ISEGORÍA. Revista de Filosofía moral y política
N.º 68, enero-junio, 2023, e23
ISSN-L: 1130-2097 | eISSN: 1988-8376
https://doi.org/10.3989/isegoria.2023.68.23

ARTÍCULOS

Justicia algorítmica y autodeterminación deliberativa

Algorithmic fairness and deliberative self-determination

Daniel Innerarity
Ikerbasque Foundation for Science (UPV/EHU) / Chair Artificial Intelligence and Democracy (European University Institute of Florence)
dinner@ikerbasque.org
ORCID iD: https://orcid.org/0000-0003-4307-8468

Resumen: Si la democracia consiste en posibilitar que todas las personas tengan iguales posibilidades de influir en las decisiones que les afectan, las sociedades digitales tienen que interrogarse por el modo de conseguir que los nuevos entornos hagan factible esa igualdad. Las primeras dificultades son conceptuales: entender cómo se configura la interacción entre los humanos y los algoritmos, en qué consiste el aprendizaje de estos dispositivos y cuál es la naturaleza de sus sesgos. Inmediatamente después nos topamos con la cuestión ineludible de qué clase de igualdad estamos tratando de asegurar, teniendo en cuenta la diversidad de concepciones de la justicia que hay en nuestras sociedades. Si articular ese pluralismo no es un asunto que pueda resolverse con una técnica agregativa, sino que requiere compromisos políticos, entonces una concepción deliberativa de la democracia parece la más apta para conseguir esa igualdad a la que aspiran las sociedades democráticas.

Palabras clave: Justicia; algoritmos; sesgos; democracia deliberativa.

Cómo citar este artículo / Citation: Innerarity, Daniel (2023) "Justicia algorítmica y autodeterminación deliberativa". *Isegoría,* 68: e23. https://doi.org/10.3989/isegoria.2023.68.23

Abstract**:** If democracy is about enabling all people to have equal opportunities to influence the decisions that affect them, digital societies need to ask how to ensure that new environments make this equality feasible. The first challenges are conceptual: understanding how the interaction between humans and algorithms is configured, what the learning of these devices consists of, and the nature of their biases. Immediately afterwards, we come up against the unavoidable question of what kind of equality, we are trying to ensure, bearing in mind the diversity of conceptions of fairness in our societies. If articulating this pluralism is not a matter that can be resolved with an aggregative technique, but requires political compromises, then a deliberative conception of democracy seems the most apt to achieve the equality to which democratic societies aspire.

Keywords: Fairness; Algorithms; Bias; Deliberative democracy.

Recibido: 2 enero 2023. *Aceptado*: 16 abril 2023.

Para abordar la cuestión de la justicia algorítmica este artículo aborda, en primer, lugar la relación entre los algoritmos y los humanos, una relación que no puede entenderse a mi juicio sin haber comprendido la naturaleza de los sesgos algorítmicos. Ese será el segundo objetivo del trabajo. Desde este entramado tecnológico configurado por algoritmos y seres humanos, ¿es posible aspirar a un ideal normativo de equidad? Mi tercer objetivo es precisamente plantear todas las dificultades a las que se enfrenta una definición universalmente compartida de justicia. Finalmente, se plantea que la concepción deliberativa de la democracia ofrece un marco en el que debería acreditarse cualquier pretensión de justicia e igualdad.

1. UNOS ALGORITMOS FRENTE A OTROS

Tenemos una gran limitación a la hora de entender intuitivamente la equidad o discriminación de los algoritmos complejos con los que interactuamos. Comparadas con las formas tradicionales de discriminación, la discriminación automatizada es más abstracta y sutil (Mittelstadt *et al*., 2016; Zarsky, 2016). Nos lo dificulta nuestro sesgo acerca de los sesgos. El principal sesgo de los humanos es creer que solo los algoritmos tienen sesgos y que basta con poner más humanos en el *loop*, manualizar la automatización o moderar los contenidos para que ese sesgo algorítmico desaparezca. El otro sesgo es del sentido contrario: nuestra mayor disposición a aceptar discriminaciones cuando son atribuidas a los algoritmos y no a los humanos (Wang, 2018). En ambos casos exoneramos del error a uno de los dos elementos (humanos o máquinas), en lugar de pensar que hay errores de unos y de otros, sesgos algorítmicos y sesgos antropológicos, por lo que sería aconsejable disponer nuestros sistemas de decisión de modo que se realice la mejor sinergia posible. «No se trata de elegir entre algoritmos digitales y un ideal platónico. Se trata más bien de elegir entre algoritmos digitales y algoritmos humanos, cada uno con sus propias ventajas e inconvenientes» (Coglianese y Lai, 2022, p. 1287). Y a veces hay que elegir entre un algoritmo digital y uno humano, o lograr un compromiso entre ambos. Los algoritmos no son una realidad completamente independiente de los humanos que los crearon, por supuesto, pero cuando hablamos de lógica generativa o *machine learning* estamos refiriéndonos a un dispositivo que en una cierta medida se desarrolla fuera del control directo de quienes los diseñaron.

Analizaremos más adelante los sesgos algorítmicos, pero no podemos tener un cuadro completo de nuestros sistemas de decisión y sus posibles arbitrariedades si no comenzamos identificando bien los sesgos humanos, para cuya mitigación disponemos de sistemas automatizados (Jolls, Sunstein y Thaler, 1998). Conviene no perder de vista que los humanos no siempre salen victoriosos de la comparación con las máquinas cuando examinamos su objetividad o transparencia. En ocasiones es más fácil corregir sesgos algorítmicos que prejuicios humanos. El juicio humano presenta una serie de limitaciones y sesgos que están bien documentados. Estas limitaciones van desde las físicas (la desmemoria o el cansancio) hasta los prejuicios ideológicos individuales (el interés propio) o sociales (el pensamiento grupal y las disfuncionalidades colectivas) (Kahneman, 2011; Thaler, 2015; Lai, 2018). Como consecuencia de ello, en no pocos casos las decisiones humanas resultan más problemáticas que sus contrapartidas digitales.

Buena parte de nuestra mala aproximación a la cuestión de los sesgos procede de que tenemos una visión muy estática de esta relación entre los humanos y las máquinas. El objetivo no es mejorar los algoritmos, sino preguntarse cómo los algoritmos interaccionan con la sociedad en su conjunto incluyendo, como actualmente ocurre, sus desigualdades estructurales. Y la intervención humana en sistemas que aprenden no deja de ser problemática. Brian Christian (2020, p. 302) señala al menos dos problemas. El primero es que cuando los humanos intervenimos el sistema aprende y corrige la idea que tiene de nuestras preferencias. Si esta corrección reduce totalmente la incertidumbre, el sistema pierde todo incentivo para responder a nuestras interrupciones. El segundo problema es que el sistema asume la lógica de que «el cliente siempre tiene la razón», pero si comprueba que los humanos nos equivocamos puede terminar creyendo que sabe mejor que nosotros lo que nos conviene y empezar a hacer oídos sordos a nuestras propuestas. Como puede verse, esta interacción es todo menos problemática y debemos entenderla en toda su amplitud y dinamicidad para evitar en lo posible errores fatales, principalmente el que supondría desconocer la inevitabilidad del error y desaprovechar las posibilidades de aprendizaje que nos ofrece a unos y otros, a humanos y a sistemas.

2. LOS SESGOS ALGORÍTMICOS

El otro conjunto importante de condicionamientos en nuestra interacción con las máquinas inteligentes tiene que ver con la naturaleza misma de los sesgos algorítmicos. Hay un cierto pánico moral que parece no entender la lógica computacional, su carácter experimental y generativo, el inevitable

sesgo generativo del *machine learning*, que no es solo una potencial fuente de error sino condición de posibilidad de su resultado. Por supuesto que hay que combatir los errores y daños efectuados por los algoritmos, pero eso es algo que no puede hacerse fuera de la lógica como calculan, es decir, sobre la base de operaciones que, como la inferencia, la intuición, las apuestas, caracterizan a su modo de razonar. Los algoritmos son un modo de evitar la arbitrariedad del juicio humano, así como un procedimiento de gestionar lo desconocido e imprevisible. ¿Para qué los hemos inventado si no para gestionar unas decisiones que nos exigían procesar una información inabarcable, en un espacio de tiempo limitado y para corregir «algunos» de nuestros sesgos que afectan menos al funcionamiento de las máquinas (que tienen a su vez otros específicos que hemos de tratar de corregir)? «Con los algoritmos actuales, las decisiones se toman al límite de lo que puede ser conocido, pero no se asume ninguna responsabilidad por las consecuencias desconocidas de la decisión» (Amoore, 2020, p. 112). Es ya un lugar común criticar los errores de los algoritmos y proponer como solución limitar sus excesos a través de códigos éticos. Pero los algoritmos no pueden ser controlados estableciendo un umbral de maldad porque la esencia de su lógica consiste en establecer tal umbral, adaptarlo y modularlo a través del tiempo (Amoore, 2020, p. 110). Algunas de sus decisiones que podemos considerar equivocadas son parte integral de su naturaleza y de sus capacidades experimentales y generativas.

Que un algoritmo aprende significa que se ajusta a las características de su entorno, para lo cual necesita tener algunas suposiciones acerca de cómo está constituido el mundo. Pese a la generalizada pretensión de que los algoritmos estén libres de sesgos, no pueden funcionar sin ellos, como los humanos, y sin la capacidad de reajustar esos prejuicios. El problema no es tanto la existencia de sesgos como su inadvertencia e incapacidad de corrección. La discriminación y el sesgo no son un accidente. Cuando el software tiene una cierta complejidad es difícil que la programación evite los «bugs» (fallos), especialmente cuando hay interacciones sistémicas y tantas combinaciones posibles de eventos que no es razonable suponer que los diseñadores puedan tomarlas a todas en consideración. «El software se libera para su uso, no cuando se sabe que es correcto, sino cuando la tasa de descubrimiento de nuevos errores se reduce a un nivel que la dirección considera aceptable» (Parnas, 1985, p. 433). Se pueden descubrir nuevos *bugs*, pero nunca hacerlos imposibles. Eliminar el «último» fallo es una broma habitual entre los informáticos. ¿Por qué es tan difícil reparar todos los errores significativos en un programa complejo?

> El problema fundamental del mantenimiento de programas es que arreglar un defecto tiene una probabilidad sustancial (20-50 %) de introducir otro. Así que todo el proceso consiste en dar dos pasos hacia delante y uno hacia atrás... Todas las reparaciones tienden a destruir la estructura... Cada vez se dedica menos esfuerzo a reparar los defectos de diseño originales; cada vez se dedica más a reparar los defectos introducidos por reparaciones anteriores (Brooks, 1975, p. 121).

Incluso después de repetidas pruebas y correcciones de errores, es difícil confiar plenamente en que el software no contenga algún fallo de diseño crucial oculto que algún día aflore inesperadamente y haga caer el sistema. Estamos aquí en la famosa «ley de Brooks», según la cual añadir recursos a un proyecto retrasado lo hace demorarse más aún. Por eso en ocasiones, más que diseñar mecanismos que busquen la neutralidad es mejor reconocer que la neutralidad es inalcanzable, hacer explícitos los problemas de tales sistemas e intentar mitigar esos sesgos mediante, por ejemplo, reglas de diversidad en la configuración de los equipos que toman decisiones, «*blind reviews*» o limitar su uso a asuntos que no tengan grandes consecuencias.

3. JUSTICIA CONTROVERTIDA

Otra de las dificultades para configurar un entorno algorítmico equitativo procede de nuestra disparidad de concepciones acerca de la justicia. Una manifestación de que la discusión acerca de los algoritmos no puede tener sino una naturaleza política es el hecho de que, aunque estuviéramos de acuerdo acerca de la necesidad de la justicia algorítmica, habría que decidir qué clase de justicia tienen que implementar los algoritmos. Los juicios acerca de la justicia están siempre bajo contestación y los criterios para medirla en el *machine learning* implican «afirmaciones cargadas de valores sobre la finalidad de un sistema, los derechos de las personas y los criterios pertinentes para la toma de decisiones» (Green y Hu, 2018, p. 2). Que nuestras concepciones de la justicia tengan una pretensión de validez universalizable no quita que su implementación en leyes o medidas de gobierno tenga que llevarse a cabo mediante procedimientos políticos donde se realizan esos compromisos.

Quienes son partidarios de la igualdad tienen visiones diferentes de la igualdad. Incluso allá donde

hay un amplio acuerdo acerca de la conveniencia de promover la igualdad no necesariamente se comparte la misma idea de igualdad. Nuestras controversias acerca de la justicia generalmente no confrontan a quienes la defienden o la desprecian, sino que obedecen a que tenemos distintas concepciones de ella, más distributiva, inclusiva o procedimental; hay quien se da por satisfecho con que sean idénticas las condiciones de partida, mientras que otros la entienden como una similitud en los resultados; hay una igualdad formal y otra de contenido; hay quien defiende una igualdad abstracta y quien argumenta en favor de una igualdad que se conseguiría mediante ciertas discriminaciones positivas; su tensión con otros valores como la libertad explica que haya tantas versiones de ella en el espacio de debate democrático. Un sistema de cuotas refuerza la igualdad en los resultados; un sistema basado en los méritos curriculares anonimizados está orientada a asegurar la igualdad de oportunidades. ¿Consiste la igualdad en asegurar que todos tienen la misma probabilidad de obtener determinado beneficio o en minimizar los perjuicios de los más desaventajados? ¿Cuáles son los criterios de justicia más apropiados para un determinado contexto? ¿Qué motivos justificarían un tratamiento diferente? ¿Qué tipos de disparidad son aceptables y cuáles no? (Binns, 2018a y 2018b). ¿Perseguimos la justicia «a través de la ceguera» (Hardt, 2014) o con el mejor conocimiento disponible de las circunstancias personales? La discriminación puede referirse a un tratamiento dispar o a un impacto dispar, que son dos aspectos completamente distintos y que se abordan con estrategias diferentes y a menudo contradictorias: intentando corregir una disparidad puede incrementarse la otra.

Además de las diferentes concepciones, la cuestión de la igualdad está llena de dilemas y paradojas. Un dilema frecuente se debe al hecho de que la justicia puede implicar que las personas similares sean tratadas similarmente, lo cual entra a menudo en tensión con la idea de paridad entre los grupos (Dwork *et al*., 2012). Un ejemplo de ello lo tenemos en el debate acerca de si los sesgos del algoritmo de reincidencia COMPAS utilizado en la administración penitenciaria pueden atribuirse a distintas ideas de la igualdad, al trato desigual o al impacto desigual, a la igualdad formal o a la igualdad de acuerdo con los resultados. Para unos el algoritmo no está sesgado porque la tasa de reincidencia es aproximadamente la misma independientemente de la raza, mientras que otros sostienen que los negros tienen de hecho más probabilidades de ser clasificados como de riesgo medio o alto de reincidencia que los blancos; para los primeros, el algoritmo no es el causante de que haya de hecho más reincidencia en unos grupos raciales que en otros, según estos últimos el algoritmo estaría sesgado porque un grupo es sometido sistemáticamente a un tratamiento más severo debido a la predicción errónea del algoritmo (Dieterich, Mendoza y Brennan, 2016).

La igualdad es un concepto controvertido que difícilmente se resuelve por una simple equiparación. Un ejemplo de lo absurda que puede llegar a ser la cuantificación lo tenemos en el hecho de que, si nuestra obsesión fuera igualar el impacto que los sistemas de predicción de reincidencia tienen en hombres y mujeres, dado que los hombres presentan unas tasas mayores, las mujeres tendrían que permanecer más tiempo en prisión, puesto que suelen reincidir menos. En este caso y en general, la cuestión de la igualdad es un asunto que debe ser interpretado y en esa misma medida es susceptible de resultar controvertido.

Otra fuente de controversias procede de las agrupaciones que llevan a cabo los algoritmos a la hora de tomar determinadas decisiones. Aquí nos topamos con el dilema de que la justicia tiene que procurar la máxima personalización posible, pero también ha de realizar la agrupación necesaria. Las predicciones a la hora de tomar una decisión son inevitables y tienen una lógica de gestión de la complejidad. Pongamos el ejemplo de una contratación laboral. Si creyéramos de verdad que cada caso es completamente diferente, entonces no podríamos hacer otra cosa que observar cómo se comporta cada persona una vez contratada, sin poder predecir nada por su pertenencia a un determinado grupo. Entonces, ¿con base a qué lo contrataríamos? No habría más que juicios *ex post,* nada podría determinarse *ex ante.* Evidentemente se trata de un absurdo que invalida cualquier instrumento de predicción y su capacidad de hacer apuestas razonables acerca de un posible comportamiento futuro. Un algoritmo (o una decisión humana preparada de acuerdo con algún tipo de agrupamiento) especifica y por tanto restringe los elementos que hay que considerar. ¿Es toda restricción una forma de rendirse a la injusticia o solo aquella que permite reducir la complejidad de las decisiones?

La cuestión de los grupos es una de las más intrincadas cuando se habla de justicia en general y de justicia algorítmica en particular. ¿Cómo deben articularse categorías, grupos e individuos para hacer frente a las discriminaciones que proceden de la pertenencia a un determinado grupo y las que se deben a ser agrupado de esa manera? La aplicación mecánica de criterios contra la discriminación a los

algoritmos puede tener efectos perjudiciales sobre aquellos grupos a los que se pretende proteger. De entrada, existe el riesgo de evaluar únicamente los criterios de injusticia en la población a la que se aplica el modelo y pasar por alto la injusticia que resulta que el modelo dote de subjetividad a unos grupos y no a otros (Mitchell *et al*., 2021). Y tampoco podemos perder de vista lo que ha podido llamarse el «*portability trap*» es decir, lo engañoso, inexacto y perjudicial que puede ser la utilización de las soluciones algorítmicas diseñadas para un contexto social en un contexto diferente (Selbst *et al*., 2019, p. 61).

Cuando formulamos la justicia algorítmica por relación a grupos de población cuya discriminación se pretende corregir, hay que tener en cuenta al menos dos cosas: la posible injusticia en el interior de esos grupos y la llamada «interseccionalidad». En cuanto a lo primero, puede ocurrir que los principios de normalización que actúan dentro de cada grupo y que el diseñador del algoritmo da por sentados ignoren las desigualdades dentro de esos grupos (Kasy y Abebe, 2021). Y en cuanto a la justicia interseccional, se trata de tener en cuenta la complejidad de los grupos sociales que se toman en consideración (Hanna *et al*., 2020, p. 8). Con el calificativo de «interseccional» se alude a algo que va más allá de la mera adición, a una desventaja multiplicada. El *Black feminism*, por ejemplo, ha llamado la atención sobre la incapacidad de entender las diferentes opresiones que sufre un determinado grupo racial desde una visión simplista u descontextualizada de las razas. Las opresiones racistas y sexistas, así como la subordinación económica están entrelazadas en la vida de las mujeres negras en el seno de unas instituciones y leyes supuestamente «*color-blind*» (Collins, 2000; Crenshaw, 2019).

4. LA AGREGACIÓN IMPOSIBLE

Supongamos que estuviéramos de acuerdo en la concepción de lo equitativo y que solo habría que establecer un procedimiento para agregar nuestros diferentes intereses. ¿Sería esto posible? ¿Es razonable buscar un procedimiento técnico que determinara la resultante equilibrada de nuestras distintas preferencias?

Esta aspiración se encuentra de entrada con la dificultad de que no estamos ante un asunto que tenga una solución «técnica», si por tal entendemos algo que nos ahorre juicios de valor y desactive el carácter controvertido de cualquier decisión pública. La justicia no consiste en una eliminación técnica de los sesgos, sino que incluye un amplio análisis social sobre el modo como es usada la inteligencia artificial en un contexto dado, de manera que sea posible una mejor auditoría de los sesgos. Al igual que la justicia no es una propiedad de los algoritmos, sino más bien de las decisiones que contienen (en su diseño, análisis o aplicación) (Ochigame *et al*., 2018, p. 4), la discriminación no es solo una cuestión algorítmica. Todo lo que tiene que ver con la justicia y la discriminación es tan contextual y controvertido que no siempre se presta a formalismos matemáticos (Selbst *et al*., 2019).

La justicia algorítmica no puede ser una implementación que satisfaga ciertos «indicadores de igualdad» incontrovertidos. Qué idea de justicia y qué otros valores deben ser considerados en un algoritmo supone un desafío de naturaleza política, no simplemente técnica, que requiere acomodar diversos intereses en conflicto; es una tarea política que no puede ser realizada por unos técnicos o por unos algoritmos que no tuvieran necesidad de contar con la opinión ciudadana, es decir, que debe llevarse a cabo democráticamente, abriendo estas definiciones a la pública discusión.

Si hubiera acuerdo acerca de qué significa «justicia», entonces el algoritmo desarrollaría una tarea puramente técnica; se trataría nada más que de encontrar el mejor modo de operacionalizar esa idea de justicia. Pero hay ciertas decisiones difíciles acerca de cómo medir la justicia que hay que tomar antes de que comience el trabajo técnico de detectar y mitigar la injusticia. ¿Podemos determinar qué significa «exactitud» y cómo se mide sin hacer algún tipo de juicio ético-político sobre los tipos de errores que pensamos que es más urgente evitar o cuál es el objetivo final de una organización? Parece claro que se trata de asuntos que deben ser objeto de discusión política y no de agregación algorítmica. El problema es que la idea de justicia es un verdadero campo de batalla democrático, un concepto elevadamente controvertido en cualquier sociedad plural. Hay más desacuerdos acerca de los valores en sí mismos que sobre los medios de conseguirlos. Ningún dispositivo tecnológico puede ahorrarnos el trabajo de discusión democrática en torno a los fines, aunque pueda facilitarnos enormemente la tarea de implementación de los objetivos que democráticamente hemos decidido perseguir.

El carácter controvertido, político y no meramente técnico de la justicia plantea otro problema adicional. Además de la dificultad de ponerse de acuerdo en torno a una idea de justicia, está la imposibilidad de satisfacer igualmente y simultáneamente esa diversidad de aspiraciones de justicia. «Las limitaciones prácticas y sociales impedirán que todas las preferencias se satisfagan al máximo

simultáneamente, lo que significa que los robots deberán mediar entre preferencias conflictivas, algo con lo que los filósofos y científicos sociales han luchado durante milenios» (Russell, 2019, p. 32). Más difícil que identificar preferencias particulares es agregarlas y hacerlas compatibles (Züger, Milan y Tanczer, 2017). Imaginemos que la tecnología nos ha permitido identificar todos los deseos, preferencias y decisiones individuales, ¿habríamos hecho innecesario cualquier elemento de mediación para la configuración de la voluntad popular? ¿Nos bastaría agregar sin deliberación las decisiones así registradas? Estaríamos así ante una variante digital del llamado «*impossibility theorem*» de Arrow (1950) por el que se declaraba como algo imposible satisfacer valores distintos (Friedler, Scheidegger y Venkatasubramanian, 2016; Berk *et al*., 2017; Miconi, 2017). Se formula así la idea de que es matemáticamente imposible que un algoritmo satisfaga simultáneamente las diversas ideas de justicia que sostenemos. No es posible recoger las diferentes preocupaciones de justicia que tenemos en una sociedad plural, ni resulta verosímil que lleguemos a un entendimiento pleno acerca de ese valor. Además, el valor de la justicia está relacionado con otros valores como la seguridad o la libertad, por lo que su formalización técnica resulta todavía más inverosímil.

Que una parte de nuestros desacuerdos sea irresoluble técnicamente y tenga una naturaleza política no es necesariamente una mala noticia. Se trata de una imposibilidad que nos obliga a explorar un modelo de decisión que tal vez tenga una gran fuerza democratizadora. El carácter controvertible de ciertos asuntos, su ambigüedad, tiene un valor político en la medida en que obliga a negociar y buscar compromisos una vez que los procedimientos de técnica algorítmica nos han dejado tirados (Coyle y Weller, 2020). Se podría hablar incluso de una cierta incompatibilidad entre la lógica de los algoritmos y la de la política. El *machine learning* optimiza la consecución de objetivos una vez que estos han sido explícitamente formulados. La política, por el contrario, se basa en una cierta ambigüedad en relación con los objetivos, gracias a la cual hay un espacio para lograr compromisos. La política es con mucha frecuencia un *trade-off* entre preferencias e intereses distintos e incluso contrapuestos. Los algoritmos son optimizadores de una determinada decisión, pero no toleran la ambigüedad. Por eso la justicia algorítmica no vendrá de que mejoremos los datos o desprejuiciemos los algoritmos sino de que sustituyamos un procedimiento de agregación de intereses y preferencias por uno de deliberación.

5. LA AUTODETERMINACIÓN DELIBERATIVA

La concepción deliberativa de la democracia parte del supuesto de que, si bien es cierto que la política está para satisfacer los intereses de las personas, esos intereses no se determinan con independencia de la reflexión sobre ellos y su compatibilidad con los de los demás. La democracia no es tanto que se tenga en cuenta nuestra opinión o se satisfaga nuestro interés como que dispongamos de un espacio público en el que configurar nuestra opinión e identificar nuestros intereses teniendo en cuenta los de los demás. Ni el interés individual está plenamente fijado, ni el interés colectivo están dados de antemano o pueden confiarse a una mera agregación de los intereses individuales; ambos tienen una dimensión de construcción pública. Hace falta establecer un marco de diálogo y negociación que permita la construcción equitativa de esa voluntad general. El problema de la gobernanza algorítmica es que registra nuestros intereses, pero no los convierte en objeto de reflexión.

Los automatismos son procesos que funcionan precisamente porque no obligan a tematizar los presupuestos sobre los que discurren. Los seres humanos, tanto en el plano personal como en el colectivo, realizamos tareas mecánicas y vivimos sin cuestionar las prioridades que una vez establecimos, pero hemos de estar abiertos a otro tipo de situaciones en las que se requiere de nosotros un cuestionamiento de las rutinas y una reorientación hacia objetivos nuevos. Una de las revitalizaciones de la democracia a finales del siglo pasado vino precisamente del concepto de «democracia reflexiva» (Beck, Lash y Giddens, 1994) con el que diversos pensadores defendían la interrupción crítica de una modernización irreflexiva. De este modo no hacían otra cosa que acentuar una propiedad de la política como actividad que se interroga por los fines frente a las rutinas administrativas. Pues bien, la gobernanza algorítmica carece por sí misma de la capacidad de cuestionarse sus objetivos o lo hace —en virtud de procesos de aprendizaje— dentro de un marco que no se ha dado a sí misma y que por ello no puede cuestionar radicalmente. La repolitización a la que debe estar abierto un sistema democrático puede estar impulsada por un algoritmo, pero no se realiza algorítmicamente.

La reflexividad es lo que hace posible la deliberación democrática, es decir, aquella forma de interacción que no es solo una negociación de nuestras preferencias e intereses, sino que permite incluso su revisión y ponderación reflexiva. El sentido de las instituciones de la mediación en una democracia consiste en establecer una distancia entre la voluntad

inmediata y la decisión política. El procedimiento para ello es la apertura de espacios en los que sea posible algo así como una desaceleración de las decisiones para permitir el libre intercambio de las opiniones y los puntos de vista. Una democracia requiere esta capacidad cuando se trata de satisfacer preferencias e intereses diversos, que no pocas veces plantean exigencias disparatadas.

A este respecto, la presencia del pueblo en la democracia algorítmica es más de *volonté de tous* que de *volonté générale*, por utilizar la terminología de Rousseau, más de agregación que de configuración, de soberanía que de democracia: nuestras preferencias de partida son tomadas en consideración, por supuesto, pero se nos priva del momento de construcción deliberativa en el que esas preferencias ya no son meramente agregadas, sino que interaccionan con otras. De este modo ya no se abre un espacio de indeterminación que permitiera una reformulación transformadora de tales preferencias atendiendo a su (in)compatibilidad con las de otros, sus argumentos y la idea de una totalidad social que puede adquirirse en ese proceso de discusión (Martí, 2021; García Marzá y Calvo, 2022). Estos sistemas no contemplan otro modelo que el de unos individuos maximizando su utilidad. El problema de la gobernanza algorítmica es que gracias a los algoritmos intervenimos en la expresión de preferencias e intereses, pero no en la construcción de una totalidad social deseable que nos habría permitido eventualmente modificarlos. Nuestra presencia en el proceso democrático algorítmico sería la de poner nuestros rastros y huellas a disposición de los sistemas de decisión, pero no la de intervenir en el diálogo en el que se ponderan esos datos y se delibera acerca de la idea de sociedad deseable a partir de ellos. En una democracia algorítmica ser ciudadano consistiría en tener el derecho a emitir deseos, pero no a ponderarlos con los de otros e incluso modificar esos deseos propios. La ciudadanía se reduciría a la generación de datos. Este modelo de gobernanza tiene al menos estas debilidades desde el punto de vista democrático: 1) que pensemos que al emitir señales digitales ya hemos expresado suficientemente lo que queremos; 2) que lo hayamos hecho sin interiorizar explícitamente la compatibilidad de nuestra voluntad con la de otros; y 3) que de este modo nos creamos eximidos de pensar en qué tipo de sociedad resultante queremos.

Nos encontraríamos ante dos tipos diferentes de racionalidad y sus correspondientes modelos de gobernanza. Una «gubernamentabilidad algorítmica» que es implícita, con criterios emanados en tiempo real de la realidad digitalizada y una «gubernamentabilidad política» explícita que resulta de una deliberación que consume tiempo (Rouvroy, 2013, p. 66). La primera de ellas, impulsada tecnológicamente, parece desafiar la interrogación, el análisis y la rendición de cuentas (Waldman, 2019, p. 72), conduciendo así a configurar un entorno político e institucional sin un debate significativo ni oportunidades de impugnación (Green y Hu, 2018). Las psicotecnologías automatizadas de la digitalización pueden debilitar la democracia en la medida en que dificulten realizar su dimensión deliberativa. En vez de constituir sujetos políticos que estén dispuestos y sean capaces de entrar en un proceso de reflexión conjunta sobre la configuración del bien común, pueden estar generando sujetos apolíticos a quienes la idea misma de una negociación democrática de intereses y preferencias les resulte incomprensible.

Pensar adecuadamente la democracia en un entorno algorítmico requiere entender en qué consiste la voluntad política, que no es la afirmación solipsista de lo que yo quiero ni la simple agregación de voluntades configuradas de un modo a-político. La paradoja que planteo es que precisamente cuando estamos tratando de establecer un marco conceptual para pensar la justicia, es decir, la no discriminación, hace falta revisar el papel que desempeña el individuo en una democracia. La justicia no exige que generemos una digitalización humano-céntrica, sino «humano-descentrada», en el sentido de que posibilitemos aquellas experiencias de comunicación, contestación y conflicto que nos hacen a los humanos seres sociales. Y puede estar ocurriendo que la personalización algorítmica en sus diversas formas (granularización, mercantilización, *microtargeting*…) esté limitando la diversidad de información, la exposición a puntos de vista alternativos y dificultar el descubrimiento de posibles preferencias.

Tal vez la justicia esté exigiendo otros objetivos que posiblemente estén en conflicto con la personalización. Solo una concepción extremadamente individualista de lo social puede consagrar el interés individual hasta el punto de considerarlo la última palabra y hacerlo en nombre de la justicia. ¿Qué hacemos cuando aquello que el usuario quiere contribuye a la injusticia? El concepto de «*post-userism*» es un planteamiento teórico que cuestiona el foco que ha dominado la interacción entre los humanos y las máquinas proponiendo la posibilidad de considerar un marco más amplio (Baumer y Brubaker, 2017). En vez de centrarse en el individuo, habría que entender la justicia en términos de distribución: cómo está distribuido el

daño o el beneficio entre los diferentes individuos y grupos. La elicitación de preferencias a partir de un conjunto fijo de alternativas es a menudo insuficiente e injusto, dado que estas preferencias reflejan los sesgos y las desigualdades existentes en una sociedad y dado que esos métodos no vienen acompañados de una deliberación democrática significativa (Robertson y Salehi, 2020).

El cambio de aproximación hacia un modelo deliberativo implica también un cambio en cuanto al modo de considerar el punto de partida (los intereses o preferencias individuales), que pasarían a ser entendidos como algo indeterminado, flexible, dinámico y cambiante. Podría conseguirse así una cierta convergencia entre la ciencia computacional y la teoría deliberativa de la democracia. Dice Stuart Russell que la inteligencia artificial ha prestado poca atención a la incertidumbre, como si hubiera siempre un perfecto conocimiento del objetivo. Esto puede valer para determinados juegos, pero para otro tipo de problemas las preferencias relevantes no son inicialmente conocidas. La idea de cómo tomar decisiones con objetivos abiertos e indeterminados es un desafío tanto para la computación como para la teoría de la democracia. Hablar de intereses y preferencias como si fueran evidencias y además de fácil implementación es una simpleza incompatible con la complejidad de los humanos y de nuestras sociedades (Innerarity, 2019 y 2023). ¿Cómo conseguimos que un robot aprenda a entender las preferencias subyacentes en el comportamiento humano, que es «irracional inconsistente, de voluntad débil y computacionalmente limitado, por lo que sus acciones no siempre reflejan sus verdaderas preferencias»? (Russell, 2019, p. 32).

La cuestión de los sesgos y la equidad algorítmica suele plantearse como si la justicia consistiera en respetar unas propiedades o intereses que las personas o grupos *tienen* y no como la generación de un marco en el que esas personas o grupos puedan relacionarse reflexivamente con sus propiedades o intereses. Se trataría por tanto de decidir de acuerdo con unas preferencias humanas cuya plasticidad permite que vayan cambiando con el tiempo y especialmente en el diálogo y conflicto con las de otros (Pettigrew, 2020). No estamos solo ante la exigencia de autogobernarnos, sino ante la posibilidad de cambiarnos. Para ello se requiere un entorno algorítmico que no se limite a registrar lo que fácticamente revelamos querer, sino que permita una autocontrolada capacidad de desafiar esa facticidad y modificarla. Y aquí la consecución de un equilibrado sistema algorítmico es de la mayor importancia, ya que tenemos que reconsiderar cómo gestionamos el posible conflicto entre la satisfacción de nuestras preferencias inmediatas y nuestra capacidad de configurar esas preferencias en el largo plazo. Los algoritmos pueden dar más peso a las preferencias de largo plazo y a las preferencias racionales sobre las inmediatas y emocionales, pero también puede ocurrir exactamente lo contrario y que un sistema algorítmico de mero registro de nuestro comportamiento impida la consideración de futuros alternativos. Un nuevo giro deliberativo de la democracia en la era de la inteligencia artificial corregiría el «hedonismo psicológico» (Gal, 2017) al que se reduce la democracia cuando se limita a la satisfacción digital de las preferencias individuales. La democracia requiere un cierto grado de «incomodidad», por ejemplo, para limitar los deseos individuales cuando afectan negativamente al conjunto de la sociedad, para asegurar la autonomía personal o para introducir consideraciones de largo plazo que puedan estar en conflicto con los intereses inmediatos. No hay verdadera autodeterminación si no podemos pensar más allá de nosotros mismos y de la actual configuración de la sociedad. Solo esta capacidad asegura la vitalidad de una sociedad democrática.

CONCLUSIÓN

La relación entre los humanos y sus tecnologías está llena de paradojas. Una de ellas es el hecho de que los seres humanos diseñamos los algoritmos y, por así decirlo, les dotamos de una vida propia para reducir nuestra arbitrariedad, lo que a su vez introduce nuevos sesgos en el escenario. Se trata de una operación en cierto modo paradójica porque unos sesgos corrigen otros y a su vez producen otros nuevos. Nos enfrentamos a un dilema que podría implicar una regresión al infinito y para cuya solución no parece tecnológicamente beneficiosa una intervención temprana de los humanos, pero mucho menos aceptable desde el punto de vista normativo y de legitimidad que dispongan los algoritmos de la última palabra. Dado el actual estado de la tecnología y, sobre todo, nuestras aspiraciones de configurar un entorno algorítmico justo, la solución deseable deberá de consistir en un equilibrio entre performatividad tecnológica e intervención humana responsable.

Si la cuestión de la justicia fuese un concepto pacíficamente compartido o una medición objetiva y calculable, entonces una agregación algorítmica podría hacerse cargo de la compatibilización de intereses y preferencias. Dado su carácter controvertido, es decir, político, la tecnología puede ayudarnos, pero no parece capaz de resolver este

problema. Y a este respecto la teoría deliberativa de la democracia ofrece el mejor marco en el que organizar la conversación para corregir la supuesta objetividad algorítmica con las plurales preferencias de los humanos.

BIBLIOGRAFÍA

Amoore, Louise (2020), *Cloud Ethics. Algorithms and the Attributes of Ourselves and Others,* Durham / London: Duke University Press.

Arrow, Kenneth J. (1950), "A Difficulty in the Concept of Social Welfare", *Journal of Political Economy* 58, 328-346. https://doi.org/10.1086/256963

Baumer, Eric y Brubaker, Jed (2017). "Post-userism", en *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, 6291-6303.

Beck, Ulrich; Lash, Scott y Giddens, Anthony (1994), *Reflexive Modernization,* Cambridge: Polity Press.

Berk, Richard; Heidari, Hoda; Jabbari, Shahin; Kearns, Michael y Roth, Aaron (2017), "Fairness in Criminal Justice Risk Assessments: The State of the Art", *arXiv preprint*, arXiv:1703.09207

Binns, Reuben (2018a), "Fairness in Machine Learning: Lessons from Political Philosophy", *Proceedings of Machine Learning Research* 81, 149-159.

Binns, Reuben (2018b), "Algorithmic Accountability and Public Reason", *Philosophy & Technology* 31, 543-556. https://doi.org/10.1007/s13347-017-0263-5

Brooks, Frederick P. (1975), *The Mythical Man-Month: Essays on Software Engineering*, Massachusetts: Addison-Wesley.

Coglianese, Cary y Lai, Alicia (2022), "Algorithm vs. Algorithm", *Duke Law Journal*, Vol. 72, University of Pennsilvania Law School, Public Law Research Paper No. 22-11, Available at SSRN: https://ssrn.com/abstract=4026207

Collins, Patricia Hill (2002), *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment,* New York: Routledge.

Coyle, Diane y Weller, Adrian (2020), "Explaining Machine Learning Reveals Policy Challenges", *Science* 386 / 6498, 1433-1434. https://doi.org/10.1126/science.aba9647

Crenshaw, Kimberlé (ed.) (2019), *Seeing Race Again: Countering Colorblindness across the Disciplines,* Berkeley: University of California Press.

Christian, Brian (2020), *The Alignment Problem: Machine Learning and Human Values,* New York: Norton & Company.

Dieterich, William; Mendoza, Christina y Brennan, Tim (2016), "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity", Northpoint Inc. Available Online at: http://go.volarisgroup.com/rs/430- MBX-989/images/ProPublica_Commentary_Final_070616.pdf

Dwork, Cynthia; Hardt, Moritz; Pitassi, Toniann; Reingold, Omer y Zemel, Richard (2012), "Fairness through awareness", *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference,* 214-226.

Friedler, Sorelle; Scheidegger, Carlos y Venkatasubramanian, Suresh (2016), "On the (Im)possibility of Fairness", *arXiv preprint*, arXiv:1609.07236.

Gal, Michal (2017), "Algorithmic challenges to autonomous choice", *Michigan Telecommunications and Thechnology Law Review* 25, 59-104. https://doi.org/10.36645/mtlr.25.1.algorithmic

García Marzá, Domingo y Calvo, Patrici (2022), "Democracia algorítmica: ¿un nuevo cambio estructural de la opinión pública?", *Isegoría*, (67/17). https://doi.org/10.3989/isegoria.2022.67.17

Green, Been y Hu, Lily (2018), "The myth in the methodology: Towards a recontextualization of fairness in machine learning", *Machine Learning: The Debates workshop at the 35th International Conference on Machine Learning*. https://www.benzevgreen.com/wpcontent/uploads/2019/02/18-icmldebates.pdf

Hanna, Alex; Denton, Emily; Smart, Andrew y Smith-Loud, Jamila (2020), "Towards a critical race methodology in algorithmic fairness", *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3351095.3372826

Hardt, Moritz (2014), "How Big Data Is Unfair: Understanding Unintended Sources of Unfairness in Data Driven Decision Making", *Medium*, September 26. https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de

Innerarity, Daniel (2019), "Democratic equality: an egalitarian defense of political mediation", *Constellations. An International Journal of Critical and Democratic Theory*, 26/4, 513-524. https://doi.org/10.1111/1467-8675.12402

Innerarity, Daniel (2023), *A theory of complex democracy. Governing in the Twenty-first century,* London: Bloomsbury.

Jolls, Chistine; Sunstein, Cass R. y Thaler, Richard (1998), "A Behavioral Approach to Law and Economics", *Stanford Law Review* 50 (5), 1471-550.

Kahneman, Daniel (2011), *Thinking, Fast and Slow,* New York: Farrar, Straus and Giroux.

Kasy, Maximilian y Abebe, Rediet (2021), "Fairness, equality, and power in algorithmic decision-making", en *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 576-586.

Lai, Alicia (2018), *Brain Bait: Effects of Cognitive Biases on Scientific Evidence in Legal Decision-Making,* A.B. thesis, Princeton University.

Martí, José Luis (2021), "New Technologies at the Service of Deliberative Democracy" en Amato, Guiliano; Barbisan, Benedetta y Pinelli, Cesar (eds.), *Rule of Law vs Majoritarian Democracy,* New York: Bloomsbury, 199-220.

Miconi, Thomas (2017), "The Impossibility of 'Fairness': A Generalized Impossibility Result for Decisions", arXiv preprint, arXiv:1707.01195

Mitchell, Shira; Potash, Eric; Barocas, Solon; D'Amour, Alexander y Lum, Kristian (2021), "Algorithmic fairness: Choices, assumptions, and definitions", *Annual Review of Statistics and Its Application*, 8, 141-163. https://doi.org/10.1146/annurev-statistics-042720-125902

Mittelstadt, Brent; Allo, Patrick; Taddeo, Mariarosaria; Wachter, Sandra y Floridi, Luciano (2016), "The ethics of algorithms: Mapping the debate", *Big Data & Society*, 3/2, July-December. https://doi.org/10.1177/2053951716679679

Ochigame, Rodrigo; Barabas, Chelsea; Dinakar, Karthik; Virza, Madars e Ito, Joichi (2018), "Beyond legitimation: Rethinking fairness, interpretability, and accuracy in machine learning", en *The Debates, at the 35th International Conference on Machine Learning*.

Parnas, David Lorge (1985), "Software Aspects of Strategic Defense Systems", *American Scientist*, September-October 1985, 432-440.

Pettigrew, Richard (2020), *Choosing for Changing Selves,* Oxford University Press.

Robertson, Samantha y Salehi, Niloufar (2020), "What if I don't like any of the choices? The limits of preference elicitation for participatory algorithm design", en *Participatory Approaches to Machine Learning Workshop*, ICML 2020. https://arxiv.org/pdf/2007.06718.pdf

Rouvroy, Antoinette (2013), "The end(s) of critique: data-behaviourism vs. due process", en Hildebrandt, Mireille y de Vries, Katja (eds.), *Privacy, Due Process and the Computational Turn. Philosophers of Law Meet Philosophers of Technology,* New York: Routledge.

Russell, Stuart (2019), "The purpose put into the machine", Brockman, John (ed.), *Possible Minds. 25 Ways of Looking at AI,* New York: Penguin, 20-32.

Selbst, Andrew D.; Boyd, Danah; Friedler, Sorelle A.; Venkatasubramanian, Suresh y Vertesi, Janet (2019). "Fairness and abstraction in sociotechnical systems", en *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3287560.3287598

Thaler, Richard H. (2015), *Misbehaving: The making of behavioral economics,* New York: Norton & Co.

Waldman, Ari Ezra (2019), "Power, Process, and Automated Decision-Making", 88 *Fordham Law Review* 613. Available at: https://ir.lawnet.fordham.edu/flr/vol88/iss2/9

Wang, Annie J. (2018), "Procedural justice and risk-assessment algorithms". *SSRN Electronic Journal* 2018: 1-31. https://doi.org/10.2139/ssrn.3170136.

Zarsky, Tal (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, 118-132. https://doi.org/10.1177/0162243915605575

Züger, Theresa; Milan, Stefania y Tanczer, Leonie Maria (2017), "Sand im Getriebe der Informationsgesellschaft: Wie digitale Technologien die Paradigmen des Zivilen Ungehorsams herausfordern und verändern", en *Politische Theorie und Digitalisierung,* Jacob, Daniel y Thiel, Thorsten (eds.), Baden-Baden: Nomos, 265-296.