidos «en el *loop*» de los procesos automatizados de

decisión podrían estar en desventaja a la hora de identificar problemas y tomar acciones correctivas (Elish, 2016).

Los algoritmos con capacidades de aprendizaje suponen retos particulares, los cuales desafían la concepción tradicional de la responsabilidad del diseñador. El modelo requiere que el sistema esté bien definido, sea comprensible y predecible; los sistemas complejos y fluidos (por ejemplo, uno con incontables reglas de procesos de decisión y líneas de código) inhibe la observancia holística de las vías y dependencias de los procesos de toma de decisión. Los algoritmos de *machine learning* son particularmente desafiantes en este sentido (Burrell, 2916; Matthias, 20004; Zarsky, 2016), como, por ejemplo, los algoritmos genéticos que se programan a sí mismos. El modelo de responsabilidad tradicional falla porque «nadie tiene suficiente control sobre las acciones de la máquina para ser capaz de asumir la responsabilidad por ellos» (Matthias, 2004, p. 177).

Allen et al. (2006, p. 14) coinciden en discutir la necesidad de una «ética maquínica»: «el diseño modular de sistemas podría significar que ninguna persona o grupo puede comprender completamente los modos en los cuales el sistema interactúa o responde a un flujo complejo de nuevos *inputs*». Desde la programación lineal tradicional a través de algoritmos autónomos, el control comportamental es transferido gradualmente desde el programador hacia el algoritmo y su ambiente operacional (Matthias, 2004, p. 182). La distancia entre el control del diseñador y el comportamiento del algoritmo crea una *brecha de responsabilidad* (Cardona, 2008) allí donde la culpa puede ser potencialmente asignada a varios agentes morales simultáneamente.

Otros pasajes de la literatura al respecto apuntan a la «ética de la automatización», la aceptación en el reemplazo o el efecto de aumento en los procesos de toma de decisión humanos con algoritmos (Naik y Bhide, 2014). Morek (2006) encuentra problemático asumir que los algoritmos puedan reemplazar habilidades profesionales de forma que se plantee una semejanza. Los profesionales poseen sabiduría implícita y habilidades sutiles (Coeckelbergh, 2013; MacIntyre, 2007) que resultan difíciles de explicitarse y quizás imposibles de ser computadas (Morek, 2006). Cuando los tomadores de decisiones algorítmicos y humanos trabajan en tándem, las normas deben prescribir cuándo y de qué manera es requerida la intervención humana, particularmente en casos como el comercio de alta frecuencia, en donde la intervención en tiempo real es imposible antes de que ocurran daños (Davis et al., 2013; Raymond, 2014).

Los algoritmos que toman decisiones pueden ser considerados agentes culpables (Floridi y Syers, 2004a, Wiltshire, 2015). La posición moral y capacidad ética del proceso de toma de decisión de los algoritmos continúa siendo una cuestión destacable en la ética de las máquinas (Allen et al., 2006; Yerson, 2008; Floridi y Syers, 2004a). Las decisiones éticas requieren agentes que evalúen cuán deseables son distintos cursos de acción, lo cual presenta conflictos entre los intereses de los involucrados (Allen et al., 2006; Wiltshire, 2015).

Para algunos, los algoritmos de *machine learning* deberían ser considerados agentes morales con algún grado de responsabilidad moral. Los requerimientos para la agencia moral deberían diferir entre humanos y algoritmos; Floridi y Syers (2004b) y Sullins (2006) argumentan, por ejemplo, que «la agencia maquínica» requiere una autonomía significativa, un comportamiento interactivo y un rol en la responsabilidad causal para ser distinguible de la responsabilidad moral, la cual implica intencionalidad. Como se ha sugerido previamente, la agencia y la responsabilidad moral están vinculadas. Asignar agencia moral a agentes artificiales podría permitir que los humanos

involucrados le echen la culpa a los algoritmos (Crnkovic y Cürüklü, 2011). Negar agencia a agentes artificiales vuelve a los diseñadores responsables por los comportamientos poco éticos de sus creaciones semiautónomas; malas consecuencias reflejan un mal diseño (Yerson y Yerson, 2014; Kraemer et al., 2011; Turilli, 2007). Ningún extremo es completamente satisfactorio debido a la complejidad de la supervisión y la volatilidad de las estructuras de toma de decisión.

Más allá de la naturaleza de la agencia moral en máquinas, el trabajo en ética maquínica también investiga cómo diseñar un mejor razonamiento moral y comportamientos dentro de algoritmos autónomos considerados agentes moral y éticamente artificiales[20] (Yerson y Yerson, 2007; Crnkovic y Cürüklü, 2011; Sullins, 2006; Wiegel y Berg, 2009). La investigación sobre esta pregunta continúa siendo altamente relevante, ya que los algoritmos pueden ser requeridos para tomar decisiones en tiempo real involucrando «intercambios comerciales difíciles [...] los cuales podrían incluir consideraciones éticas complejas» sin la presencia de un operador (Wiegel y Berg, 2009, p. 234).

La automatización de tomas de decisión crea problemas de consistencia ética entre humanos y algoritmos. Turilli (2007) argumenta que los algoritmos deberían ser limitados «por el mismo conjunto de principios éticos» que los de los otrora trabajadores humanos para asegurar la consistencia al interior de los estándares éticos de una organización. De todas maneras, los principios éticos utilizados por decisores humanos podrían ser difíciles de definir y volver computables. La ética de las virtudes es también pensada para proveer conjuntos de reglas para estructuras de decisión algorítmicas, las cuales son fácilmente computables. Un modelo ideal de agentes morales artificiales basados en virtudes heroicas es el de Wiltshire (2015), en el que los algoritmos son entrenados para ser heroicos y, por ende, morales.[21]

Otros abordajes no requieren que los principios éticos sirvan de pilares de los marcos de procesos de toma de decisión algorítmicos. Bello y Bringsjord (2012) insisten en que el razonamiento moral en los algoritmos no debería estructurarse alrededor de los principios éticos clásicos porque no refleja la manera en que los humanos interactuamos realmente en los procesos morales de toma de decisión. Más bien, esas arquitecturas computacionales cognitivas —que permiten a las máquinas «leer mentes», o atribuir estados mentales a otros agentes— son necesarias. Yerson y Yerson (2007) sostienen que los algoritmos pueden diseñarse para imitar los procesos de toma de decisión éticos humanos, modelados en base a investigación empírica sobre cómo interactúan las instituciones, los principios y el razonamiento. Como mínimo, este debate revela que no existe aún una visión consensuada sobre cómo relocalizar de manera práctica los deberes éticos y sociales

20 Puede hacerse otra distinción entre agentes artificiales morales y agentes artificiales éticos. Los agentes artificiales morales no poseen verdadera «inteligencia artificial» o la capacidad de reflexión requerida para decidir y justificar un curso de acción ético. Los agentes artificiales éticos pueden calcular la mejor acción frente a dilemas éticos utilizando principios éticos (Moor, 2006) o marcos derivados de los mismos. En contraste, la moralidad artificial solo requiere que las máquinas actúen 'como si' fueran agentes morales, volviendo éticamente justificadas las decisiones de acuerdo con criterios predefinidos (Moor, 2006). La construcción de moralidad artificial es vista como el desafío alcanzable más inmediato e inminente para la ética maquínica, ya que no requiere inteligencia artificial (Allen et al., 2006). Dicho esto, la pregunta acerca de «si es posible crear agentes completamente éticos» continúa ocupando a los eticistas maquínicos (Tonkens, 2012, p. 139).

21 Tonkens (2012) argumenta que los agentes que utilizan marcos basados en la virtud encontrarán su creación inadmisible debido al sentido empobrecido de virtudes que una máquina puede desarrollar. En resumen, los desarrollos de carácter de humanos y máquinas son demasiado disímiles para ser comparados. Él predice que a menos que los agentes autónomos sean tratados como agentes completamente morales comparables a humanos, las injusticias sociales existentes se exacerbarán, puesto que las máquinas autónomas están privadas de libertad para expresar su autonomía al ser obligadas a servir a las necesidades de su diseñador. Esta preocupación apunta a un problema aún mayor en la ética maquínica vinculado a si los algoritmos y máquinas con autonomía de procesos de toma de decisión continuarán siendo tratadas como herramientas pasivas en oposición a agentes (morales) activos (Wiegel y Berg, 2009).

desplazados por la automatización (Shackelford y Raymond, 2014).

Más allá del diseño filosófico elegido, Friedman y Nissenbaum (1996) sostienen que los desarrolladores tienen la responsabilidad de diseñar para contextos diversos guiados por distintos marcos morales. Siguiendo esta idea, Turilli (2007) propone el desarrollo colaborativo de requerimientos éticos para sistemas computacionales para fundar un protocolo ético operacional. Así, podría confirmarse una consistencia entre el protocolo (estructura del proceso de toma de decisión) y los principios éticos explícitos del diseñador o la organización (Turilli y Floridi, 2009).

Puntos de futuras investigaciones

Tal como demuestra la discusión previa, el mapa propuesto (ver la sección «mapa de la ética de los algoritmos») puede ser utilizado para organizar el discurso académico actual en torno a las preocupaciones éticas sobre algoritmos, sobre fundaciones epistémicas y éticas puras. Tomando prestado un concepto del desarrollo de *software*, el mapa podría quedar perpetuamente «en *beta*». A medida que se identifiquen nuevos tipos de preocupaciones éticas respecto de los algoritmos, o si uno de los seis descritos puede ser dividido en dos o más tipos, el mapa es susceptible de revisión. Nuestra intención ha sido describir el estado del discurso académico sobre la ética de los algoritmos, y proponer una herramienta organizativa para futuros trabajos en el campo que salden brechas lingüísticas y disciplinarias. Esperamos que el mapa mejore la precisión de la forma en que las preocupaciones éticas se describen a futuro y que funcione como un recordatorio de las limitaciones de soluciones meramente metodológicas, técnicas o sociales a los desafíos que presentan los algoritmos. Como lo indica el mapa, las preocupaciones éticas son multidimensionales y por lo tanto requieren soluciones multidimensionales.

Si bien el mapa provee la estructura conceptual esencial que necesitamos, aún debe ser completado a medida que proliferen los estudios críticos sobre algoritmos. Los siete temas identificados en las secciones precedentes identifican el lugar que tiene la «ética algorítmica» en el mapa. Con esto en mente, en esta sección recorremos un número de tópicos que no reciben atención sustancial en la literatura vinculada a los efectos transformadores y la trazabilidad de los algoritmos. Estos tópicos podrían considerarse como direcciones futuras de investigación para la ética de los algoritmos a medida que se expanda y madure el campo.

En relación con los efectos transformadores, los algoritmos cambian la forma en que se construye, se maneja y se protege la identidad y los mecanismos de protección de datos (ver sección «Efectos transformadores que generan desafíos a la privacidad informacional»). La privacidad informacional y la identificabilidad están fuertemente vinculadas; un individuo posee privacidad en la medida en que posee control sobre los datos y la información sobre él. Por su parte, los algoritmos transforman la privacidad representando la identificabilidad como algo menos importante, y de allí que sea necesaria una teoría sobre la privacidad que responda a la poca importancia que se le da a la identificación y la individualidad.

Van Wel y Royakkers (2004) instan a una reconceptualización de los datos personales en la que la se otorguen protecciones equivalentes de privacidad a las «características grupales» en reemplazo de las «características individuales» a la hora de generar conocimiento sobre un individuo o de realizar acciones sobre él. Se necesitará mucho trabajo a futuro para describir de qué manera opera la privacidad a nivel grupal, en ausencia de identificabilidad (Mittelstadt y Floridi, 2016; Taylor et al., 2017). Se requieren mecanismos del mundo real para reforzar la privacidad en analíticas. Una propuesta mencionada en

la literatura revisada es el desarrollo de técnicas de minería de datos que preserven la privacidad y que no requieran acceso a registros individuales e identificables (Agrawal y Srikant, 2000; Fule y Roddick, 2004). Ya se está realizando investigaciones al respecto para detectar discriminación en la minería de datos (Barocas, 2014; Calders y Vewer, 2010; Hajian et al., 2012), aunque limitadas a la detección de discriminación *ilegal*. Harán falta más mecanismos de detección de daños producidos por algoritmos a usuarios en desventaja, en formas indirectas y poco obvias que exceden las definiciones legales de lo que es la discriminación (Syvig et al., 2014; Tufekci, 2015). No puede asumirse que los algoritmos discriminan de acuerdo con las características observables o comprensibles para los humanos.

Con respecto a la trazabilidad, existen dos desafíos clave para el reparto de responsabilidades hacia los algoritmos. Primero, a pesar de la fructífera producción literaria ocupada en la agencia y responsabilidad moral de los algoritmos, no se ha prestado suficiente atención a la responsabilidad *distribuida,* o responsabilidad compartida a través de una red de actores humanos y algorítmicos simultáneamente (Simon, 2015). La literatura revisada (ver «trazabilidad que conduce a responsabilidad moral») trata la potencial agencia moral de los algoritmos, pero no describe los métodos y principios para distribuir culpas o responsabilidades a través de una red mixta de actores humanos y algorítmicos.

Segundo, existe una confianza sustancial en los algoritmos, en algunos casos generando la *des-responsabilización* de los actores humanos, o una tendencia a «ocultarse detrás de la computadora» asumiendo que los procesos *automatizados* son correctos por *default* (Zarzky, 2016, p. 121). Delegar la toma de decisión a los algoritmos puede desviar la responsabilidad de los decisores humanos. Pueden observarse efectos similares en redes mixtas de humanos y sistemas informacionales como ya ha sido estudiado en burocracias, caracterizados por la reducción en los sentimientos de responsabilidad personal y la ejecución de otrora injustificables acciones (Arendt, 1997). Los algoritmos que involucran interesados de múltiples disciplinas pueden, por caso, hacer que cada parte asuma que los otros cargarán con las responsabilidades éticas generadas por las acciones de los algoritmos (Davis et al., 2013). El *machine learning* añade otra capa de complejidad entre los diseñadores y las acciones manejadas por el algoritmo, lo cual podría debilitar la culpa otorgada sobre los primeros. Se necesita investigación adicional para comprender la prevalencia de estos efectos en los sistemas de toma de decisión algorítmicos y para discernir cómo minimizar la inadvertida justificación de actos perjudiciales.

El mal funcionamiento y la resiliencia son dos problemas relacionados. La necesidad de distribuir la responsabilidad se siente agudamente cuando los algoritmos funcionan mal. Los algoritmos poco éticos podrían ser pensados como artefactos de *software* que funcionan mal o que no operan como se preveía. Existe una distinción útil entre errores de diseño (tipos) y errores de operación (*tokens*), y entre la imposibilidad de operar como se prevé (disfunción) y la presencia de efectos colaterales no intencionados (mal funcionamiento) (Floridi et al., 2014). El mal funcionamiento se distingue del mero efecto colateral negativo a través de la *evitabilidad,* o hasta dónde el tipo de algoritmo existente cumple con la función prevista sin los efectos en cuestión. Estas distinciones aclaran aspectos éticos de los algoritmos que están estrictamente relacionados con su funcionamiento, tanto en abstracto (por ejemplo, cuando observamos una performance en bruto) o como parte de un sistema mayor de tomas de decisión, revelando la interacción

multifacética entre el comportamiento pretendido y el generado.

Ambos tipos de mal funcionamiento implican distintas responsabilidades para los algoritmos y los desarrolladores de *software*, usuarios y artefactos. Es necesario trabajar para diferenciar de manera justa la responsabilidad por la disfunción y el mal funcionamiento al interior de un equipo de desarrolladores y en complejos contextos de uso. También sería necesario el trabajo orientado a especificar los requerimientos para la resistencia al mal funcionamiento como un ideal ético en el diseño de algoritmos. El *machine learning* en particular acumula desafíos únicos, ya que alcanzar un comportamiento esperable o «correcto» no implica la ausencia de errores[22] (Burrell, 2016) o acciones nocivas y *loops* de retroalimentación. Los algoritmos, particularmente aquellos involucrados en robótica, podrían, por ejemplo, volverse interrumpibles por seguridad al mismo tiempo que pueden ser desalentadas acciones nocivas sin que al algoritmo se le inste a engañar al usuario humano para evitar interrupciones (Orseau y Armstrong, 2016).

Para finalizar, si bien se reconoce que cierto grado de transparencia es requisito para la trazabilidad, cómo operacionalizar esa transparencia sigue siendo una pregunta abierta, particularmente para el *machine learning*. La mera modelación transparente del código de un algoritmo es insuficiente para asegurar el comportamiento ético. Los requerimientos regulatorios o metodológicos para que los algoritmos sean *explicables* o *interpretables* demuestran los desafíos que los controladores de datos enfrentan (Tutt, 2016). Un camino posible para la explicabilidad es la auditoría algorítmica llevada adelante por procesadores de datos (Zarsky, 2016), reguladores externos (Pasquale, 2015; Tutt, 2016; Zarsky, 2016), o investigadores empíricos (Kitchin, 2016, Neyly, 2016), usando estudios de auditorías *ex post* (Adler et al., 2016; Diakopoulos, 2015; Kitchin, 2016; Romei y Ruggeri, 2014; Syvig et al., 2014), estudios etnográficos en desarrollo y testeo (Neyly, 2016), o mecanismos de reporte diseñados al interior del algoritmo (Vellido et al., 2012). Para todo tipo de algoritmos, la auditoría es una precondición necesaria de cara a *verificar* el correcto funcionamiento. Para algoritmos de analítica con un previsible impacto humano, auditar podría crear un registro de procedimientos *ex post* de decisiones complejas para desentrañar decisiones problemáticas o inexactas, o para detectar discriminación o daños similares. Se necesita un mayor trabajo en el diseño de mecanismos de auditoría para algoritmos que sean ampliamente aplicables y de bajo impacto (Adler et al., 2016; Syvig et al., 2014) y que se basen en lo ya trabajado en transparencia e interpretabilidad del *machine learning* (Kim et al., 2015; Lous et al., 2013).

Todos los desafíos destacados en esta revisión son susceptibles de ser trabajados. Como con cualquier artefacto tecnológico, para obtener soluciones prácticas se requerirá la cooperación entre investigadores, desarrolladores y quienes diseñan políticas. Un área que también requiere más desarrollo es la transformación de políticas existentes y futuras aplicables a algoritmos en mecanismos y estándares regulatorios realistas. El Reglamento General de Protección de Datos (RGPD) de la Unión Europea en particular es indicativo de los desafíos a ser enfrentados globalmente en la regulación de algoritmos.[23] El RGPD estipula un número de responsabilidades para los controladores de datos y los derechos de los sujetos relevantes para

22 Excepto para casos triviales, la presencia de falsos positivos y falsos negativos en el trabajo de los algoritmos, particularmente en el de *machine learning*, es inevitable.

23 Es importante notar que esta regulación se aplica incluso en controladores de datos o procesadores que no están establecidos al interior de la UE, allí donde el monitoreo (incluida la predicción y perfilización) de comportamientos se enfoque en sujetos que estén localizados en la

los algoritmos de toma de decisión. Sobre el primero, al emprender una perfilización, los controladores serán obligados a evaluar las potenciales consecuencias de sus actividades de procesamiento de datos vía un asesoramiento en impacto de la protección de datos (art. 35 (3)(a)). Además de evaluar los riesgos de privacidad, los controladores de los datos también deberán comunicar estos riesgos a las personas involucradas. De acuerdo con los art. 13 (2) (f) y 15 (2) (g), los controladores de datos están obligados a informar a los sujetos sobre los métodos de perfilización existentes, su *significado* y sus *consecuencias previstas.* El art. 12 (1) obliga a que se utilice *lenguaje claro y llano* para informar sobre estos riesgos.[24] Sumado a esto, el Considerando 71 estipula que los controladores de datos tienen la obligación de explicar la lógica de la manera en que se tomó una decisión. Finalmente, el art. 22 (3) estipula que el deber de los controladores es «implementar medidas acordes para salvaguardar los derechos, libertades e intereses legítimos del sujeto interesado en estos datos» cada vez que se aplique una toma de decisión automatizada. Esta obligación, sin embargo, es bastante vaga y opaca.

Acerca de los derechos de los sujetos implicados, el RGPD generalmente toma un enfoque de autodeterminación. A los sujetos implicados se les garantiza el derecho a objetar los métodos de perfilización (art. 21) y el derecho a no ser sujeto exclusivo de procesos individuales automatizados de toma de decisión[25] (art. 22). En estos casos y otros similares, la persona involucrada tiene derecho a objetar los métodos utilizados o debería al menos tener derecho a «obtener intervención humana» para poder expresar sus puntos de vista y «recusar la decisión» (art. 22[3]).

A primera vista estas disposiciones trasladan el control al sujeto implicado y le permiten decidir la forma en que sus datos son utilizados. A pesar de que el RGPD muestra gran potencial de mejorar la protección de datos, un número no menor de excepciones limita los derechos de los sujetos.[26] El RGPD puede ser un mecanismo inoperante o poderoso para proteger a los sujetos, dependiendo de su eventual interpretación legal: el fraseo de la regulación permite que ambas condiciones sean verdaderas. Las autoridades supervisoras y sus futuros juicios determinarán la efectividad del nuevo marco regulatorio.[27] De todas formas, es necesario más esfuerzo en paralelo para proveer guías normativas y mecanismos prácticos que pongan en vigencia nuevos derechos y responsabilidades.

Estas no son tareas regulatorias mundanas. Por ejemplo, las disposiciones plantea-

UE (art. 3(2)(b) y Considerando 24).

24 En casos en los que el consentimiento informado sea requerido, el art. 7 (2) estipula que el no cumplimiento del art. 12 (1) hace que el consentimiento dado no sea legalmente vinculante.

25 El considerando 71 explica que las tomas de decisión automatizadas e individuales deben entenderse como un método «que produce efectos legales en él o ella o afecta significativamente a él o ella, como los rechazos automáticos de una aplicación a un crésssdito online o prácticas de búsquedas laborales electrónicas sin intervención humana» e incluye perfilización que permite «predecir aspectos relacionados con la performance laboral del sujeto implicado, situación económica, salud, preferencias o intereses personales, comportamiento o confiabilidad, locación o movimientos».

26 El art. 21(1) explica que el derecho a oponerse a los métodos de elaboración de perfiles puede limitarse «si el responsable del tratamiento demuestra que existen motivos legítimos imperiosos para el tratamiento que prevalecen sobre los intereses, los derechos y las libertades del interesado o para el reconocimiento, el ejercicio o la defensa de reclamaciones judiciales». Además, el art. 23(1) estipula que los derechos consagrados en los arts. 12 a 22 —incluido el derecho a oponerse a la toma de decisiones automatizada— pueden restringirse en casos como «la seguridad nacional, la defensa; la seguridad pública; [...]; otros objetivos importantes de interés público general de la Unión o de un Estado miembro, en particular un interés económico o financiero importante de la Unión o de un Estado miembro, incluidos los asuntos monetarios, presupuestarios y fiscales, la salud pública y la seguridad social; [...]; la prevención, la investigación, la detección y el enjuiciamiento de las infracciones deontológicas de las profesiones reguladas; [...]». En consecuencia, estas excepciones también se aplican al derecho de acceso (art. 15. derecho a obtener información si se trata de datos personales), así como al derecho a ser olvidado (art. 17).

27 El art. 83 (5)(b) confiere a las autoridades de control la facultad de imponer multas de hasta el 4% del volumen de negocios anual total a nivel mundial en caso de que se hayan infringido los derechos de los interesados (artículos 12 a 22). Esta ventaja puede utilizarse para imponer el cumplimiento de las normas y mejorar la protección de los datos.

das con anterioridad podrían interpretarse de manera que signifiquen que las decisiones automatizadas deben ser *explicables* a los sujetos implicados. Dada la dependencia y conectividad de los algoritmos y los *datasets* en sistemas de información complejos, y su tendencia a errores y sesgos en los datos y los modelos a ser escondidos a lo largo del tiempo (ver sección «Evidencia errónea que conduce a sesgos»), la «*explicabilidad*»[28] podría ser particularmente disruptiva para las industrias de datos intensivos. En el futuro habrá que elaborar requisitos prácticos que establezcan un equilibrio adecuado entre los derechos de los interesados a ser informados sobre la lógica y las consecuencias de la elaboración de perfiles y la carga impuesta a los responsables del tratamiento. Alternativamente, puede ser necesario limitar la automatización o determinados métodos analíticos en contextos particulares para cumplir los requisitos de transparencia especificados en el RGPD (Tutt, 2016; Zarsky, 2016). Ya existen restricciones comparables en la Ley de Información Crediticia de Estados Unidos, que prohíbe de hecho el *machine learning* en la calificación crediticia, ya que los motivos de la denegación del crédito deben ponerse a disposición de los consumidores a demanda (Burrell, 2016).

Conclusiones

Los algoritmos median cada vez más la vida digital y la toma de decisiones. El trabajo descrito aquí ha hecho tres contribuciones para aclarar la importancia ética de esta mediación: (1) una revisión de la discusión existente sobre los aspectos éticos de los algoritmos; (2) un mapa prescriptivo para organizar la discusión; y (3) una evaluación crítica de la literatura para identificar las áreas que requieren más trabajo para desarrollar la ética de los algoritmos.

La revisión realizada aquí se limitó principalmente a la literatura que trata explícitamente de los algoritmos. En consecuencia, solo se abordaron brevemente trabajos relevantes realizados en campos relacionados, en áreas como la ética de la inteligencia artificial, los estudios de vigilancia, la ética de la informática y la ética de las máquinas.[29] Aunque lo ideal sería resumir los trabajos en todos los campos representados en la literatura revisada y, por tanto, en cualquier ámbito en el que se utilicen algoritmos, el alcance de tal ejercicio es prohibitivo. Por lo tanto, debemos aceptar que puede haber lagunas en el tratamiento de los temas discutidos solo en relación con tipos específicos de algoritmos, y no para los algoritmos en sí mismos. A pesar de esta limitación, el mapa es deliberadamente amplio e iterativo para organizar el debate en torno a la ética de los algoritmos, tanto del pasado como del futuro.

El debate sobre un concepto tan complejo como el de «algoritmo» se enfrenta inevitablemente a problemas de abstracción o de «hablar por hablar» debido a que no se especifica un nivel de abstracción para el debate y, por lo tanto, se limita el conjunto pertinente de observables (Floridi, 2008). Todavía no existe una «ética de los algoritmos» madura, en parte porque el concepto «algoritmo» describe una gama prohibitiva de *software* y sistemas de información. A pesar de esta limitación, de la literatura

28 Se prefiere «explicabilidad» a «interpretabilidad» para destacar que la explicación de una decisión debe ser comprensible no solo para los científicos de datos o los responsables del tratamiento, sino para los sujetos implicados (o algún representante) afectados por la decisión.

29 Los diversos ámbitos de investigación y desarrollo descritos comparten una característica común: todos hacen uso de algoritmos informáticos. Esto no quiere decir que campos complejos como la ética de las máquinas y los estudios de vigilancia queden subsumidos en la etiqueta «ética de los algoritmos». Más bien, cada ámbito tiene cuestiones que no se originan en el diseño y la funcionalidad de los algoritmos que se utilizan. Por lo tanto, estas no se consideran parte de una «ética de los algoritmos», a pesar de la inclusión en un mismo campo. La «ética de los algoritmos» no pretende, por tanto, sustituir a los campos de investigación existentes, sino más bien identificar las cuestiones compartidas por un número diverso de dominios que se derivan de los algoritmos informáticos que utilizan.

surgieron varios temas que indican cómo se puede discutir la ética de manera coherente cuando se centra en los algoritmos, independientemente del trabajo específico.

El mapeo de estos temas en el marco propuesto ha demostrado ser útil para distinguir entre los tipos de preocupaciones éticas generadas por los algoritmos, que a menudo se confunden en la literatura. Las distintas preocupaciones epistémicas y normativas se tratan a menudo como un grupo. Esto es comprensible, ya que las diferentes preocupaciones forman parte de una red de interdependencias. Algunas de estas interdependencias están presentes en la bibliografía que hemos revisado, como la conexión entre la parcialidad y la discriminación (ver secciones «Pruebas erróneas que conducen a la parcialidad» y «Resultados injustos que conducen a la discriminación») o el impacto de la opacidad en la atribución de responsabilidades (ver secciones 'Pruebas inescrutables que conducen a la opacidad' y «Trazabilidad que conduce a la responsabilidad moral»).

El mapa propuesto pone de manifiesto otras áreas, como el efecto polifacético de la presencia y ausencia de deficiencias epistémicas en las ramificaciones éticas de los algoritmos. Además, el mapa demuestra que la solución de los problemas en un nivel no aborda todos los tipos de preocupaciones; una decisión algorítmica perfectamente auditable, o que se basa en pruebas concluyentes, escrutables y bien fundadas, puede, sin embargo, causar efectos injustos y transformadores, sin formas obvias de rastrear la culpabilidad en la red de actores que contribuyen. Mejores métodos para producir pruebas para algunas acciones no tienen por qué descartar todas las formas de discriminación, por ejemplo, e incluso pueden utilizarse para discriminar más eficazmente. De hecho, hasta se pueden concebir situaciones en las que los algoritmos menos finos en su diseño pueden tener efectos menos objetables.

Y lo que es más importante, como ya se ha subrayado repetidamente, en principio no podemos evitar los residuos epistémicos y éticos. Normalmente se puede esperar que herramientas algorítmicas cada vez mejores descarten muchas deficiencias epistémicas obvias, e incluso nos ayuden a detectar problemas éticos bien conocidos (por ejemplo, la discriminación). Sin embargo, todo el espacio conceptual de los desafíos éticos que plantea el uso de algoritmos no puede reducirse a los problemas relacionados con deficiencias epistémicas y éticas fácilmente identificables. Con la ayuda del mapa trazado aquí, futuros trabajos deberían esforzarse por hacer explícitas las numerosas conexiones implícitas de los algoritmos en la ética y en otros ámbitos.

Referencias

Adler, P., Falk C. y Friedler, S. A. et al. (2016). Auditing black-box models by obscuring features. *arXiv:1602.07043* [*cs*, *stat*]. Available at: http://arxiv.org/abs/1602.07043 (accessed 5 March 2016).

Agrawal, R. y Srikant, R. (2000). Privacy-preserving data mining. ACM *Sigmod Record*, ACM, pp. 439-450. Available at http://dl.acm.org/citation.cfm?id=335438 (accessed 20 August 2015).

Allen, C., Wallach, W. y Smith, I. (2006). Why machine ethics? *Intelligent Systems*, IEEE, *21*(4). Available at http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1667947 (accessed 1 January 2006).

Ananny, M. (2016). Toward an ethics of algorithms convening, observation, probability, and timeliness. *Science, Technology & Human Values*, *41*(1), 93-117.

Anderson, M. y Anderson, S. L. (2007). Machine ethics: creating an ethical intelligent agent. *AI Magazine*, *28*(4), 15.

Anderson, M. y Anderson S. L. (2014). Toward ethical intelligent autonomous healthcare agents: a case-supported principle-based behavior paradigm. Available at: http://doc.gold.ac.uk/aisb50/AISB50-S17/AISB50-S17-Anderson-Paper.pdf (Accessed 24 August 2015).

Anderson, S. L. (2008). Asimov's 'Three laws of robotics' and machine metaethics. *AI and Society*, *22*(4), 477-493.

Applin, S. A. y Fischer M. D. (2015). New technologies and mixed-use convergence: how humans and algorithms are adapting to each other. *In 2015* IEEE *international symposium on technology and society* (ISTAS) (pp. 1-6). Dublin, Ireland: IEEE.

Arendt, H. (1971). *Eichmann in Jerusalem*: *a report on the banality of evil*. Viking Press.

Barnet, B. A. (2009). Idiomedia: the rise of personalized, aggregated content. *Continuum*, *23*(1), 93-99.

Barocas, S. (2014). Data mining and the discourse on discrimination. Available at https://dataethics.github.io/proceedings/DataMiningandtheDiscourseOnDiscrimination.pdf (Accessed 20 December 2015).

Barocas, S. y Selbst, A. D. (2015). *Big data's disparate impact*, SSRN Scholarly paper. Rochester, NY: Social Science Research Network. Available at http://papers.ssrn.com/abstract=2477899 (Accessed 16 October 2015).

Bello, P. y Bringsjord, S. (2012). On how to build a moral machine. *Topoi*, *32*(2), 251-266.

Birrer, F. A. J. (2005). Data mining to combat terrorism and the roots of privacy concerns. *Ethics and Information Technology*, *7*(4), 211-220.

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and Information Technology*, *15*(3), 209-227.

Brey, P. y Soraker, J. H. (2009). *Philosophy of computing and information technology*, Elsevier.

Burrell, J. (2016). How the machine 'thinks': understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 1-12.

Calders, T. y Verwer S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, *21*(2), 277-292.

Calders, T., Kamiran, F. y Pechenizkiy, M. (2009). Building classifiers with independency constraints. In *Data mining workshops*, *2009.* ICDMW*'09.* IEEE *International Conference on Miami*, USA, IEEE, pp. 13-18.

Cardona, B. (2008). 'Healthy ageing' policies and anti-ageing ideologies and practices: on the exercise of responsibility. *Medicine*, *Health Care and Philosophy*, *11*(4), 475-483.

Coeckelbergh, M. (2013). E-care as craftsmanship: virtuous work, skilled engagement, and information technology in health care. *Medicine*, *Health Care and Philosophy*, *16*(4), 807-816.

Cohen, I. G., Amarasingham, R., Shah, A. et al. (2014). The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Affairs 33*(7), 1139-1147.

Coll, S. (2013). Consumption as biopower: governing bodies with loyalty cards. *Journal of Consumer Culture*, *13*(3), 201-220.

Crawford, K. (2016). Can an algorithm be agonistic? Ten scenes from life in calculated publics. *Science*, *Technology & Human Values*, *41*(1), 77-92.

Crnkovic, G. D. y Çürüklü, B. (2011). Robots: ethical by design. *Ethics and Information Technology*, *14*(1), 61-71.

Danna, A. y Gandy, O. H. Jr. (2002). All that glitters is not gold: digging beneath the surface of data mining. *Journal of Business Ethics*, *40*(4), 373-386.

Datta, A., Sen, S. y Zick, Y. (2016). Algorithmic transparency via quantitative input influence. In *Proceedings of 37th* IEEE *symposium on security and privacy*, San José, USA. Available at http://www.ieee-security.org/TC/SP2016/papers/0824a598.pdf (accessed 30 june 2016).

Davis, M., Kumiega, A., Van Vliet, B. (2013). Ethics, finance, and automation: a preliminary survey of problems in high frequency trading. *Science and Engineering Ethics*, *19*(3), 851-874.

De Vries K. (2010). Identity, profiling algorithms and a world of ambient intelligence. *Ethics and Information Technology*, *12*(1), 71-85.

Diakopoulos, N. (2015). Algorithmic accountability: journalistic investigation of computational power structures. *Digital Journalism*, *3*(3), 398-415.

Diamond, G. A., Pollock, B. H. y Work J. W. (1987). Clinician decisions and computers. *Journal of the American College of Cardiology*, *9*(6), 1385-1396.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the* ACM, *55*(10), 78-87.

Dwork, C., Hardt, M., Pitassi. T. et al. (2011). Fairness through awareness. *arXiv:1104.3913* [*cs*]. Available at: http://arxiv.org/abs/1104.3913 (Accessed 15 February 2016).

Elish, M. C. (2016). Moral crumple zones: cautionary tales in human-robot interaction (WeRobot 2016). SSRN. Available at http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2757236 (accessed 30 June 2016).

European Commission. (2012). *Regulation of the European Parliament and of the Council on the Protection of Individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)*. European Commission. Available at http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf (Accessed 2 April 2013).

Feynman, R. (1974). 'Cargo cult science', by Richard Feynman. Available at http://neurotheory.columbia.edu/~ken/cargo_cult.html (Accessed 3 September 2015).

Floridi, L. (2008). The method of levels of abstraction. *Minds and Machines*, *18*(3), 303-329.

Floridi, L. (2011). The informational nature of personal identity. *Minds and Machines 21*(4), 549-566.

Floridi, L. (2012). Big data and their epistemological challenge. *Philosophy & Technology*, *25*(4), 435-437.

Floridi, L. (2014). *The fourth revolution*: *how the infosphere is reshaping human reality*. OUP.

Floridi, L. y Sanders, J. W. (2004a). On the morality of artificial agents. *Minds and Machines*, *14*(3). Available at http://dl.acm.org/citation.cfm?id=1011949.1011964 (accessed 1 august, 2004).

Floridi, L. y Sanders J. W. (2004b) On the morality of artificial agents. *Minds and Machines*, *14*(3). Available at http://dl.acm.org/citation.cfm?id=1011949.1011964 (accessed 1 August, 2004).

Floridi, L., Fresco, N., Primiero, G. (2014). On malfunctioning software. *Synthese 192*(4), 1199-1220.

Friedman, B., Nissenbaum, H. (1996). Bias in computer systems. ACM *Transactions on Information Systems* (TOIS), *14*(3), 330-347.

Fule, P. y Roddick, J. F. (2004). Detecting privacy and ethical sensitivity in data mining results. In *Proceedings of the 27th Australasian conference on computer science*, *26*, 159-166. Dunedin, New Zealand, Australian Computer Society, Inc., pp. Available at http://dl.acm.org/citation.cfm?id=979942 (Accessed 24 August 2015).

Gadamer, H. G. (2004). *Truth and method*. Continuum International Publishing Group.

Glenn, T. y Monteith, S. (2014). New measures of mental state and behavior based on data collected from sensors, smartphones, and the internet. *Current Psychiatry Reports*, *16*(12), 1-10.

Goldman, E. (2006). Search engine bias and the demise of search engine utopianism. *Yale Journal of Law & Technology*, *8*, 188-200.

Granka, L. A. (2010). The politics of search: a decade retrospective. *The Information Society*, *26*(5), 364-374.

Grindrod, P. (2014). *Mathematical underpinnings of analytics*: *theory and applications*. OUP.

Grodzinsky, F. S, Miller, K. W, Wolf, M. J. (2010). Developing artificial agents worthy of trust: 'Would you buy a used car from this artificial agent?'. *Ethics and Information Technology*, *13*(1), 17-27.

Hacking, I. (2006). *The emergence of probability*: *a philosophical study of early ideas about probability*, *induction and statistical inference*. Cambridge University Press.

Hajian, S. y Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. IEEE *Transactions on Knowledge and Data Engineering*, *25*(7), 1445-1459.

Hajian, S., Monreale, A., Pedreschi, D. et al. (2012) Injecting discrimination and privacy awareness into pattern discovery. In *Data mining workshops* (ICDMW)*, 2012* IEEE *12th international conference on* Brussels, Belgium, IEEE, pp. 360-369. Available at http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6406463 (Accessed 3 November 2015).

Hildebrandt, M. (2008). Defining profiling: a new type of knowledge? In Hildebrandt M, Gutwirth, S. (eds.) *Profiling the European Citizen*, the Netherlands (pp. 17-45). Springer. Available at http://link.springer.com/chapter/10.1007/978-1-4020-6914-7_2 (Accessed 14 May 2015).

Hildebrandt, M. (2011). Who needs stories if you can get the data? ISPs in the era of big number crunching. *Philosophy & Technology*, *24*(4), 371-390.

Hildebrandt, M. y Koops, B-J. (2010). The challenges of ambient law and legal protection in the profiling era. *The Modern Law Review*, *73*(3), 428-460.

Hill, R. K. (2015). What an algorithm is. *Philosophy & Technology*, 29(1), 35-59.

Illari, P. M. y Russo, F. (2014). *Causality*: *philosophical theory meets scientific practice*, Oxford University Press.

Introna, L. D. y Nissenbaum, H. (2000). Shaping the Web: why the politics of search engines matters. *The Information Society*, *16*(3), 169-185.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

James, G., Witten, D., Hastie, T. et al. (2013). *An introduction to statistical learning* (vol. 6). Springer.

Johnson, J. A. (2006). *Technology and pragmatism*: *from value neutrality to value criticality*, SSRN Scholarly Paper. Social Science Research Network. Available at http://papers.ssrn.com/abstract=2154654 (Accessed 24 August 2015).

Johnson, J. A. (2013). *Ethics of data mining and predictive analytics in higher education*, SSRN Scholarly Paper. Social Science Research Network. Available at http://papers.ssrn.com/abstract=2156058 (Accessed 22 July 2015).

Kamiran, F. y Calders, T. (2010). Classification with no discrimination by preferential sampling. In *Proceedings of the 19th machine learning conf. Belgium and the Netherlands*, Leuven, Belgium. Available at http://wwwis.win.tue.nl/~tcalders/pubs/benelearn2010 (Accessed 24 August 2015).

Kamishima, T., Akaho, S., Asoh, H. et al. (2012). Considerations on fairness-aware data mining. In IEEE *12th International Conference on Data Mining Workshops*, *Brussels*, *Belgium*. 378-385. Available at http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6406465 (Accessed 3 November 2015).

Kim, B., Patel, K, Rostamizadeh, A, et al. (2015). Scalable and interpretable data representation for high-dimensional, complex data. AAAI. 1763-1769.

Kim, H., Giacomin, J. y Macredie, R. (2014). A qualitative study of stakeholders' perspectives on the social network service environment. *International Journal of Human-Computer Interaction*, *30*(12), 965-976.

Kitchin, R. (2016). Thinking critically about and researching algorithms. *Information, Communication & Society*, *20*(1), 14-29.

Kornblith, H. (2001). *Epistemology*: *internalism and externalism*. Blackwell.

Kraemer, F., van Overveld, K. y Peterson, M. (2011). Is there an ethics of algorithms? *Ethics and Information Technology*, *13*(3), 251-260.

Lazer, D., Kennedy, R, King, G. et al. (2014). The parable of Google flu: traps in big data analysis. *Science*, *343*(6176), 1203-1205.

Leese, M. (2014). The new profiling: algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union. *Security Dialogue*, *45*(5), 494-511.

Levenson, J. L., Pettrey, L. (1994). Controversial decisions regarding treatment and DNR: an algorithmic guide for the uncertain in decision-making ethics (GUIDE). *American Journal of Critical Care*: *an official publication*, *American Association of Critical-Care Nurses*, *3*(2), 87-91.

Lewis, S. C., Westlund, O. (2015). Big data and journalism. *Digital Journalism*, *3*(3), 447-466.

Lomborg, S., Bechmann, A. (2014). Using APIs for data collection on social media. *Information Society*, *30*(4), 256-265.

Lou, Y., Caruana, R., Gehrke, J. et al. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th* ACM SIGKDD *international conference on knowledge discovery and data mining*. Chicago, USA, ACM, pp. 623-631.

Louch, M. O., Mainier, M. J. y Frketich D. D. (2010). An analysis of the ethics of data warehousing in the context of social networking applications and adolescents. In *2010* ISECON *Proceedings*, 27(1392). Nashville, USA.

Lupton, D. (2014). The commodification of patient opinion: the digital patient experience economy in the age of big data. *Sociology of Health & Illness*, *36*(6), 856-869.

MacIntyre, A. (2007). *After virtue*: *a study in moral theory* (3.ª ed.). Gerald Duckworth & Co Ltd. Revised edition.

Macnish, K. (2012). Unblinking eyes: the ethics of automating surveillance. *Ethics and Information Technology*, *14*(2), 151-167.

Mahajan, R. L., Reed J. y Ramakrishnan, N. et al. (2012). *Cultivating emerging and black swan technologies.* ASME *2012 International Mechanical Engineering Congress and Exposition*, *Houston*, USA. 549-557.

Markowetz, A., Błaszkiewicz, K., Montag, C. et al. (2014). Psycho-informatics: Big data shaping modern psychometrics. *Medical Hypotheses*, *82*(4), 405-411.

Matthias, A. (2004). The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175-183.

Mayer-Schönberger, V. y Cukier, K. (2013). *Big data*: *a revolution that will transform how we live*, *work and think*. John Murray.

Mazoué, J. G. (1990). Diagnosis without doctors. *Journal of Medicine and Philosophy*, *15*(6), 559-579.

Miller, B. y Record, I. (2013). Justified belief in a digital age: on the epistemic implications of secret internet technologies. *Episteme*, *10*(2), 117-134.

Mittelstadt, B. D. y Floridi, L. (2016). The ethics of big data: current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics*, *22*(2), 303-341.

Mohler, G. O., Short, M. B., Brantingham, P. J. et al. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, *106*(493), 100-108.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *Intelligent Systems,* IEEE, *21*(4). Available at http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1667948 (Accessed 1 January 2006).

Morek, R. (2006). Regulatory framework for online dispute resolution: a critical view. *The University of Toledo Law Review*, *38*, 163.

Naik, G., Bhide, S. S. (2014). Will the future of knowledge work automation transform personalized medicine? *Applied & Translational Genomics*, *Inaugural Issue*, *3*(3), 50-53.

Nakamura, L. (2013). *Cybertypes*: *race, ethnicity, and identity on the Internet*. Routledge.

Newell, S. y Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: a call for action on the long-term societal effects of 'datification'. *The Journal of Strategic Information Systems*, *24*(1), 3-14.

Neyland, D. (2016). Bearing accountable witness to the ethical algorithmic system. *Science, Technology & Human Values*, *41*(1), 50-76.

Orseau, L. y Armstrong, S. (2016). Safely interruptible agents. Available at http://intelligence.org/files/Interruptibility.pdf (Accessed 12 September 2016).

Pariser, E. (2011). *The filter bubble*: *what the Internet is hiding from you*. Viking.

Pasquale, F. (2015). *The black box society: the secret algorithms that control money and information*. Harvard University Press.

Patterson, M. E. y Williams, D. R. (2002). *Collecting and analyzing qualitative data: hermeneutic principles, methods and case examples. Advances in tourism Application Series*. Sagamore Publishing, Inc. Available at http://www.treesearch.fs.fed.us/pubs/29421 (Accessed 7 November 2012).

Portmess, L. y Tower, S. (2014). Data barns, ambient intelligence and cloud computing: the tacit epistemology and linguistic representation of big data. *Ethics and Information Technology*, *17*(1), 19.

Raymond, A. (2014). The dilemma of private justice systems: big data sources, the cloud and predictive analytics. *Northwestern Journal of International Law & Business, Forthcoming*. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2469291 (accessed 22 July 2015).

Romei, A. y Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, *29*(5), 582-638.

Rubel, A., Jones, K. M. L. (2014). *Student privacy in learning analytics*: *an information ethics perspective*. SSRN *Scholarly Paper*. Social Science Research Network. Available at http://papers.ssrn.com/abstract=2533704 (Accessed 22 July 2015).

Sametinger, J. (1997) *Software Engineering with Reusable Components*. Springer Science & Business Media.

Sandvig, C., Hamilton, K., Karahalios, K., et al. (2014). Auditing algorithms: research methods for detecting discrimination on internet platforms. *Data and Discrimination*: *Converting Critical Concerns into Productive Inquiry*. Available at http://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sandvig.pdf (Accessed 13 February 2016).

Schermer, B. W. (2011). The limits of privacy in automated profiling and data mining. *Computer Law & Security Review*, *27*(1), 45-52.

Shackelford, S. J. y Raymond, A. H. (2014). Building the virtual courthouse: ethical considerations for design, implementation, and regulation in the world of Odr. *Wisconsin Law Review*, 3, 615-657.

Shannon, C.E. y Weaver, W. (1998). *The Mathematical Theory of Communication*. University of Illinois Press.

Simon, J. (2010). The entanglement of trust and knowledge on the web. *Ethics and Information Technology*, *12*(4), 343-355.

Simon, J. (2015). Distributed epistemic responsibility in a hyperconnected era. In L. Floridi (ed.), *The onlife manifesto*. Springer International Publishing, pp. 145-159. Available at http://link.springer.com/chapter/10.1007/978-3-319-04093-6_17 (Accessed 17 June 2016).

Stark, M. y Fins, J. J. (2013). Engineering medical decisions. *Cambridge Quarterly of Healthcare Ethics*, *22*(4), 373-381.

Sullins, J. P. (2006). When is a robot a moral agent? Available at http://scholarworks.calstate.edu/xmlui/bitstream/handle/10211.1/427/Sullins%20Robots-Moral%20Agents.pdf?sequence=1 (Accessed 20 August 2015).

Sweeney, L. (2013). Discrimination in online ad delivery. *Queue*, *11*(3), 10:10 10:29.

Swiatek, M. S. (2012). Intending to err: the ethical challenge of lethal, autonomous systems. *Ethics and Information Technology*, *14*(4). Available at https://www.scopus.com/inward/record.url?eid=2-s2.0-84870680328&partnerID=40&md5=018033cfd83c46292370e160d4938ffa (Accessed 1 January 2012).

Taddeo, M. (2010). Modelling trust in artificial agents, a first step toward the analysis of e-trust. *Minds and Machines*, *20*(2), 243-257.

Taddeo, M. y Floridi, L. (2015). The debate on the moral responsibilities of online service providers. *Science and Engineering Ethics*. 1-29.

Taylor, L., Floridi, L., Van der Sloot, B. (2017). *Group privacy: new challenges of data technologies* (1st ed.). Springer.

Tene, O. y Polonetsky, J. (2013a). Big data for all: privacy and user control in the age of analytics. Available at http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/nwteintp11&section=20 (Accessed 2 October 2014).

Tene, O. y Polonetsky, J. (2013b). Big data for all: privacy and user control in the age of analytics. Available at http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/nwteintp11&section=20 (Accessed 2 October 2014).

Tonkens, R. (2012). Out of character: on the creation of virtuous machines. *Ethics and Information Technology*, *14*(2), 137-149.

Tufekci, Z. (2015). Algorithmic harms beyond Facebook and Google: emergent challenges of computational agency. *Journal on Telecommunications and High Technology Law*, *13*, 203.

Turilli, M. (2007). Ethical protocols design. *Ethics and Information Technology*, *9*(1), 49-62.

Turilli, M. y Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology*, *11*(2), 105-112.

Turner, R. (2016). The philosophy of computer science. Spring 2016. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Available at http://plato.stanford.edu/archives/spr2016/entries/computer-science/ (Accessed 21 June 2016).

Tutt, A. (2016). *An* FDA *for algorithms*. SSRN *Scholarly Paper*. Social Science Research Network Available at http://papers.ssrn.com/abstract=2747994 (Accessed 13 April 2016).

Valiant, L. G. (1984). A theory of the learnable. *Communications of the Journal of the* ACM, 27, 1134-1142.

Van den Hoven, J. y Rooksby, E. (2008). Distributive justice and the value of information: a (broadly) Rawlsian approach. In Van den Hoven, J. y Weckert, J. (eds.), *Information technology and moral philosophy* (pp. 376-396). Cambridge University Press.

Van Otterlo, M. (2013). A machine learning view on profiling. In M. Hildebrandt, De Vries, K. (eds.), *Privacy*, *due process and the computational turn-philosophers of law meet philosophers of technology* (pp. 41-64). Routledge.

Van Wel, L. y Royakkers, L. (2004). Ethical issues in web data mining. *Ethics and Information Technology*, *6*(2), 129-140.

Vasilevsky, N. A., Brush, M. H. y Paddock, H. et al. (2013). On the reproducibility of science: unique identification of research resources in the biomedical literature. *PeerJ*, *1*, e148.

Vellido, A., Martín-Guerrero, J. D. y Lisboa, P. J. (2012). Making machine learning models interpretable. In ESANN *2012 proceedings* (pp. 163-172), Bruges, Belgium.

Wiegel, V. y Van den Berg, J. (2009). Combining moral theory, modal logic and mas to create well-behaving artificial agents. *International Journal of Social Robotics 1*(3), 233-242.

Wiener, N. (1988). *The human use of human beings*: *cybernetics and society*. Da Capo Press.

Wiltshire, T. J. (2015). A prospective framework for the design of ideal artificial moral agents: insights from the science of heroism in humans. *Minds and Machines 25*(1), 57-71.

Zarsky, T. (2013). Transparent predictions. *University of Illinois Law Review* 2013(4). Available at http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2324240 (Accessed 17 June, 2016).

Zarsky, T. (2016). The trouble with algorithmic decisions an analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science*, *Technology & Human Values*, *41*(1), 118-132.