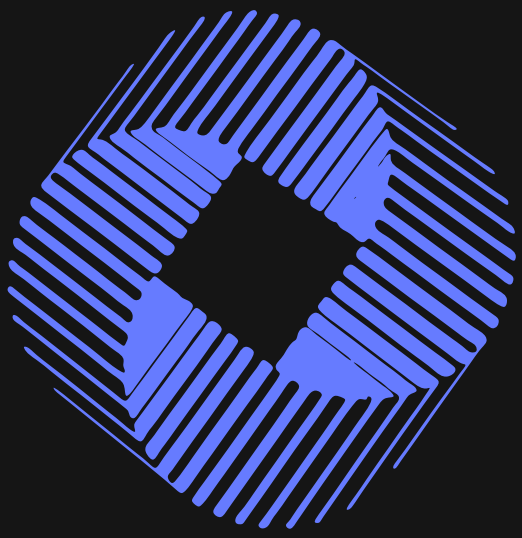
TecnoPsique.

LECTURAS / TRADUCCIONES



Subjetividades algorítmicas

Eric P. S. Baumer · Alex S. Taylor

Jed R. Brubaker · Micki McGee

Traducción de trabajo al español para lectura académica. Se conservan las notas de la versión recibida.

Subjetividades algorítmicas

Eric P. S. Baumer · Alex S. Taylor · Jed R. Brubaker · Micki McGee

Traducción de trabajo al español para lectura académica. Se conserva la numeración de citas del original. Las referencias bibliográficas se mantienen en su idioma de publicación para facilitar su localización. Fuente: *ACM Transactions on Computer-Human Interaction*, 31(3), artículo 35, agosto de 2024. DOI: 10.1145/3660344.

Resumen

Este artículo examina cómo las subjetividades cobran vida en los sistemas algorítmicos. En primer lugar, revisamos la bibliografía relacionada para precisar nuestra concepción de las «subjetividades» como emergentes de una trama enmarañada de procesos y actores. A continuación, ofrecemos dos estudios de caso que ejemplifican la emergencia de subjetividades algorítmicas: uno referido al modelado computacional de tópicos en blogs escritos por madres y padres de niñas y niños dentro del espectro autista, y otro referido a la moderación algorítmica de contenidos en redes sociales. A partir de estos casos, articulamos una serie de cualidades que caracterizan a las subjetividades algorítmicas. También comparamos y contrastamos esas cualidades con diversos conceptos afines de la bibliografía previa, a fin de mostrar de qué manera las subjetividades algorítmicas constituyen una contribución teórica novedosa y ofrecen una lente focal para futuras investigaciones empíricas y para el diseño. En suma, el artículo señala cómo ciertos mundos están siendo producidos y/o posibilitados mediante sistemas algorítmicos, e invita a la Interacción Humano-Computadora (HCI) a considerar qué otros mundos podrían ser posibles.

Conceptos CCS: computación centrada en lo humano → teoría, conceptos y modelos de HCI; teoría, conceptos y paradigmas del diseño de interacción.

Palabras clave adicionales: subjetividad, algoritmos, HCI reflexiva.

1. Introducción

Numerosas tecnologías interactivas incorporan hoy sofisticadas formas de aprendizaje automático (*machine learning*, ML) y/o inteligencia artificial (IA). Los ejemplos abarcan desde el análisis de sentimientos en reseñas en línea [199], pasando por la curaduría de los muros de noticias en redes sociales [71], hasta la publicidad dirigida algorítmicamente [145, 188]. Muchas de estas tecnologías se basan en inferencias automatizadas acerca de atributos de usuarios individuales, entre ellos edad, raza, género, redes de amistad, preferencias mediáticas, opiniones políticas, creencias religiosas y otros rasgos [97]. Un corpus de investigación emergente estudia los tipos de interacciones y experiencias a los que estos sistemas dan lugar, así como las maneras de diseñarlos [5, 10, 16, 36, 48, 61, 71, 72, 80].

Este artículo realiza una contribución teórica a esa bibliografía en crecimiento al considerar cómo las subjetividades **se hacen** mediante algoritmos. En particular, propone el concepto de **subjetividades algorítmicas**, un enfoque que nos invita a reconocer, atender y dar cuenta de las maneras en que algoritmos, seres humanos y sistemas más amplios en los que se encuentran insertos se co-constituyen mutuamente, no solo como entidades sino también como categorías conceptuales. Este enfoque no toma «humano» y «algoritmo» como conceptos prefigurados, sino como categorías que únicamente llegan a existir a través de sus conexiones e interacciones mutuas.

Este uso del término «subjetividades» se alinea con perspectivas desarrolladas por Foucault [77, 78], Deleuze y Guattari [58], Ahmed [4] y otros autores. Sostiene que los sí-mismos, en lugar de ser independientes y preexistentes, son continuamente puestos en existencia por constelaciones de actores, estructuras y procesos heterogéneos. En vez de presuponer que los sujetos preceden a sus subjetividades, reconocemos una relación mucho más compleja que entreteje qué son exactamente los sujetos y los fenómenos de la experiencia. Dicho de manera

sencilla: en lugar de sostener que los seres humanos **tienen** experiencias, nuestra perspectiva se concentra en las maneras en que esas experiencias llegan a producir y definir aquello que concebimos como humano.

A primera vista, este uso de «subjetividades» podría parecer un simple reconocimiento de la pluralidad de experiencias subjetivas que tienen las personas, o incluso un intento posmoderno de cuestionar la realidad afirmando que todo está socialmente construido. Nuestra intención, sin embargo, es establecer una distinción ontológica importante. Buscamos llamar la atención sobre cómo los sujetos llegan a existir a través de una multiplicidad de experiencias subjetivas. Para reiterarlo: esta conceptualización no presupone personas como actores predefinidos que luego experimentan el mundo, ni afirma que las personas no existan realmente. Entiende, más bien, que los actores humanos surgen siempre de las relaciones en las que están entrelazados; están siempre sujetos —si se nos permite el juego de palabras— a la constelación de entidades y procesos a través de la cual emergen las experiencias subjetivas.

Basándose en estos y otros trabajos afines de la teoría crítica [4, 58, 77, 78], los estudios poshumanistas y más-que-humanos [12, 13, 60, 91, 92] y la HCI crítica [14, 15, 17, 185], este artículo ofrece una contribución principalmente teórica al ampliar la teoría de la subjetividad para dar cuenta del papel de los sistemas algorítmicos en el modo en que diversas subjetividades «cobran vida». De manera semejante a la formulación de Ahmed [3, 4] según la cual «los cuerpos toman forma», utilizamos aquí la expresión «cobrar vida» para connotar los procesos de co-constitución mutua antes descritos. En el entretejido de los algoritmos, las interfaces mediante las cuales se los encuentra y las experiencias subjetivas correspondientes, no existe un único actor o proceso responsable de crear o construir subjetividades. Estas llegan a existir, en cambio, a través de las interacciones entre detalles técnicos de implementación, relaciones afectivas, sistemas de clasificación y otros componentes.

Los nudos que enlazan las subjetividades y lo algorítmico son considerados aquí constitutivos tanto del algoritmo como de lo humano, no solo en cuanto entidades individuales sino también en cuanto categorías conceptuales: es decir, en cuanto a cómo humanos y algoritmos son definidos y constituidos. Dada la prevalencia de los algoritmos, la noción de subjetividades algorítmicas ayuda a superar las limitaciones de conceptualizar a las personas y la subjetividad mediante la lente de los «usuarios» [17], y permite considerar cómo se desplazan las condiciones y experiencias de uso en contextos algorítmicos.

Consideremos el siguiente ejemplo. La práctica de un usuario que pone «Me gusta» y luego «Oculta» una misma publicación de su muro de Facebook [71] demuestra cómo estas subjetividades llegan a existir y operan conjuntamente, a veces en tensión. Los algoritmos que subyacen a la curaduría del muro de Facebook —y determinan qué se le muestra al usuario—, junto con funciones interactivas como los botones de «Me gusta», establecen condiciones para la experiencia subjetiva. Ofrecen al usuario un rango estrecho de respuestas posibles y, a su vez, configuran el alcance de sus experiencias. Así, «dar Me gusta y ocultar» establece una relación con un contenido que ha sido ordenado y priorizado algorítmicamente. Dar literalmente el pulgar hacia arriba a una publicación y, al mismo tiempo, degradarla supone responder afectivamente a un resultado determinado de manera automática [véase también 36]. Sin embargo, aunque sea de manera mínima, estas acciones

permiten que alguien sea definido algorítmicamente por las contradicciones inherentes a sus decisiones. Lo decisivo es que el sistema y el usuario son constitutivos el uno del otro: se vuelven entidades específicas y se comprenden como un ensamblaje singular a través de sus interrelaciones; de allí nuestro emparejamiento entre lo algorítmico y las subjetividades.

El «Me gusta» de una red social es apenas un ejemplo de estas complejidades. Un número creciente de técnicas algorítmicas opera sobre muy diversos tipos de datos: textos, imágenes, datos de sensores, redes, geolocalizaciones y muchos otros. En este artículo abordamos ejemplos centrados en el lenguaje y en cómo su procesamiento computacional —denominado con frecuencia procesamiento del lenguaje natural, PLN— enactúa subjetividades de maneras muy particulares. Nuestro objetivo es utilizar el procesamiento del lenguaje como una vía para comenzar a examinar más ampliamente las subjetividades algorítmicas.

A través de una serie de ejemplos concretos, consideramos cómo prestar atención a los tipos de subjetividades que cobran vida mediante sistemas algorítmicos puede orientarnos hacia un diseño más receptivo y responsable. Un diseño de ese tipo es sensible a las interacciones individuales de las personas con sistemas y servicios computacionales; al mismo tiempo, reconoce su compleja imbricación con estructuras sociotécnicas más amplias de clasificación, automatización y agencia, entrelazadas con la proliferación tecnológica y capaces de producir daños [30, 135, 137].

El primero de nuestros dos estudios de caso se refiere al uso de una forma específica de PLN —el modelado de tópicos [24, 25]— para analizar blogs escritos por madres y padres de niñas y niños dentro del espectro autista. El caso se sitúa en el contexto histórico de la clasificación del autismo en el *Manual diagnóstico y estadístico de los trastornos mentales* (DSM), en particular en su cuarta y quinta edición (DSM-IV y DSM-5). Lo utilizamos para ilustrar las relaciones complejas y con frecuencia sutiles entre palabras/lenguaje y subjetividades [de manera semejante a 39, 91], así como el modo en que ciertas subjetividades cobran vida como parte de un ensamblaje de actores. Dicho ensamblaje crea maneras de ver a esos actores y estructuras y delimita los tipos de afirmaciones que pueden formularse acerca de ellos, a menudo con consecuencias potencialmente importantes.

Nuestro segundo estudio de caso amplía esta cuestión de las consecuencias de las subjetividades algorítmicas mediante el ejemplo de la moderación algorítmica de contenidos. Este caso muestra cómo el lenguaje, parcialmente a través del procesamiento computacional, puede entretejerse con innumerables facetas adicionales de los sistemas algorítmicos. También ofrece la oportunidad de investigar las complejas conexiones y tensiones entre experiencias subjetivas —por ejemplo, que el propio contenido sea moderado automáticamente— y detalles técnicos de implementación, como los pesos de un vector de características.

Al sintetizar ambos casos, recurrimos a una variedad de trabajos previos para articular elementos comunes entre distintas instancias de subjetividades algorítmicas. Primero, la capacidad de los algoritmos para inferir o efectuar predicciones distingue estos ensamblajes de aquellos que involucran otras tecnologías informáticas [cf. 159, 172, 189]. Segundo, es a través del entrelazamiento de algoritmos dentro de esos ensamblajes que diversos actores y sistemas se producen mutuamente y que se ejercen poder y autoridad [cf. 13, 38, 79, 112]. Tercero, esta coproducción mutua implica que los algoritmos desempeñen un papel casi paradójico en la **humanización** de los sujetos, es decir, en el establecimiento y la puesta en

acto de definiciones de lo que significa ser humano [cf. 37, 92, 106]. Si bien la bibliografía anterior ofrece conceptos y teorías útiles para comprender estas cualidades, la contribución específica de este artículo consiste en subrayar las consecuencias que surgen de su combinación singular. A través de ellas mostramos que la orientación hacia las subjetividades algorítmicas permite esclarecer cómo determinados mundos [cf. 186] están siendo producidos y/o posibilitados mediante sistemas algorítmicos.

A partir de esta reflexión, el artículo concluye preguntando cómo podrían ser las cosas de otro modo. Aunque su contribución es principalmente teórica, las ideas presentadas también pueden orientar el diseño. Si los sistemas algorítmicos actuales posibilitan determinados mundos, ¿qué otros mundos resultan menos posibles? ¿Qué otras condiciones de posibilidad querríamos enactuar? ¿Qué nuevas subjetividades, ensamblajes de actores o formas de agencia podrían llegar a existir? ¿Qué tipos de subjetividades queremos que los historiadores futuros puedan leer en nuestros sistemas computacionales?

2. ¿Qué son las subjetividades?

Antes de presentar los estudios de caso, esta sección desarrolla el concepto de subjetividades, introduce con mayor detalle la noción de subjetividades algorítmicas y señala dónde se cruza con el diseño de sistemas interactivos y con preocupaciones sociotécnicas más amplias.[1]

2.1. El concepto de subjetividades

Aunque podemos experimentar el mundo de múltiples maneras, es habitual percibirnos como individuos delimitados y comprender las experiencias subjetivas como algo «sentido» mediante procesos interiores [161]. Consideremos cómo, cuando alguien habla en voz alta, los movimientos físicos de sus cuerdas vocales —dentro de su cuerpo— desencadenan movimientos en los tímpanos de otra persona —también dentro de su cuerpo— [136]. Estas experiencias viscerales producen la impresión de que las personas son individuos discretos, cada uno compuesto por una interioridad separada del mundo exterior que lo rodea.

Esta orientación también se refleja en la manera en que la HCI concibe las interacciones humano-computadora o las experiencias de usuario. Pensemos, por ejemplo, en cómo los investigadores de HCI estudian el apoyo social en línea [por ejemplo, 54, 118, 132, 187, 192]. Nuestro campo reconoce las diferencias, en términos de posibilidades de diseño y experiencia subjetiva, entre recibir un emoji de corazón como respuesta a una publicación pública de Facebook y recibirlo mediante Messenger. Sin embargo, se conserva la idea de que quien recibe es una unidad: un sí-mismo distinto y persistente que percibe e interpreta el mismo mundo a través de esos entornos diferentes.

Mediante la teoría de la subjetividad, Bardzell y Bardzell [14] describen cómo la dependencia de este sí-mismo singular y separado genera problemas en el contexto de la HCI y el diseño. Retomando a Foucault y enfoques posmodernos sobre la subjetividad, destacan que la HCI

1 Aquí «subjetividad» no designa lo opuesto a «objetividad», sino que remite a teorías sobre la experiencia subjetiva y la constitución del sí-mismo [por ejemplo, 4, 14, 58, 60]. Esta definición es distinta de la subjetividad entendida como una forma de sesgo, que numerosos trabajos intentan reducir en nombre de la equidad o la justicia [por ejemplo, 43, 51, 98, 142].

confunde con demasiada frecuencia al usuario con un sí-mismo singular y coherente. Proponen hablar de «sujetos de información» y distinguen entre **posiciones de sujeto** —las condiciones estructurales de uso ofrecidas por un sistema— y **subjetividades** —las experiencias vividas y las performances de quienes ocupan esas posiciones—. Bardzell y Bardzell sostienen que abordar el concepto de «usuario» a través de la subjetividad resulta útil porque evita colapsar de manera simplista a una persona y a un usuario como si fueran equivalentes; Baumer y Brubaker [17] desarrollan posteriormente este argumento.

Así como alguien puede mantener múltiples direcciones de correo electrónico para propósitos distintos —por ejemplo, diversos papeles profesionales y personales—, considerar cada una como un usuario distinto, con independencia de que detrás exista la misma persona, ayuda a diferenciar realidades estructurales y contextuales específicas, esto es, posiciones de sujeto. El sí-mismo separado situado detrás del «usuario», sostienen, no es tan relevante para el diseño como la posición estructural desde la que ocurre el uso y las experiencias subjetivas con las que se conecta.

Si el sí-mismo separado ofrece beneficios limitados para el diseño, una perspectiva contrapuesta podría sostener que ni siquiera somos actores unitarios. La teoría posestructuralista y posmoderna tardía ha argumentado que es la constelación de subjetividades la que da lugar a posiciones de sujeto, y no a la inversa. Somos constituidos continuamente como sujetos diferenciados a través de las condiciones y relaciones singulares que experimentamos. Aquí nuestros sí-mismos son múltiples, no puramente interiores, y siempre están figurados o enactuados en y a través de las diversas relaciones con las que nos encontramos [4, 58, 60]. Un enfoque constitutivo de la subjetividad presenta al sí-mismo —definido mediante experiencias subjetivas internas— como apenas una entre muchas maneras de comprender a un actor humano y sus agencias, y quizá una manera bastante empobrecida. Por ello, nuestro trabajo se concentra en las relacionalidades de estas constelaciones que producen posiciones de sujeto. En vez de actores que buscan un escenario, somos performances resultantes del escenario y de los ensamblajes dentro de los cuales ese escenario existe.

Desde esta perspectiva, Facebook, TikTok y el correo electrónico proporcionan conjuntos diferentes de afinidades y responsabilidades que permiten la cohesión de distintos tipos de ensamblajes y sí-mismos. En las redes sociales, por ejemplo, las redes difusas de millones de usuarios e interacciones se vuelven posibles mediante subjetividades experimentadas a menudo en términos de métricas simples como el «Me gusta». Aunque nuestras experiencias en esas plataformas son siempre más complejas que las métricas por sí solas [89], aquí buscamos destacar las interrelaciones entre esas métricas, los componentes técnicos correspondientes de nuestros ensamblajes y nuestras experiencias en las plataformas.

2.2. Definición de las subjetividades algorítmicas

Es este enfoque relacional de las subjetividades el que retomamos al introducir y desarrollar el concepto de subjetividades algorítmicas. Proponemos el término como una **herramienta para pensar** [174] que permite observar cómo los algoritmos, en particular, interactúan dentro de sistemas más amplios para crear las condiciones de determinadas experiencias subjetivas y para que ciertos tipos de actores habiten y operen en ellas. Entendidas de este modo, las subjetividades algorítmicas invierten la ortodoxia de la HCI: ya no se trata de

observar cómo un usuario experimenta un mundo impregnado de tecnología, sino cómo actores tales como los usuarios llegan a existir a través de lo experiencial.

Este enfoque difiere de trabajos previos que han considerado las relaciones entre subjetividad y algoritmos. Blackwell [23, p. 193], por ejemplo, sostiene que «las definiciones operativas de la IA siempre deben construirse en relación con los humanos». Su preocupación principal, sin embargo, reside en cómo estas nociones humanas de (inter)subjetividad configuran la evaluación de los sistemas de IA —sobre todo las funciones objetivo mediante las cuales son diseñados—, más que en las clases de experiencias humanas subjetivas que surgen alrededor de la interacción con dichos sistemas.

De manera relacionada, Fisher [74] pregunta cómo los sistemas algorítmicos proporcionan a las personas una forma particular de conocerse a sí mismas. No obstante, su argumento se apoya en una formulación de la subjetividad «como ámbito cuasi trascendental —a la vez utópico y efectivo— de libertad y autenticidad individuales» [74, p. 1]. Dicho de otro modo, se ocupa más de la constitución del humano como sujeto (político) legítimo que de la co-constitución de humano y algoritmo como categorías conceptuales.

Más próximo a nuestro artículo, Armano et al. [11] se interesan por las condiciones del capitalismo de plataformas. Consideran, por ejemplo, cómo ciertos «modos específicos de sentir», a través de las plataformas, se convierten en formas de subjetividad, maneras implícitas de seleccionar opciones y, en último término, de mirar el mundo [11, p. 3]. Sin embargo, prestan menor atención a la interacción entre esas «maneras de mirar el mundo» y los detalles técnicos de los componentes algorítmicos de las plataformas. Por ello, nuestra forma de definir y articular las subjetividades algorítmicas ofrece una perspectiva singular sobre estos fenómenos.

Esta perspectiva es importante por al menos dos razones. En primer lugar, invita a ir más allá de las experiencias individuales del usuario y a prestar una atención minuciosa a las condiciones y relaciones que hacen posibles determinadas subjetividades. Ayuda, por tanto, a examinar las interacciones humano-computadora mucho más allá de la experiencia subjetiva individual de un usuario o incluso de grupos, tanto si están en un mismo lugar como si se encuentran distribuidos. Al hacerlo advertimos que aquello que solemos concebir como «el usuario» y como «las interacciones» está constituido, en realidad, mediante entrelazamientos heterogéneos de estructuras más amplias: interfaces y algoritmos, ciertamente, pero también sistemas de normas, valores, agencia y otros elementos que organizan y administran los múltiples mundos que experimentamos [17, 177].

En segundo lugar, esta perspectiva permite especular acerca de las condiciones de posibilidad para hacer esos mundos de otro modo. Esto va más allá de formular preguntas contrafácticas sobre cómo podrían ser otros mundos posibles [116, 162; véase también 186]. La reducción de la experiencia humana al «Me gusta» de Facebook, por ejemplo, afirma un mundo singular —un mundo que representa al sí-mismo subjetivo en términos de determinadas formas sociales, económicas y políticas de vida [véase, por ejemplo, 57]— y excluye otros mundos posibles. La perspectiva aquí defendida propone considerar cómo no solo la existencia de un «Me gusta», sino también su implementación técnica, restringe posibilidades al volver menos probables ciertos mundos. Pensar en términos de subjetividades algorítmicas proporciona, sugerimos, una base para diseñar mundos impregnados de algoritmos que puedan ser

distintos y crear condiciones en las que se tenga en cuenta una mayor variedad y diversidad de actores [93].

2.3. Intersección con el diseño

Para el diseño, la relevancia de esta visión de los sí-mismos subjetivos y las subjetividades algorítmicas se vincula con aquello que la HCI considera su marco de análisis, y con el alcance y la escala de aquello sobre lo que intervenimos al diseñar sistemas interactivos. Bardzell y Bardzell [14] distinguen entre posiciones de sujeto y subjetividades, en gran medida con fines de claridad. Sin embargo, los sistemas algorítmicos ponen de relieve su carácter acoplado, y es precisamente ese acoplamiento lo que atiende este artículo.

Si bien Bardzell y Bardzell ofrecen una fundamentación teórica valiosa, nuestro trabajo la desarrolla para concentrarse en cómo emergen y se transforman las subjetividades, especialmente en relación con el diseño de sistemas algorítmicos. Dentro de estos sistemas, interacciones aparentemente triviales —como clics en un panel táctil o la velocidad de un deslizamiento— pueden ser aprovechadas y analizadas para retroalimentar el diseño. Este acoplamiento entre agregación, análisis y retroalimentación comienza a ordenar y regular formas de expresión subjetiva y, en correspondencia, a dar forma a versiones legítimas del sí-mismo subjetivo [34, 126]. La importancia de lo que sigue para la HCI consiste tanto en mostrar por qué resulta crucial ampliar el alcance y la escala de lo que estudiamos y diseñamos, como en ofrecer ejemplos y conceptos para hacerlo.

3. Estudios de caso

Esta sección rastrea el hacerse de las subjetividades algorítmicas mediante dos estudios de caso. El primero aborda la aplicación del análisis computacional a blogs escritos por madres y padres de niñas y niños dentro del espectro autista. Tras una breve revisión del trasfondo histórico de la clasificación del autismo, el caso considera cómo los sistemas algorítmicos que emplean PLN pueden desempeñar un papel específico en el modo en que ciertas subjetividades cobran vida.

El segundo estudio analiza la moderación algorítmica de contenidos en línea. Señala las maneras en que las subjetividades algorítmicas surgen de ensamblajes complejos de datos —no solo lingüísticos, sino también multimodales y multifacéticos— y algunas de sus consecuencias.

Ninguno de los dos casos pretende ampliar el conocimiento general de sus respectivos ámbitos de aplicación, es decir, del trastorno del espectro autista (TEA) o de la moderación de contenidos. Su propósito es explorar el papel de los sistemas algorítmicos en la producción de subjetividades. Sintetizar los casos, tanto en sus semejanzas como en sus diferencias, permite enumerar las cualidades singulares que caracterizan la forma en que estas situaciones surgen y operan.

3.1. Modelado de tópicos y autismo: clasificación

Esta sección ilustra los papeles particulares que desempeña el procesamiento algorítmico de textos al hacer cobrar vida a ciertos tipos de subjetividad dentro de sistemas de clasificación más amplios. En particular, el papel problemático del lenguaje en la historia de la

clasificación y el diagnóstico del TEA ayuda a mostrar por qué el artículo se concentra en el PLN. Después de revisar parte del trasfondo histórico sobre la importancia del lenguaje para clasificar el autismo, interrogamos las subjetividades que surgen en el contexto de un tipo específico de algoritmo de PLN.

3.1.1. Trasfondo histórico

Los procedimientos de clasificación y sus consecuencias desempeñan desde hace mucho tiempo un papel relevante en la salud y la medicina, especialmente en la salud mental. Los ejemplos abarcan desde la clasificación internacional de enfermedades [30], pasando por la construcción social y el tratamiento cambiantes de la «locura» o la «insania» [76], hasta la reificación de usuarios «saludables» y «trastornados» en análisis de datos de redes sociales mediante aprendizaje automático [44].

Este estudio se concentra específicamente en la clasificación y el diagnóstico del TEA. El DSM, elaborado por la Asociación Estadounidense de Psiquiatría, proporciona una taxonomía de categorías de trastornos mentales, criterios diagnósticos vinculados con ellas y estadísticas correspondientes. El manual ha atravesado numerosas revisiones, desde el DSM-I de 1952 hasta el DSM-5 de 2013.

Una de esas revisiones, el DSM-IV de 1994 [7], contenía un error tipográfico en los criterios diagnósticos del autismo [105, p. 520]. En concreto, incluía «un deterioro grave y generalizado en el desarrollo de la interacción social recíproca **o** de las habilidades de comunicación verbal y no verbal, o cuando se presentan conductas, intereses y actividades estereotipados» [7, p. 77, énfasis añadido]. Como resultado, los pacientes que presentaban deterioro en el desarrollo de la interacción social recíproca **o** deterioro en el desarrollo de las habilidades de comunicación verbal y no verbal eran diagnosticados con Trastorno Generalizado del Desarrollo No Especificado. Esta categoría incluía el autismo atípico, es decir, «presentaciones que no satisfacen los criterios del trastorno autista» [7, p. 78].

El error fue corregido en el DSM-IV-TR de 2000 [8], donde los criterios pasaron a decir: «un deterioro grave y generalizado en el desarrollo de la interacción social recíproca **asociado con** deterioro de las habilidades de comunicación verbal o no verbal, o con la presencia de conductas, intereses y actividades estereotipados» [8, p. 84, énfasis añadido]. Este cambio significaba, en la práctica, que para recibir un diagnóstico una persona debía presentar tanto deterioro de la interacción social recíproca como deterioro de la comunicación —verbal o no verbal— o conductas, intereses y actividades estereotipados. Entre 2007 y 2012 se realizaron nuevas revisiones, que condujeron a la publicación del DSM-5 en 2013.

Esta situación contribuyó a la incertidumbre acerca de la denominada «epidemia» de autismo de fines de la década de 1990. ¿Había aumentado realmente la prevalencia de personas con la constelación de conductas y hábitos mentales que había llegado a denominarse autismo? ¿O se trataba de un error clasificatorio, precipitado en parte por un descuido de corrección editorial en el marco de otros ajustes al DSM-IV-TR? Aunque probablemente estas preguntas nunca obtengan una respuesta definitiva [82], la legitimidad de los criterios diagnósticos se convirtió en objeto de escrutinio público y profesional.

Estas complejidades del lenguaje, la clasificación y sus repercusiones no son aquí nuestro centro principal. Este contexto de controversia —o quizá, más precisamente, de crisis—

ontológica y epistemológica proporciona el trasfondo necesario para comprender la naturaleza del caso.

3.1.2. Subjetividades modeladas mediante tópicos

El caso surge de las propias experiencias de investigación de algunos autores y fue decisivo para advertir la necesidad de pensar la brecha teórica existente alrededor de las subjetividades y los sistemas algorítmicos. Se centra en una colección de blogs escritos por madres y padres de niñas y niños dentro del espectro autista. Los blogs fueron identificados a partir del conocimiento de esta comunidad por parte de McGee y mediante una revisión iterativa de sus contenidos. En total, el corpus comprende 31.976 documentos —entradas de blog— con 17.273.079 palabras; el promedio es de 540,2 palabras por documento y la mediana, de 430.

Dos de los autores, McGee y Baumer, participaron en esfuerzos por analizar y comprender las experiencias relatadas en estos blogs [20]. Debido al volumen del conjunto de datos, el método principal fue el modelado de tópicos [24, 25]. En apariencia, esta técnica proporciona una manera de identificar temas latentes en un corpus documental. Un tema o «tópico» se representa como una distribución de probabilidad sobre palabras. Un tópico puede asignar probabilidades altas a términos como *humano*, *genoma*, *ADN* y *secuencia*, mientras que otro puede hacerlo con *computadora*, *modelos*, *información* y *datos* [24].

Estas distribuciones de probabilidad —los tópicos— son inferidas a partir de un corpus de documentos no etiquetados; el modelo no recibe información previa acerca de cuáles deberían ser los temas. El único parámetro fijado por una persona es la cantidad de tópicos; todo lo demás es no supervisado. El modelo también asigna a cada documento una proporción de tópicos. Siguiendo el ejemplo anterior, un artículo de una revista de bioinformática podría estar compuesto en un 40 % por el primer tópico, en un 30 % por el segundo y en un 30 % por otros tópicos.

Siguiendo trabajos recientes [19, 104, 127, 131, 138, 156], McGee y Baumer aplicaron el modelado de tópicos como una lente interpretativa sobre los datos. Los resultados no fueron tratados como un mapa que representara directamente el corpus, sino como una brújula capaz de indicar direcciones en las que podría ser fructífero mirar [de manera semejante a 70].

Al revisar los tópicos resultantes, advertimos un patrón curioso: muchos iban más allá de ser estrictamente «temáticos». De manera análoga a otros usos interpretativos del modelado de tópicos [157, 179, 180], varios se alineaban estrechamente con determinados discursos, sobre todo con perspectivas o argumentos específicos. Uno de ellos, por ejemplo, asignaba probabilidades altas a palabras como *espectro*, *trastorno*, *autismo*, *diagnóstico* y *trastornos*. Los documentos con una elevada proporción de ese tópico discutían los cambios del DSM-5 en el diagnóstico del TEA, precisamente los cambios descritos en el trasfondo histórico. Además, la mayoría expresaba una posición bastante crítica frente a la transición del DSM-IV-TR al DSM-5.[2]

2 Debido a la facilidad con la que es posible buscar en Internet frases textuales, los autores omiten citas literales de los documentos para proteger la identidad de los blogueros [18].

Así, los tópicos se convierten en lentes mediante las cuales lectores humanos producen sentido sobre los documentos. Una entrada del corpus fue clasificada como un 57,8 % relativa al tópico mencionado —espectro, trastorno, autismo, etc.—. El texto de la publicación consistía enteramente en una copia de los criterios del DSM-IV-TR, bajo el título aproximado «Un momento: ¿mi hijo realmente tiene autismo?».[3] La entrada podría interpretarse de maneras diversas e incluso contradictorias: como crítica de los criterios diagnósticos, como expresión de frustración o confusión, como demostración de la facilidad con la que podían aplicarse esos criterios, como solicitud de ayuda a otras familias para diagnosticar al hijo del autor, entre otras posibilidades.

Sin embargo, saber que el texto posee una proporción elevada de aquel tópico proporciona al lector, antes de comenzar a leer, una indicación de que la publicación —e implícitamente su autor— adopta una posición crítica ante las revisiones del DSM-5. Dicho de otro modo, el modelo de tópicos interviene en el modo en que cobran vida determinadas experiencias subjetivas y relacionalidades alrededor del DSM-5, calculadas algorítmicamente.

Ahora bien, estas subjetividades algorítmicas —por ejemplo, una posición crítica o escéptica frente a los criterios del DSM-5— cobran vida porque una persona lee el modelo de una manera determinada. Dentro del modelo, cada tópico se formula como una distribución de probabilidad sobre palabras, y cada documento como una distribución de probabilidad sobre tópicos. Estas formulaciones dan lugar a un tipo particular de subjetividad, diferente de las descripciones que probablemente ofrecerían personas mediadas por herramientas distintas del modelado de tópicos. Es improbable que investigadores humanos o autores de blogs que sostienen esa postura describan su posición en términos de distribuciones de probabilidad calculables.

La subjetividad cobra vida, por tanto, mediante un ensamblaje de manipulaciones computacionales e interpretaciones humanas del procesamiento. La orientación de este ensamblaje probablemente difiere de la manera en que seres humanos por sí solos o un modelo computacional aislado conceptualizarían y organizarían los mismos datos [cf. 37].

Para ilustrar aún más las cualidades específicas de este ensamblaje humano-modelo, consideremos las posibilidades inferenciales que ofrece el modelado de tópicos [121]. El modelo puede utilizarse para predecir los tópicos presentes en un documento nuevo que no formó parte del corpus de entrenamiento. Ese documento podría ser otra entrada de los mismos blogs, una publicación de otro autor o un texto diferente, como las declaraciones públicas de representantes electos. Las palabras del documento permiten inferir la probabilidad de que contenga uno de los tópicos identificados en el corpus original.

El modelo también puede identificar computacionalmente semejanzas entre autores. Al agregar las probabilidades temáticas de todos los documentos escritos por un mismo bloguero, sería posible calcular la similitud entre cualquier par de autores o incluso identificar grupos. Estas relaciones no se basan únicamente en interacciones directas, sino también en aspectos latentes del uso lingüístico. Por último, los resultados podrían introducirse con facilidad en otros análisis predictivos. La aparición de ciertos tópicos en el blog de una familia podría predecir, por ejemplo, que tiene un hijo no verbal, que lo

3 El título fue deliberadamente modificado para proteger la identidad del autor; véase la nota anterior.

matriculó en una escuela pública o que recibió asistencia sanitaria financiada por el Estado. Estas posibilidades inferenciales constituyen, como desarrollaremos más adelante, una cualidad distintiva de las subjetividades algorítmicas.

Además, las subjetividades algorítmicas no surgen únicamente de interacciones individuales entre humano y algoritmo. Llegan a existir al entrelazarse con estructuras y sistemas mucho más amplios. El trasfondo histórico delimita algunos de ellos: diagnósticos clínicos, compañías de seguros, burocracias estatales, estrategias pedagógicas, entre otros. La incorporación del procesamiento computacional del lenguaje podría alterar de manera rápida y significativa los ensamblajes, procesos y autoridades mediante los cuales diversas entidades enactúan sujetos autistas.

Tales alteraciones tienen consecuencias potencialmente relevantes. Es posible imaginar herramientas semejantes al modelado de tópicos utilizadas como barómetro de la opinión pública o incluso con fines diagnósticos. Keyes [106], en efecto, analiza un amplio corpus de investigaciones sobre IA aplicada al diagnóstico del autismo que recurre a visión computacional —por ejemplo, análisis de la marcha—, procesamiento de señales —como el análisis de grabaciones de audio— y otras técnicas.[4]

En Estados Unidos, un diagnóstico oficial de autismo puede dar acceso a Medicaid, al Ingreso Suplementario de Seguridad, al reembolso de servicios por seguros privados y a otras prestaciones sociales de cuidado. Pero también podría descalificar a un estudiante para acceder a una plaza financiada en escuelas orientadas a quienes son clasificados como personas con «dificultades del habla y el lenguaje» o «alteraciones emocionales». Pueden producirse efectos perjudiciales cuando estos recursos se asignan a partir de designaciones oficiales y no de las necesidades educativas, sociales o emocionales de cada persona [41, 191].

El punto del caso no es, sin embargo, el diagnóstico específico que un algoritmo podría asignar a una persona, ni determinar si un bloguero o progenitor sería clasificado como crítico de ciertos criterios diagnósticos, ni reflexionar en general sobre la construcción social del TEA. El caso destaca las maneras particulares en que surgen subjetividades alrededor de la identificación algorítmica de esa posición crítica. Dicha identificación presenta —y quizá reduce— la crítica de las revisiones del DSM a un patrón estadísticamente reconocible de uso del lenguaje.

El crítico del DSM y el sistema algorítmico se co-construyen de maneras muy específicas. En el modelado de tópicos, estas maneras dependen de distribuciones de probabilidad sobre tokens de palabras que coaparecen y habilitan numerosas inferencias. A la vez, esas inferencias, distribuciones, palabras-token y operaciones son interpretadas desde un contexto sociohistórico, mientras recodifican simultáneamente ese contexto como algo comprensible mediante tokens, distribuciones probabilísticas e inferencias. Examinar estas relaciones permite esclarecer las cualidades singulares de las subjetividades algorítmicas.

---

4 Esta tendencia no se limita al autismo. Se han aplicado técnicas de aprendizaje automático para detectar la enfermedad de Parkinson mediante el análisis del habla [122], y textos y otros datos de redes sociales se han utilizado para detectar diversas condiciones de salud mental [44].

3.2. Moderación algorítmica

La moderación de contenidos —las decisiones acerca de qué puede y qué no puede publicarse en grupos en línea— revela una compleja red de subjetividades que el desarrollo de formas algorítmicas e híbridas de análisis lingüístico y textual extiende, negocia e infraestructuriza. La moderación es un factor decisivo para el éxito de las comunidades en línea [109] y ha recibido una atención considerable por parte de la investigación [por ejemplo, 45, 46, 64, 73, 81, 88, 101, 102, 117, 129, 181, 182]. Las noticias recientes también abundan en casos donde se escrutó el fracaso de la moderación, como grupos de odio, acoso en línea o extremismo violento [1].

A primera vista, la moderación podría parecer sencilla: se identifica y elimina el contenido que vulnera las normas de una comunidad. Una observación más atenta revela que casi siempre es mucho más compleja. Para sostener este argumento, la sección entreteje ejemplos y planteos de la bibliografía previa, tanto sobre moderación de contenidos en general como sobre moderación algorítmica en particular. Esta estrategia complementa el uso de las propias investigaciones y experiencias de los autores en el caso anterior y muestra cómo las preocupaciones por las subjetividades algorítmicas emergen en una amplia variedad de trabajos sobre moderación.

En la mayoría de las grandes plataformas, un ejército de moderadores recorre flujos de contenido en línea que fue denunciado por personas que objetan su presencia. Estos moderadores toman decisiones valorativas basadas en estándares y políticas elaborados por equipos que deben operacionalizar el umbral entre contenidos aceptables e inaceptables. Pero esos estándares no están completamente estandarizados. Las Normas Comunitarias públicas de Facebook explican, por ejemplo, que el acoso, aun cuando puede adoptar muchas formas, no se tolera «porque impide que las personas se sientan seguras y respetadas en Facebook». Al mismo tiempo, las políticas reconocen categorías diferentes para las figuras públicas —a fin de permitir la crítica pública— y para las personas de entre 13 y 18 años —para quienes las protecciones son mayores—.

Además, los equipos de moderadores suelen estar distribuidos globalmente para «seguir al sol» y trabajan con frecuencia mediante contratistas. Informes recientes describen condiciones laborales extremas, donde los moderadores son gestionados y medidos para asegurar que la empresa contratista cumpla sus compromisos [133].

La escala del contenido y la limitación de recursos humanos han dado lugar a enfoques híbridos, en los que algoritmos marcan contenidos potencialmente inapropiados para su revisión por personas. Equipos de ingenieros desarrollan herramientas algorítmicas de apoyo y trabajan con otros actores [83, 160] para decidir qué tipos de contenido deben priorizarse según su volumen, la gravedad de la infracción y la viabilidad técnica.

Algunas infracciones se detectan con relativa facilidad y se delegan por completo a algoritmos. Administradores de pequeñas comunidades de Reddit [107] y Discord [102] pueden utilizar formas simples de «automoderación» basadas en reglas. Ya en 2011, Facebook permitió a los administradores de páginas filtrar automáticamente publicaciones que contenían determinadas palabras obscenas [49]. En estos casos, las herramientas algorítmicas automatizan parte del trabajo humano necesario para moderar esos espacios. Estas técnicas

básicas funcionan simplificando el posible significado —y, por tanto, la adecuación— del contenido y la respuesta considerada apropiada [55].

Sin embargo, muchas infracciones son matizadas y requieren decisiones caso por caso que probablemente excedan la capacidad de cualquier herramienta automática. Situaciones como el discurso de odio, las fotografías de madres amamantando o la documentación de violencia en zonas de conflicto [1] muestran que el «contenido apropiado» está ligado al contexto, el tiempo y la cultura [103]. Haimson et al. [88] señalan, por ejemplo, que la eliminación de contenido —tanto humana como automatizada— afecta desproporcionadamente a determinadas poblaciones, entre ellas conservadores políticos, personas trans y personas negras. Sostienen que las prácticas de moderación deben abordar más directamente las zonas grises —contenidos que no son claramente aceptables ni claramente objetables— y advierten que «existe una enorme diferencia entre silenciar la desinformación y el discurso de odio de conservadores y silenciar el contenido personal vinculado con la identidad de usuarios trans y negros» [88, p. 24].

Incluso contenidos más cotidianos presentan matices al decidir qué medida adoptar. En su estudio sobre moderadores voluntarios, Seering et al. [169] encontraron que la complejidad suele aparecer al determinar el nivel de sanción —desde una advertencia hasta la expulsión— cuando alguien infringe normas específicas de una comunidad. A la vez, puede formarse una línea muy fina y borrosa entre matiz e inconsistencia, y esta última alimentar percepciones de injusticia [117; véase también 40].

Estas complejidades se agravan por los distintos grados de transparencia y opacidad. Las respuestas ante la posibilidad de una curaduría algorítmica desconocida en plataformas como Facebook o X —antes Twitter— suelen adoptar la forma de sorpresa u oposición [62, 72, 147]. En Yelp, las reacciones fueron todavía más cargadas afectivamente: algunos reseñadores describieron la opacidad del filtrado algorítmico como «furtiva», «engañosa», «inductora a error» e incluso «posible censura» [73, p. 6]. Las repercusiones significativas derivan, por consiguiente, no solo de la existencia de moderación algorítmica, sino también de los detalles de su implementación y de sus conexiones con diversas relacionalidades.

De estas situaciones surgen al menos dos cuestiones. En primer lugar, revelan una tensión potencial entre la experiencia subjetiva individual de ser moderado y la preocupación más amplia de los moderadores por crear una esfera de interacción «segura y respetuosa». En segundo lugar, es posible que estos sistemas sean internamente consistentes, pero de maneras que no resultan legibles para los usuarios humanos cuyo contenido moderan. Las preguntas por la legibilidad y la interpretabilidad inciden directamente en los tipos de subjetividades que cobran vida en estos sistemas.

También importa cuál es la **unidad de moderación**. ¿Se modera una pieza específica de contenido? ¿A su autor? ¿A la comunidad en la que fue producida o publicada? En este punto, algoritmos bienintencionados pueden desviarse. Aprovechar el historial de un usuario puede resultar atractivo, pero plantea problemas. Creadores de YouTube han informado que, cuando se modera uno de sus videos —por ejemplo, incorporándole una restricción de edad—, su contenido posterior recibe menos tráfico aun cuando no contenga nada que justifique esa moderación. En otro ejemplo, el troleo se predice mejor por el contexto social que por el historial de la persona [47]. Del mismo modo, quienes acosan suelen haber sido también objeto de acoso [96]. Un algoritmo diseñado para identificar únicamente al autor y al blanco

en piezas individuales de contenido infractor solo podría hacer cobrar vida a posiciones de sujeto mutuamente excluyentes. Se perdería esa multiplicidad matizada.

La unidad de moderación puede ser, alternativamente, la comunidad en la que se produjo el contenido. Reddit, por ejemplo, pone en cuarentena comunidades enteras cuando existe una toxicidad excesiva o información y desinformación cuestionables [150]. Sin embargo, estas cuarentenas han despertado inquietudes acerca de una posible censura [111, 120]. Moderar más allá de una única pieza de contenido o de un usuario puede exceder la capacidad de los algoritmos —o, al menos, aquello que estamos dispuestos a confiarles— incluso cuando trabajan junto con equipos humanos.

Los enfoques híbridos ponen de relieve, por tanto, el modo en que moderadores humanos y algorítmicos están implicados unos con otros. La moderación se vuelve intensamente dependiente de las relaciones sociotécnicas a las que los algoritmos pueden responder. Las condiciones bajo las cuales debería o podría ocurrir la moderación deben volverse legibles para un algoritmo. Dicho de otro modo, moderar algorítmicamente exige operacionalizar de maneras específicas numerosas subjetividades diferentes. La persona acosada, quien acosa, el moderador y las distintas audiencias que presencian la conducta deben ser conceptualizados de formas codificables computacionalmente. La manera en que esto se realiza desplaza tanto las propias subjetividades como sus relaciones mutuas.

La moderación proporciona asimismo otro ejemplo en el que las subjetividades en juego tienen consecuencias. Son distintas de las consecuencias de los casos del DSM y el autismo, pero no por ello menos importantes. La moderación de plataformas en línea es, en efecto, la moderación de nuestra plaza pública contemporánea [181]. Sus consecuencias no son nada más que incontables videos virales y nada menos que la libertad de expresión y el funcionamiento de la democracia.

4. Cualidades de las subjetividades algorítmicas

Toda comprensión significativa de las subjetividades algorítmicas requiere una orientación procesual. Podemos trazar una analogía entre estudiar el sustantivo **infraestructura** —tratarla como una entidad diferenciada que se analiza en sí misma— y estudiar el verbo **infraestructurar** —examinar los procesos mediante los cuales las infraestructuras llegan a existir— [171]. De forma análoga, no postulamos las subjetividades algorítmicas como entidades separadas que deban estudiarse por sí mismas. Nos interesa cómo los algoritmos y los sistemas algorítmicos **hacen subjetividades**, es decir, los procesos mediante los cuales estas llegan a ser.

Al sintetizar los estudios de caso, esta sección articula la contribución del artículo como una extensión de la teoría de la subjetividad capaz de abordar las cualidades particulares que tipifican ese hacerse de las subjetividades algorítmicas.

Debido al carácter teórico de la contribución, la mayor parte de la sección recurre a bibliografía conceptual previa para articular estas cualidades. En numerosos casos, tales trabajos proporcionan vocabularios valiosos para describir partes del fenómeno y su funcionamiento. Sin embargo, pocos de esos conceptos —si es que alguno— permiten por sí solos dar cuenta de la totalidad. La combinación de cualidades característica de las

subjetividades que cobran vida dentro y alrededor de sistemas algorítmicos exige la multiplicidad de perspectivas teóricas movilizadas aquí.

Aunque el artículo se concentra principalmente en las «implicaciones para la teoría» [68], también atiende a aspectos de estos procesos relacionados con el diseño de sistemas algorítmicos. A diferencia de otros trabajos, no ofrecemos casos que indiquen exactamente qué diseños deben implementarse [por ejemplo, 197], ni directrices prescriptivas sobre cómo realizar el trabajo de diseño [por ejemplo, 9]. Proponemos, en cambio, orientaciones conceptuales capaces de sensibilizar a los diseñadores respecto del papel singular de los sistemas algorítmicos en la producción de subjetividades.

4.1. Inferir

La inferencia desempeña un papel central en los sistemas algorítmicos. Un sistema de moderación ajustado puede inferir qué contenidos serán probablemente denunciados como objetables. Un modelo de tópicos entrenado puede inferir los temas presentes en un documento nuevo. Es posible inferir numerosos atributos potencialmente sensibles de un usuario de redes sociales a partir de información aparentemente inocua [97, 170]. Como mostraron los casos anteriores, esta capacidad tiene consecuencias importantes. Sugerimos que tales predicciones inferenciales constituyen una característica definitoria del modo en que se hacen las subjetividades algorítmicas.

Trabajos previos [por ejemplo, 30, 77] han destacado que las actividades informacionales de etiquetar y clasificar son, simultáneamente, actos epistémicos de conocimiento —¿qué clase de cosa es este punto de datos?— y ejercicios de poder político. Clasificar sitúa el punto de datos dentro de una estructura de conocimiento existente, con su propia historia, compromisos valorativos y fundamentos políticos. Cuando aquello que se clasifica es una persona, especialmente si la clasificación es realizada por alguien o algo distinto de ella misma, el acto se convierte en un ejercicio de poder en la medida en que la define de maneras que pueden coincidir o no con su propia autodefinición.

Pero las inferencias algorítmicas van más allá de clasificar o etiquetar puntos de datos, humanos o no. Las subjetividades que cobran vida por medio de sistemas algorítmicos pueden no haber existido previamente como clase o etiqueta. En nuestros casos, un bloguero puede indicar implícita o indirectamente, mediante una manera particular de escribir, que sostiene cierta posición crítica frente a los criterios de autismo del DSM. Sería posible etiquetarlo como crítico sin modelado de tópicos —a partir de afirmaciones explícitas, por ejemplo—, pero el modelo hace cobrar vida a esa subjetividad crítica de una forma específica.

Esta animación algorítmica presenta la subjetividad como manifestada en patrones estadísticos de elección de palabras, sugiere que observar el uso lingüístico de una persona permitiría inferir hasta qué grado enactúa esa subjetividad y puede opacar o excluir otras interpretaciones posibles del mismo lenguaje. La interacción de los sistemas algorítmicos supera así la asignación de categorías ya existentes —antivacunas, trol [47], víctima [96], etc. — y produce subjetividades que parecen familiares, pero están al mismo tiempo fundadas en inferencias mecanicistas [cf. 58].

Estas inferencias mecanicistas suelen basarse no solo en datos directamente observables, sino también en los llamados **espacios latentes de características** [por ejemplo, 123]. En el caso del TEA, cada tópico —descrito como una distribución de probabilidad sobre palabras—

funciona como una dimensión de uno de esos espacios. Debido al significado semántico de las palabras, estas distribuciones son relativamente interpretables para una persona. Las palabras de alta probabilidad de un tópico —*espectro*, *trastorno*, *autismo*, *diagnóstico*, etc.— pueden leerse como una descripción breve de su contenido. Representar un documento en términos de esos tópicos latentes proporciona, de manera semejante, una descripción semántica humanamente interpretable. Las dimensiones se denominan «latentes» porque son inferidas en lugar de observarse directamente en los datos.

Muchos sistemas algorítmicos emplean espacios latentes, pero no todos son tan interpretables. Los grandes modelos de lenguaje (LLM) [por ejemplo, 63, 123] utilizan dimensiones latentes que con frecuencia carecen de una descripción evidente o semánticamente significativa. Las representaciones de palabras y documentos se transforman en un espacio de *embeddings*, denominado a veces «espacio semántico», compuesto habitualmente por unos cientos de dimensiones. Esto contrasta con los enfoques tradicionales de representación documental, que suelen utilizar una dimensión por cada palabra del vocabulario, es decir, decenas de miles de dimensiones.

Aunque estos espacios pueden contener más dimensiones latentes que un modelo de tópicos, reducen de manera significativa el número de dimensiones empleado para representar un documento. Los mapeos desde palabras o documentos hacia dichas representaciones se infieren optimizando iterativamente su capacidad para predecir tokens enmascarados —una palabra omitida de una oración— [63] sobre conjuntos textuales masivos, como rastreos de millones o miles de millones de páginas web. Las representaciones resultantes suelen reducir de manera considerable la dispersión general, lo cual contribuye a su mejor rendimiento en tareas posteriores de PLN: análisis de sentimiento, reconocimiento de entidades, clasificación textual y otras. Al mismo tiempo, disminuyen marcadamente la capacidad humana de determinar qué significa una dimensión latente concreta.

Esta dificultad interpretativa tiene diversas consecuencias. Explica en parte por qué se han desarrollado tantas técnicas destinadas simplemente a detectar sesgos en LLM [por ejemplo, 28, 84, 140, 143]: la escasa interpretabilidad del espacio de representación vuelve difícil advertirlos. También ayuda a comprender cómo puede dirigirse publicidad en línea a consumidores de cannabis aunque la plataforma no ofrezca «consumo de cannabis» como opción explícita de segmentación [29; véase también 125, 170]. En síntesis, los sistemas algorítmicos infieren no solo a partir de patrones estadísticos de datos observados, sino también mediante representaciones transformadas de esos datos que son a su vez inferidas y, por tanto, con frecuencia ilegibles para los humanos.

Las subjetividades algorítmicas, sin embargo, no están compuestas solamente por puntos de datos que analizar o modelos que manipular. Pensemos en los procesos mediante los cuales cobran vida subjetividades autistas a través de ensamblajes del DSM, prácticas clínicas, instituciones educativas y burocracias estatales. No es posible «ejecutar» esa subjetividad como un modelo para predecir la probabilidad de distintos resultados —por ejemplo, el modo en que determinadas opciones terapéuticas afectarían el nivel educativo futuro, el ingreso anual o la satisfacción vital—, ni calcular automáticamente su grado de ajuste con una persona concreta.

Las subjetividades algorítmicas no realizan, por sí mismas, predicciones o inferencias acerca del mundo. Pero la capacidad de efectuar predicciones cuantitativas —de inferir— se

convierte en una característica distintiva cuando se entrelaza con ensamblajes más amplios de personas, instituciones y estructuras de significado.

4.1.1. Diseñar para la inferencia

Estas capacidades inferenciales se vinculan con desafíos del diseño de sistemas algorítmicos [69, 114, 197, 198]. Los diseñadores deben anticipar no solo los resultados que el algoritmo hará aparecer durante una interacción, sino también los ecosistemas sociotécnicos más amplios en los que ese contenido puede emerger. Además, los propios modelos cambian en respuesta a flujos de datos en permanente transformación.

Dos líneas de pensamiento resultan útiles para comprender y diseñar alrededor de la inferencia. La primera proviene de trabajos sobre modernismo [159, 189], descrito a menudo mediante cuatro principios: calculabilidad, eficiencia, predictibilidad y control jerárquico —para una síntesis, véase [35, p. 950]—. La inferencia está estrechamente conectada con todos ellos: para inferir, los cuerpos deben volverse calculables; los modelos permiten predecir las acciones de una persona; y así sucesivamente. Comprender las subjetividades algorítmicas exige, por tanto, comprender cómo los cuerpos llegan a ser calculados, eficientizados, predichos y controlados jerárquicamente.

Considerar los sistemas algorítmicos como una empresa modernista ayuda a orientar la atención analítica y de diseño. El modelado de tópicos, por una parte, traduce perspectivas diferentes —como la crítica a las revisiones del DSM-5— en algo identificable mediante cálculo, en este caso a través de patrones estadísticos de coaparición de palabras. Estas inferencias permiten a la vez predecir cómo podrían comportarse o reaccionar determinados blogueros ante acontecimientos futuros, por ejemplo nuevas revisiones del DSM [de modo semejante a 86, 153].

Por otra parte, las predicciones habilitan múltiples posibilidades de diseño: desde herramientas que permitan a un bloguero reflexionar sobre el recorrido de su familia mediante el análisis de su escritura [similar a 32], hasta sistemas que propongan conexiones entre varios blogs para distintos propósitos [similar a 20]. Los diseñadores pueden atender a los principios del modernismo preguntándose qué cosas son —y cuáles no son— hechas calculables, eficientes, predecibles y controlables por cada posibilidad de diseño. Esto proporciona un lenguaje conceptual para dar cuenta de las diversas formas en que el diseño algorítmico interviene en la producción de subjetividades.

La segunda línea recurre a la noción de **sujeto escalable** [172]. Como refinamiento del «doble de datos» [87], Stark [172] destaca cómo las huellas digitales de una persona se emplean para crear modelos matemáticos y computacionales. Estos modelos y sus usos representan una confluencia singular entre ciencias psicológicas e informática, con consecuencias importantes tanto para la comprensión como para el control de las personas.

Stark recurre a ejemplos que van desde las pruebas A/B y el experimento de contagio emocional de Facebook [108] hasta herramientas de salud mental destinadas a asistir a pacientes con trastornos del estado de ánimo y la conducta, como Ginger.io. Todos estos casos, sostiene, asumen que las relaciones entre variables observadas en el agregado se aplicarán también a un individuo —la denominada «falacia ecológica» [144]—. Los sujetos escalables se crean, en parte, cuando observaciones de puntos de datos pasados —a menudo seres humanos y sus actividades— se utilizan para extraer conclusiones sobre puntos nuevos, es

decir, para inferir acerca de ellos. Son precisamente esas inferencias las que hacen posible la escalabilidad del sujeto escalable.

Este aparato conceptual [172] resulta útil, sugerimos, para razonar acerca de los medios y consecuencias de la inferencia y quizá diseñar en torno a ellos. Es evidente, por ejemplo, que la moderación de publicaciones y personas —como acosador, blanco, contenido tóxico, observador, etc.— se realiza mediante inferencias algorítmicas. La lente de la escalabilidad invita a que los diseñadores atiendan a los mecanismos por los cuales se producen esas inferencias.

En la moderación, una publicación podría proyectarse primero sobre un espacio latente de características. Su distancia[5] dentro de ese espacio serviría luego para inferir la probabilidad de que constituyera un caso de acoso o toxicidad. El mecanismo descansa sobre un supuesto de escalabilidad: que las manifestaciones de toxicidad producen publicaciones que, proyectadas en ese espacio, quedan geométricamente próximas. Concentrarse en los mecanismos permite preguntar si esa escalabilidad debería mantenerse en ese contexto o si existen otros modos de identificar contenidos tóxicos o intimidatorios.

Aunque la noción de sujeto escalable es analíticamente útil para examinar la inferencia y quizá diseñar alrededor de ella, ofrece menos orientación sobre la experiencia subjetiva individual. En efecto, «en las descripciones del agregado se pierden las maneras en que los sujetos individuales comprenden su propia escalabilidad» [172, p. 213]. Sin embargo, las subjetividades algorítmicas implican no solo experiencias subjetivas de individuos humanos, sino también sus interacciones con actores multiescalares. Comprenderlas requiere complementar el sujeto escalable con otros dispositivos teóricos capaces de dar cuenta de esos entrelazamientos.

4.2. Entrelazar

Las subjetividades algorítmicas cobran vida mediante interacciones entre y dentro de ensamblajes heterogéneos de actores y procesos. No afirmamos, por tanto, que los algoritmos produzcan subjetividades por sí solos. Un sistema de moderación puede marcar un contenido cargado; un modelo de tópicos puede asignar un tema significativo a un documento. Pero estos procedimientos computacionales no funcionan como entidades completamente independientes y separadas, ni en el análisis ni en la práctica.

Las subjetividades que nos interesan surgen a través de algoritmos imbricados o entrelazados [13, 79] en agregados heterogéneos de entidades, acciones e interpretaciones. Cobran vida, por ejemplo, mediante el entrelazamiento de algoritmos de moderación con procesos que articulan usuarios humanos, sistemas automatizados, organismos productores de políticas, trabajo humano mal remunerado, normas de interacción interpersonal aceptable y muchos otros elementos. No es el algoritmo *per se* el que anima las subjetividades algorítmicas, sino su funcionamiento como parte de estructuras siempre mucho más amplias y su entrelazamiento con ellas. Este entrelazamiento acompaña, además, la escala y el alcance extraordinariamente masivos en que tales subjetividades operan.

---

5 Pueden utilizarse distintas medidas de distancia o proximidad, entre ellas la distancia euclidiana, la similitud coseno, las métricas espectrales y muchas otras.

El entrelazamiento plantea al menos tres preocupaciones interconectadas. Todas implican, de diferentes maneras, preguntas acerca de cómo actores individuales se relacionan dentro de estos ensamblajes.

La primera pregunta es: ¿qué ocurre con la agencia? Si los actores individuales ya no pueden concebirse como entidades enteramente separadas, ¿cómo pensar la agencia con la que actúan? Latour [112] aborda dos afirmaciones contrapuestas sobre la regulación de armas. Una sostiene que «las armas matan personas»; la otra, que «las personas matan personas, no las armas». Ambas distribuyen de modo diferente la agencia y, por consiguiente, la responsabilidad. Latour resuelve la aparente tensión señalando que ninguna es completamente correcta. Al entrar en relación, el arma y la persona que la sostiene se combinan para convertirse en otro tipo de actor: «un ciudadano-arma, un arma-ciudadano» [112, p. 32]. Por ello, «no son ni las personas ni las armas las que matan»; «la responsabilidad de la acción debe compartirse entre los diversos actantes» [112, p. 34].[6]

Una lógica semejante permite comprender la agencia en los sistemas algorítmicos. El descenso de gradiente estocástico (SGD), algoritmo habitual para entrenar modelos de aprendizaje automático, se convierte en un actor diferente cuando se aplica a datos en los que contenidos acosadores o tóxicos fueron etiquetados. Los sistemas de moderación no matan a nadie como lo hace un arma, pero sí intervienen en el marcado de contenidos y la expulsión de usuarios. Decir que «el sistema» —o incluso el modelo entrenado— expulsa o marca oculta complejidades fundamentales. El modelo, el algoritmo SGD, los datos de entrenamiento, los moderadores humanos cuyo trabajo produjo esos datos, los programadores que construyen el sistema y las organizaciones —con frecuencia corporativas— que recolectan los datos y supervisan la operación no son responsables por separado, porque la agencia no pertenece a una sola de esas entidades. Como sugiere Latour, la responsabilidad debe distribuirse entre ellas.

Distribuir la responsabilidad plantea una segunda pregunta: en esta sopa entrelazada de actores, ¿cómo definimos entidades individuales? Si la responsabilidad se reparte entre «actantes», ¿cómo determinamos quiénes o qué son esos actantes? ¿Tiene siquiera sentido intentarlo? Una vía para responder recurre al concepto de entrelazamiento de Barad [13]. Las entidades, sostiene, no necesariamente existen como tales antes de sus relaciones. Se vuelven co-constitutivas unas de otras a través de aquello que Barad denomina **intra-acciones**.

Como ilustración, la física cuántica establece que es posible conocer la posición o el momento de una partícula, pero no ambos simultáneamente. Barad muestra que este relato refuerza una dualidad entre mundo —la partícula— y representación —las mediciones de su posición, momento, etc.—, al presuponer una partícula que ya posee ambas propiedades. En cambio, propone que antes de la observación «la partícula sencillamente no existía en ningún estado fijo» [99, p. 929], sino en una condición de potencialidad indeterminada. No significa que no existiera en absoluto, sino que aspectos de su existencia permanecían indeterminados.

Mediante el fenómeno de observación, la partícula, su posición o su momento —pero no ambos—, el aparato de medición y el observador humano cognoscente llegan a existir en una relación particular estabilizada. Barad denomina a esto **corte agencial**: el momento/proceso

6 Latour emplea «actante» para evitar supuestos antropomórficos implícitos acerca de cualquier entidad capaz de actuar.

por el cual entidades indeterminadas se estabilizan, aunque sea temporalmente, en relaciones mutuamente constitutivas.

Trabajos recientes, especialmente los de Frauenberger [79], han propuesto que las teorías del entrelazamiento ofrecen una nueva metáfora generativa para la HCI. El entrelazamiento permite reconceptualizar de manera fundamental la «interacción»: «las cosas y las personas, como fenómenos, se constituyen mutuamente mediante su intra-acción; es decir, los límites entre humanos y máquinas no están predeterminados, sino que son enactuados» [79, p. 9]. Así como una partícula es llevada a un estado fijo —y excluida de otros— mediante su intra-acción con un aparato de medición, un observador y otros componentes, un sistema sociotécnico existe en un estado de indeterminación hasta que se entrelaza de una manera específica.

Frauenberger ilustra esta cuestión con *Flow*, un dispositivo hipotético que «muestra la comodidad o la ansiedad de los participantes de una conversación a partir de datos de diversos sensores» [79, p. 12]. Diferentes modos de evaluarlo enactúan cortes agenciales distintos: «un estudio mediante entrevistas convertirá a Flow en un artefacto cultural; una prueba controlada con usuarios en el laboratorio, en una herramienta funcional; y un estudio longitudinal mediante diarios podría convertirlo en un sentido artificial de las personas» [79, p. 15].

Este pensamiento ayuda a dar cuenta de las complejas relacionalidades de los sistemas algorítmicos. El acosador, la víctima, el contenido tóxico y quienes observan llegan a constituirse como entidades con determinadas propiedades y relaciones a raíz de los cortes agenciales realizados por los sistemas de moderación. Esto no significa que los observadores no existieran antes del marcado del contenido. Estaban allí, pero asumieron el papel y entraron en las relaciones de «observadores», con ciertas propiedades y no otras posibles, en parte mediante las intra-acciones del sistema de moderación con ellos y con los demás actores.

Aplicar estas nociones de indeterminación a los sistemas algorítmicos abre una tercera pregunta: ¿cómo pensar las dinámicas de poder y el ejercicio de autoridad? Couldry y Mejias [53] sostienen que la recolección y el análisis de datos operan mediante dinámicas de poder semejantes al colonialismo. Del mismo modo que los gobiernos coloniales apropiaban tierras, cuerpos y recursos naturales para maximizar beneficios, las empresas tecnológicas cuantifican la interacción social como datos para extraer valor. Así como el capitalismo transformó históricamente la actividad humana en la mercancía trabajo, el colonialismo de datos transforma la experiencia de la socialidad humana en la mercancía dato. En esta formulación, las empresas que recolectan y analizan datos se convierten en los principales enactores de una autoridad impulsada por datos.

Burrell y Fourcade [38] sugieren de manera similar que el poder de efectuar estas determinaciones reside en aquello que denominan la **élite de la codificación**:

> *«La élite de la codificación es una nebulosa de desarrolladores de software, directores ejecutivos de empresas tecnológicas, inversores y profesores de informática e ingeniería, entre otros, que a menudo circulan sin esfuerzo entre estos papeles influyentes […]. Las personas más valoradas en este mundo son las que tocan y comprenden el código informático. Las más poderosas son quienes poseen el código y pueden contratar a otros para desplegarlo según sus criterios» [38, p. 217].*

Si «el código es ley» [115], quienes escriben el código hacen las leyes y, por ello, poseen el poder de gobernar.

Pero desde la lente del entrelazamiento, la élite de la codificación no existe como entidad prefigurada. Se constituye de ese modo por sus interconexiones dentro de sistemas mucho más amplios de «empresas emergentes, [...] grandes firmas, laboratorios de investigación financiados por el Estado, aulas» y otros espacios [38, p. 217]. El enfoque del entrelazamiento reconoce su poder considerable, pero sostiene que no deriva únicamente de la capacidad de escribir y desplegar código ni de recolectar datos y extraerles valor [53]. La élite llega a ser tal mediante sus intra-acciones con sistemas algorítmicos, donde el diseño y la implementación constituyen modalidades de intra-acción. La élite, los observadores y los sistemas de moderación son mutuamente co-constitutivos, junto con sus propiedades, incluidas las diferencias de poder.

Las cuestiones de poder no se reducen, entonces, a determinar quién realiza los cortes agenciales. No deberíamos afirmar que la élite de la codificación es responsable de crear observadores u otras posiciones de sujeto en situaciones de acoso en línea. Debemos preguntar, más bien, cómo se configuran las condiciones de posibilidad. Según Barad [13], antes de entrar en entrelazamientos co-constitutivos, las entidades residen en un estado de indeterminación. Pero esa indeterminación no debe concebirse, por así decirlo, como una distribución previa uniforme.[7] Una entidad no tiene idéntica probabilidad de constituirse como acosador o como víctima en una situación determinada. Los detalles de los ensamblajes sociotécnicos operantes —pesos de vectores de características, normas y políticas de moderación, conjuntos de entrenamiento, denuncias de usuarios, etc.— configuran la mayor o menor probabilidad de que las entidades se co-constituyan de ciertos modos.

Dar cuenta de las dinámicas de poder exige examinar cómo las configuraciones actuales de cuerpos, tecnologías y organizaciones hacen que diversas configuraciones futuras sean más o menos probables. Dado un sistema particular de moderación y sus antecedentes, ¿quién tiene más probabilidades de ser constituido como acosador, observador o víctima? ¿Qué procesos aumentan o disminuyen esas probabilidades? Alcanzar la rendición de cuentas necesaria para abordar estas preguntas requiere, entre otras cosas, estrategias que utilicen el entrelazamiento para orientar el diseño de sistemas algorítmicos.

4.2.1. Diseñar para el entrelazamiento

Retomemos un punto anterior: las subjetividades no remiten aquí a conceptos internos o mentales del sí-mismo individual; tampoco nacen únicamente de encuentros emocionales con un modelo de tópicos, un sistema de moderación [181] o una publicación de Facebook [36]. Utilizamos «subjetividades algorítmicas» para destacar cómo cuerpos transitorios o mutables llegan a existir mediante los impulsos, repliegues y sintonizaciones afectivas [100] que los algoritmos hacen posibles. Las controversias por la clasificación del autismo en el DSM-5 desencadenan indudablemente respuestas subjetivas muy personales, y las familias están profundamente afectadas por cambios textuales sutiles entre versiones del manual.

---

7 En estadística bayesiana, una distribución previa describe la probabilidad de los valores posibles de una variable antes de incorporar observaciones. En una distribución previa uniforme, todos los valores son igualmente probables.

Nuestro punto, sin embargo, es que a medida que los algoritmos ocupan papeles más prominentes en estos ensamblajes, emerge una naturaleza correspondientemente algorítmica en las subjetividades probables o incluso posibles. Las operaciones de cálculo de un modelo de tópicos, que traducen y enumeran lo escrito por las personas, proporcionan un registro afectivo para comprender las controversias del DSM. A través de sus interacciones lingüísticas en una plataforma de discusión, los actores definen y son definidos por el potencial calculable de determinados tópicos y por la posibilidad de computar respuestas afectivas o subjetivas a través de ellos. Prestamos atención a los algoritmos por la escala de los ensamblajes que contribuyen a constituir y por el poder que ejercen dentro de esos entrelazamientos [5].

Estas consideraciones conceptuales sugieren formas de adaptar los métodos de diseño. Pensemos en las **personas** y los **escenarios** [42, 50, 134, 146]. Habitualmente, una persona se define mediante una combinación de atributos individuales: datos demográficos, objetivos profesionales, intereses personales, formación, vínculos familiares, etc. Nielsen [134], por ejemplo, describe una persona de salud electrónica llamada «Gitte»:

> *«40 años. Médica generalista. Madre de dos hijos de cinco y siete años. Casada con un académico. Vive en una ciudad provincial grande. Trabaja en una práctica médica compartida. […] Asiste a un coro una vez por semana y corre regularmente en un club deportivo. Es miembro de un club de lectura. Mantiene una actitud conservadora/minimalista hacia la tecnología y las herramientas profesionales en general. […] Es inteligente, delgada y procura vestirse con prendas de buen diseño, joyería exquisita y el cabello recién cortado» [134, pp. 178-179].*

Descripciones como esta concentran implícitamente la atención del diseñador en atributos individuales. Podríamos, en cambio, articular una persona menos por esos atributos que por sus entrelazamientos. ¿Con qué entidades tecnológicas, políticas, culturales, económicas y sociales se encuentra entrelazado ese individuo? ¿Qué colecciones heterogéneas de intra-acciones producen esos vínculos? ¿Qué posibilidades existían antes de ellas? ¿De qué manera la persona está sujeta a esos entrelazamientos y es a la vez constitutiva de ellos? ¿Cómo «Gitte», como actor, se co-constituye con la práctica médica donde trabaja o con el coro en el que canta? Este enfoque probablemente exigiría articular no solo una serie de personas individuales, sino las relaciones entre ellas.

Del mismo modo, el alcance de un escenario se desplaza desde una interacción particular hacia una intra-acción específica: los cortes agenciales mediante los cuales las entidades involucradas adquieren el estado de esas entidades particulares —y no de otras— con esas propiedades y relaciones —y no con otras posibles—.

La noción de entrelazamiento podría incluso desestabilizar la idea misma de **centrar** el diseño. La HCI suele describirse como «centrada en el usuario» o «centrada en lo humano» [17, 75, 152, 177]. Nuestra concepción abre al menos dos posibilidades. Primero, los diseñadores podrían resistir desde el comienzo estos centramientos y explorar el entrelazamiento como locus conceptual del diseño: una HCI «centrada en el entrelazamiento» [cf. 79]. Segundo, podrían resistir el concepto de centramiento en cuanto tal. Elegir cualquier concepto o entidad —humano, usuario, entrelazamiento— como centro disminuye implícitamente otros centros posibles. El entrelazamiento podría orientar un diseño «descentrado» que no otorgue primacía a ninguna entidad o concepto. Ambas opciones resuenan con llamados a enfoques de diseño más-que-humanos [184, 185].

Esto no significa que los seres humanos carezcan de importancia. Proponemos desplazar el foco desde atributos específicos de entidades individuales hacia entrelazamientos sistémicos. Las preguntas del diseñador cambiarían entonces de «¿qué tareas facilita o dificulta este diseño?» a «¿qué entrelazamientos singulares vuelve más o menos probables?»; de «¿cuán útiles, placenteras, confusas o gratificantes son las interacciones de los usuarios?» a «¿qué papeles desempeña el diseño al entrelazarlos con estructuras y sistemas sociotécnicos más amplios?»; y de «¿cómo hará reaccionar esta decisión a los usuarios?» a «¿qué formas de ser permitirá o excluirá?». No tomamos una posición definitiva acerca de si el entrelazamiento debe ocupar el centro del diseño o constituir una consideración entre muchas en un diseño descentrado.

En ambos casos, estos desplazamientos pueden ayudar a resistir el impulso de tratar a humanos, usuarios o comunidades como entidades prefiguradas. En lugar de dar por supuesto al humano como centro, entender el diseño como entrelazamiento llama la atención sobre el modo en que el diseño —como artefacto técnico y como proceso discursivo— co-constituye la humanidad de esos humanos. Esto anuncia otro desplazamiento potencialmente fundamental, referido al «humano» de la HCI.

4.3. Humanizar

Como hemos mostrado, la HCI desarrolló maneras cada vez más matizadas de hablar de los humanos y sus interacciones con las computadoras [por ejemplo, 17, 75, 163, 177]. Al mismo tiempo, los enfoques centrados en lo humano postulan implícitamente «humano» como una categoría prefigurada. La perspectiva de este artículo propone atender a los procesos mediante los cuales llegan a existir las categorías de humano, algoritmo y otras semejantes.

Diversos trabajos sugieren que la manera en que una máquina «piensa» difiere fundamentalmente de la manera en que «piensa» un humano, o al menos que sus capacidades y procesos difieren en escala y afecto [2, 176]. Burrell [37, p. 6], por ejemplo, muestra que una red neuronal entrenada para reconocer dígitos manuscritos no descompone la tarea en subtareas inteligibles para las personas, tales como identificar una barra horizontal, una forma oval cerrada o una línea diagonal. Las características visuales a las que atiende parecen extrañas y casi irreconocibles para un observador humano. Burrell admite que también existe «una forma de opacidad en el proceso humano —en gran medida subconsciente— de reconocer caracteres» [37, p. 7]. Sin embargo, pocas personas que intenten descifrar una cifra escrita con mala caligrafía dirían que buscan los patrones amorfos que utiliza la red.

Figura 1. Visualización de las características utilizadas por una red neuronal entrenada para reconocer dígitos manuscritos [37]. Burrell [37] sostiene que estos patrones se parecen muy poco a aquello en lo que reparan los humanos al realizar la misma tarea.

Los estudios de caso anteriores ofrecen ejemplos semejantes. El modelado de tópicos representa las relaciones entre temas como mezclas de distribuciones de probabilidad, pero esta representación guarda escaso parecido con la manera en que un lector humano encuentra temas en un corpus [véase también 156, 179]. La moderación de foros y redes puede depender del acoplamiento entre humanos y procesos algorítmicos, pero ese acoplamiento produce fricciones. Para las máquinas, la aplicación de estándares y políticas se decide mediante umbrales. Así, dunas de arena pueden ser marcadas como desnudez [128] y

fotografías periodísticas como pornografía infantil [178], errores más probables en una moderación puramente algorítmica que con moderadores humanos. Las personas, en cambio, se ven obligadas a leer e interpretar relatos de vidas concretas; las interacciones personales con un usuario pueden orientar exactamente qué sanción decide aplicar un moderador [169].

A primera vista, estos ejemplos parecerían demostrar dos categorías claras —humano y máquina—, cada una con su estilo de pensamiento. Proponemos, en cambio, que el funcionamiento de los sistemas algorítmicos contribuye, casi paradójicamente, a definir qué constituye la categoría de humano. Esto sucede al menos de dos maneras.

Primero, los sistemas codifican y enactúan definiciones de «humano». La bibliografía previa se concentra sobre todo en la dimensión discursiva de esta operación. Keyes [106], por ejemplo, investiga valiosamente la IA aplicada al diagnóstico del autismo, que suele tratar la conducta social y las habilidades comunicativas —o la ausencia inferida de ambas— como indicadores. Dicho de otro modo, «el autismo es tratado como opuesto a los rasgos que "hacen" que una persona sea una persona» [106, p. 14]. Estos sistemas codifican así una formulación del autismo —y de la humanidad— en la que «las personas autistas son representadas como asociales, fundamentalmente carentes de capacidad para conocer y comprender y, por consiguiente, carentes de agencia y de condición de persona» [106, p. 3].

Sugerimos que lógicas semejantes operan tanto en los detalles técnicos de implementación como en el funcionamiento práctico. El usuario de Facebook se define tanto por sus interacciones con entidades de la plataforma —amistades, organizaciones, anuncios, etc.— como por la agregación de sus «Me gusta». Del mismo modo, los sistemas de recomendación codifican al usuario principalmente como un actor económico que intenta maximizar la selección de objetos de acuerdo con sus gustos y preferencias [168]. Estos y otros sistemas construyen —y reconstruyen continuamente— qué significa ser humano.

Segundo, la propia noción de inteligencia artificial postula determinados tipos de inteligencia como «artificiales» [23, 110, 124, 158]. Esta terminología sugiere una manera de pensar específicamente artificial o algorítmica. Al nombrarla así, definimos implícitamente el pensamiento y el ser humanos como diferentes, quizá incluso opuestos, al pensamiento y al ser algorítmicos. La operación de estos sistemas define entonces lo humano mediante un contraejemplo.

Esto no significa que la distinción entre algoritmos y humanos sea innecesaria. No somos objetos homogéneos [94] ni actantes uniformes dentro de una red [113]. La perspectiva de las subjetividades algorítmicas propone, más bien, no aceptar estas categorías como prefiguradas y atender a cómo **se hacen**, es decir, a los procesos mediante los cuales llegan a establecerse diferencias entre lo algorítmico y lo humano.

4.3.1. Diseñar para el significado humano

El funcionamiento de los sistemas algorítmicos está profunda e inextricablemente entretejido con los valores humanos incorporados en su diseño y con aquellos que se interpretan a partir de ellos [22, 23]. Como el cyborg de Haraway [92], los ensamblajes resultantes son creaciones híbridas, simultáneamente familiares y extrañas. Incorporan elementos que en la superficie reconocemos —uso del lenguaje, datos demográficos, conexiones sociales—, pero que dentro de contextos algorítmicos adquieren significados y relevancias sutilmente diferentes.

Para Haraway y para trabajos poshumanos o más-que-humanos posteriores [por ejemplo, 31, 100, 175], esta hibridez invita a una lectura situada de significados, valores, cuerpos subjetivos y entrelazamientos. Patrones de uso lingüístico —determinados coloquialismos, insultos racistas o combinaciones de emojis— pueden incrementar la probabilidad de que un sistema de moderación marque un contenido [95, 166]. Pese a su similitud superficial, esas características significan cosas diferentes para un clasificador y para una persona que lee o escribe. Aquello que un humano interpreta como connotación de acoso, el algoritmo lo codifica como pesos dentro de un vector de características.

El modelado de tópicos ofrece un ejemplo privilegiado. En su núcleo se basa en contar. Sus distribuciones de probabilidad se ajustan al conjunto de datos según patrones en la cantidad de veces que grupos de palabras coaparecen en documentos. Los números importan —«cuentan», por así decirlo—, pero puede resultar casi absurdo afirmar que la palabra más frecuente de un documento es también la más importante [141]. De hecho, la palabra más habitual en casi cualquier texto en inglés es *the* [56]. Esta desconexión entre frecuencia e importancia motiva el uso de listas de **palabras vacías** o *stop words* [149]: términos tan frecuentes que transmiten escaso significado aislado y se omiten durante el procesamiento [119, p. 27]. Aunque estas listas son más comunes o extensas en ciertas técnicas que en otras, revelan una divergencia fundamental entre los conteos y distribuciones del modelo y los significados atribuidos por lectores humanos [véase también 155].

Además, las implementaciones computacionales tratan las palabras no como unidades significativas *per se*, sino como tokens únicos. Para un modelo de tópicos, «autismo» y «autista» son tan diferentes como «manzana» y «naranja».[8] Una aparición de «autismo» solo tiene significado en cuanto afecta la inferencia de las distribuciones probabilísticas subyacentes. Existe, por tanto, una desconexión fundamental entre el procesamiento computacional del lenguaje y su interpretación humana.

Podría ser productivo hacer visibles estas desconexiones durante el diseño. Al desarrollar un sistema interactivo basado en modelado de tópicos [por ejemplo, 20, 66], incorporar los resultados a prototipos tempranos puede ayudar a diseñadores, usuarios y participantes de codiseño a interpretarlos. Pero también puede generar la impresión de que el sistema comprende de manera humana las relaciones entre palabras como «autismo» y «autista». Una alternativa sería acompañarlas con el identificador aleatorio que el modelo asigna a cada token: por ejemplo, «token2475-autismo» y «token1382-autista». Esto resaltaría de forma simple la diferencia entre representación computacional e interpretación humana sin impedir una retroalimentación productiva.

Los detalles de esta diferencia son específicos del modelado de tópicos, pero el problema se aplica, con otras particularidades técnicas, a los demás enfoques computacionales del lenguaje.

Mientras el artículo se preparaba para su publicación, se desarrollaron modelos informados por la semántica distribucional [26, 27] y los mecanismos de atención [183] que abordaban parte de estas dificultades. En los LLM y sus aplicaciones —ChatGPT [139] como ejemplo destacado, pero también *word embeddings* composicionales como word2vec [123],

8 Técnicas como la derivación o la lematización pueden representar «autismo» y «autista» como relacionados, aunque hacerlo suele reducir la calidad de los resultados [167].

representaciones bidireccionales de codificadores transformer [63] y otros transformadores generativos preentrenados [148]— la representación de un token depende del contexto en que aparece, es decir, de los tokens próximos. Además, los tokens no suelen coincidir con palabras naturales completas, sino con segmentos subléxicos. «Autism» podría representarse mediante «aut» e «ism», mientras que «autistic» podría dividirse en «aut», «ist» e «ic». A diferencia del modelado tradicional, estos modelos producen representaciones muy próximas para pares semejantes.

Esto se logra mediante tokenización subléxica y espacios semánticos de dimensionalidad moderadamente alta, semejantes a los espacios latentes antes descritos [63, 123]. Una palabra, segmento, oración o documento puede representarse como vector en ese espacio. La similitud se calcula mediante el coseno del ángulo entre vectores. Las representaciones se aprenden entrenando sobre corpus masivos —miles de millones de palabras procedentes de rastreos web, libros y otros textos—, de modo que las palabras utilizadas en contextos similares queden ubicadas en regiones próximas del espacio semántico.

A primera vista, este enfoque aprovecha hábilmente la afirmación del primer Wittgenstein [193] según la cual una palabra solo tiene significado en el contexto de un enunciado o proposición lógica. El Wittgenstein tardío [194], sin embargo, desarrolló la noción de juegos de lenguaje: una palabra, oración o documento obtiene su significado de su **uso** en el contexto de una actividad humana. La palabra inglesa «Fly!» significa cosas radicalmente distintas cuando la pronuncia un cliente irritado levantando una copa de Chablis —«¡mosca!»—, un lanzador en un campo de béisbol —«¡elevada!»— o un mago gris colgando de un precipicio —«¡huyan!»—. El problema supera la polisemia y la desambiguación léxica, así como el hecho de que cambien las palabras circundantes. Los procesos mediante los cuales los humanos construyen y contrastan activamente esos significados [151] guardan probablemente poca semejanza con el cálculo de cosenos entre representaciones vectoriales de tokens subléxicos.

Sin embargo, son esas distancias vectoriales las que pueden decidir si un contenido es marcado para moderación. Por ello, las investigaciones futuras sobre experiencias alrededor de sistemas algorítmicos deben considerar explícitamente la relación entre la pluralidad de significados, los mecanismos computacionales que contribuyen a producirlos y las fuentes de agencia implicadas en su enactuación [15, 65, 165, 195]. Como escribe Barad [13, p. 353], «los fenómenos —sean lagartos, electrones o humanos— existen únicamente como resultado y como parte de la intra-actividad continua del mundo, de su diferenciación dinámica y contingente en relacionalidades específicas».

Ningún fenómeno o entidad puede tratarse reductivamente como aislado ni observarse desde fuera de su contexto. Debe comprenderse como un devenir continuo a través de relaciones entrelazadas. Nuestro análisis se concentra en los sistemas algorítmicos y en cómo las relaciones entre distribuciones de probabilidad, pesos de vectores, etiquetas, ejercicios de poder y atribuciones de significado desempeñan un papel singular en la coproducción de la categoría de «humano».

Atender a estas dinámicas es tan importante para quienes investigan sistemas algorítmicos como para quienes los diseñan. Los investigadores deberían detenerse antes de interpretar los componentes de estos sistemas del mismo modo en que los humanos interpretamos elementos similares en contextos no algorítmicos. Una característica léxica muy informativa para un clasificador de toxicidad no posee necesariamente el mismo significado que para una

persona que lee esa palabra en una publicación. Los diseñadores, por su parte, pueden y deberían atender a las maneras en que el funcionamiento interno de los sistemas define implícitamente qué constituye la humanidad. Las características usadas para curar un muro de noticias no solo influyen en nuestra percepción de cercanía con ciertas personas [72]; también pueden reconfigurar cómo concebimos la constitución, enactuación y performance de la cercanía humana.

5. Implicaciones y conclusión

Una reacción intuitiva ante muchos problemas que los algoritmos hacen aparecer sería retroceder [a la manera de 21] y reducir hasta qué punto se entrelazan con la producción de subjetividades. Podríamos repensar la distribución del trabajo e intentar devolver a los humanos el peso de la agencia. Las soluciones más tecnocéntricas, en cambio, proponen aumentar la transparencia algorítmica y construir herramientas que expliquen cómo se producen clasificaciones y decisiones [67, 164]. Estas sugerencias se apoyan en la existencia de una perspectiva neutral, objetiva, semejante al «ojo de Dios» [90]: el supuesto de que es posible eliminar los sesgos de los esquemas clasificatorios [28] y de las elecciones que estos ofrecen [52].

Ninguna de ambas respuestas toma realmente en serio la proliferación de procesos que intentamos describir. Presuponen mundos en los que humanos y máquinas son separables y donde las decisiones pueden adoptarse a partir de datos o hechos desprendidos de la experiencia humana situada. Nuestros casos, junto con una extensa historia de investigación, muestran que ya siempre proliferamos en mundos que entrelazan inexorablemente lo humano y lo maquínico. Definir cuerpos y experiencias nunca carece de una política acerca de qué cuenta como normativamente valioso. Prohibir el discurso de odio y a quienes lo producen, y entrenar máquinas para hacerlo, implica participar en los procesos que definen el odio y la manera en que se lo performa.

Nuestra propuesta final tiene dos partes y abre, creemos, un espacio para la investigación y el diseño en HCI. En primer lugar, debemos atender a los mundos que están siendo posibilitados mediante las subjetividades algorítmicas. En segundo lugar, debemos atender a los mundos que querríamos hacer posibles.

Esto exige más que reemplazar «usuario» por «sujeto» o «subjetividad» [14, 17]. Requiere una reorientación general: dejar de ver al usuario como dado y comprenderlo —junto con numerosas relaciones adicionales— como algo coproducido continuamente en y mediante el diseño. La HCI debería atender a los procesos concretos por los que ocurre esta coproducción.

El artículo no ofrece un punto final conclusivo ni instrucciones prescriptivas precisas. Pretende aportar un avance sobre el cual otros puedan construir conceptual, empírica, metodológica y teóricamente. La lista de cualidades presentada probablemente no sea definitiva. Existen numerosos sistemas no abordados directamente: visión computacional y reconocimiento facial [170, 173], asignación automática de tareas en el trabajo de plataformas [6, 190, 196], curaduría de muros de redes sociales [71, 72], evaluación y predicción de riesgo en justicia penal [33, 48], entre muchos otros. Todos constituyen ámbitos relevantes para desarrollar las contribuciones conceptuales del artículo.

La palabra **desarrollar** es crucial. Entre estos contextos habrá diferencias en los detalles técnicos de la inferencia, en los entrelazamientos entre clasificación y algoritmos y en la co-constitución de la categoría «humano». Los trabajos futuros deberán explorar la adecuación y los conocimientos producidos por métodos diversos, tanto tradicionales como novedosos. También existirán otras cualidades, además de inferir, entrelazar y humanizar, hacia las cuales investigadores y diseñadores dirigirán productivamente su atención.

La primera parte de nuestra propuesta de cierre consiste, en cualquier caso, en preguntar: **¿qué subjetividades cobran vida en este caso particular y mediante qué entrelazamientos relacionales se hicieron posibles?**

No convocamos a un estudio específico ni a una serie cerrada, sino a un programa amplio de investigación. Es improbable que este programa gire alrededor de una única conceptualización fija de las subjetividades algorítmicas. Presentar una formulación definitiva socavaría la capacidad de trabajos futuros para dar cuenta del carácter continuo de estos entrelazamientos. Las maneras siempre cambiantes en que los sistemas penetran nuevas facetas de la vida pública y privada exigen una evolución igualmente permanente de nuestros conceptos. Ningún trabajo aislado ofrecerá una explicación definitiva o completamente exhaustiva; deberá articular cómo su conexión con el programa más amplio desarrolla nuestra comprensión de estos devenires continuos.

Además de estudiar los sistemas y sus subjetividades, buscamos llamar la atención sobre sus relaciones con el diseño. Si aceptamos que el diseño de algoritmos e interfaces participa en el hacer de los cuerpos —en hacer cobrar vida a ciertas subjetividades—, debemos reconocer que podemos considerar qué otras subjetividades deseamos producir y cómo. Evitamos prescribir cuáles serían preferibles. Ofrecemos herramientas conceptuales para explorar esas preguntas.

El cyborg de Haraway [92] proporciona de nuevo un ejemplo instructivo. Puede leerse legítimamente como un imaginario tecnocéntrico que amplifica versiones muy masculinizadas y occidentales del progreso tecnocientífico. Pero la pregunta célebre de Haraway es qué otros imaginarios podrían ser posibles. ¿Qué posibilidades técnicas e infraestructurales del cyborg abren condiciones para formas más equitativas y distribuidas de florecimiento global y modesto?

Análogamente, esperamos que nuestro trabajo proporcione a los diseñadores un salto especulativo que invite a preguntar: **¿qué más?** La segunda parte de la propuesta consiste en interrogar qué otras relaciones y subjetividades podrían recibir aquí una oportunidad. ¿Cómo diseñar de maneras más abiertas —que permitan una variedad mayor de entrelazamientos— y más generosas —que reconozcan que todos los actores han intervenido en el resultado—? ¿Cómo puede el diseño [a la manera de 186] abrir espacios para imaginar y enactuar otros mundos posibles [116, 162]?

Despret [59, p. 129] concibe este encuadre expansivo y más-que-humano de agencias y subjetividades entrelazadas como «una aventura en cuyo transcurso las subjetividades se superponen, se transforman, se actualizan y se extienden». La subjetividad nunca queda aislada en el individuo, sino atrapada en el intercambio entre actores humanos y no humanos. Sostenemos que incluso considerar —y mucho más perseguir— otros imaginarios exige aplicar perspectivas semejantes a la comprensión y el diseño de subjetividades algorítmicas.

Para la HCI, esto abre la posibilidad de imaginar las condiciones que podrían ofrecer los sistemas que construimos. Cerramos con tres sugerencias.

Primero, las subjetividades algorítmicas constituyen un área importante de investigación empírica. Los estudios deben atender no solo a experiencias subjetivas vividas, sino a cómo los entrelazamientos de actores heterogéneos actúan conjuntamente para hacer posibles ciertas subjetividades y no otras. La pregunta no es únicamente qué subjetividades anima un sistema, sino **cómo hace subjetividades**: mediante qué procesos infiere, entrelaza y humaniza experiencias y entidades dentro de ensamblajes más amplios.

Segundo, ese trabajo empírico puede desarrollar iterativamente una comprensión colectiva del papel de las decisiones de diseño. Aquí «diseño» no designa solo interfaces e interacciones, el foco habitual de la HCI. Incluye los esquemas clasificatorios en que se basan los sistemas, los formalismos matemáticos que codifican modelos y las políticas organizacionales o estatales que determinan cómo y por quién pueden utilizarse. Ampliar de esta manera el alcance del diseño es necesario para tomar en serio las implicaciones de las subjetividades algorítmicas.

Tercero, el objetivo central debe ser imaginar cómo las cosas podrían ser distintas. ¿Qué elecciones diferentes podrían realizarse en la selección de características, el uso de espacios latentes, los hiperparámetros, el ajuste y la optimización? ¿Cómo volverían esas decisiones menos probables ciertos mundos y más probables otros? ¿Cómo construir sistemas que no solo configuren las interacciones de las personas, sino también los imaginarios que hacen posibles los mundos?

Comprender cómo los múltiples aspectos del diseño hacen cobrar vida a subjetividades algorítmicas alcanza su mayor impacto cuando permite ver —y quizá enactuar— esos otros mundos. Esperamos que este artículo facilite el trabajo de quienes imaginen condiciones alternativas de posibilidad.

Agradecimientos

Los autores agradecen a Jeff Bardzell, Shaowen Bardzell, Michael Ann DeVito, Jessica Fueston y a los revisores anónimos por sus comentarios constructivos sobre versiones anteriores.

Referencias

Las referencias se conservan en su idioma original para mantener la precisión bibliográfica y facilitar su búsqueda.

[3] Sara Ahmed. 2006. Queer Phenomenology: Orientations, Objects, Others. Duke University Press, Durham, NC. [4] Sara Ahmed. 2010. The Promise of Happiness. Duke University Press, Durham, NC. [5] Ali Alkhatib and Michael Bernstein. 2019. Street-Level Algorithms: A Theory at the Gaps between Policy and Decisions. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). ACM, New York, NY, 530. DOI: https://doi.org/10.1145/3290605.3300760 [6] Ali Alkhatib, Michael S. Bernstein, and Margaret Levi. 2017. Examining Crowd Work and Gig Work Through the Historical Lens of Piecework. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). ACM, New York, NY, 4599–4616. DOI: https://doi.org/10.1145/3025453.3025974 [7] American Psychiatric Association. 1994. Diagnostic and Statistical Manual of Mental Disorders: DSM-IV (4th. ed., 7th. print ed.). American Psychiatric Association, Washington, DC. [8] American Psychiatric Association. 2000. Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR (4th. ed., text revision ed.). American Psychiatric Association, Washington, DC. [9] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). ACM, New York, NY, 1–13. DOI: https://doi.org/10.1145/3290605.3300233 [10] Mike Ananny. 2011. The Curious Connection Between Apps for Gay Men and Sex Offenders. The Atlantic (April 2011). [11] Emiliana Armano, Marco Briziarelli, Joseph Flores, and Elisabetta Risi. 2022. Platforms, Algorithms and Subjectivities: Active Combination and the Extracting Value Process – An Introductory Essay. In Digital Platforms and Algorithmic Subjectivities, Vol. 24. Emiliana Armano, Marco Briziarelli, and Elisabetta Risi (Eds.), University of Westminster Press, 1–18. Retrieved from https://www.jstor.org/stable/j.ctv319wpvm.4 [12] Karen Barad. 2003. Posthumanist Performativity: Toward an Understanding of How Matter Comes to Matter. Signs 28, 3 (2003), 801–831. DOI: https://doi.org/10.1086/345321 [13] Karen Barad. 2007. Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning. Duke University Press, Durham, NC. [14] Jeffrey Bardzell and Shaowen Bardzell. 2015. The User Reconfigured: On Subjectivities of Information. In Decennial Aarhus Conference on Human Centered Computing. Aarhus, Denmark, 133–144. DOI: https://doi.org/10.7146/aahcc. v1i1.21298 [15] Shaowen Bardzell. 2010. Feminist HCI: Taking Stock and Outlining an Agenda for Design. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). ACM, New York, NY, 1301–1310. [16] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. California Law Review 104, 3 (2016), 671–732. DOI: https://doi.org/10.15779/Z38BG31 [17] Eric P. S. Baumer and Jed R. Brubaker. 2017. Post-userism. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). ACM, New York, NY, 6291–6303. DOI: https://doi.org/10.1145/3025453.3025740 [18] Eric P. S. Baumer and Micki McGee. 2019. Speaking on Behalf of: Representation, Delegation, and Authority in Computational Text Analysis. In Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES). ACM, New York, NY, 163–169. DOI: https://doi.org/10.1145/3306618.3314292 [19] Eric P. S. Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K. Gay. 2017. Comparing

Grounded Theory and Topic Modeling: Extreme Divergence or Unlikely Convergence? Journal of the Association for Information Science and Technology (JASIST) 68, 6 (June 2017), 1397–1410. DOI: https://doi.org/10.1002/asi.23786 [20] Eric P. S. Baumer, Drew Siedel, Lena McDonnell, Jiayun Zhong, Patricia Sittikul, and Micki McGee. 2020. Topicalizer: Reframing Core Concepts in Machine Learning Visualization by Co-Designing for Interpretivist Scholarship. Human– Computer Interaction 35, 5–6 (April 2020), 452–480. DOI: https://doi.org/10.1080/07370024.2020.1734460 [21] Eric P. S. Baumer and M. Six Silberman. 2011. When the Implication Is Not to Design (Technology). In Pro- ceedings of the ACM Conference on Human Factors in Computing Systems (CHI). New York, NY, 2271–2274. DOI: https://doi.org/10.1145/1978942.1979275 [22] Alan F. Blackwell. 2015. Interacting With an Inferred World: The Challenge of Machine Learning for Humane Computer Interaction. Aarhus Series on Human Centered Computing 1, 1 (Oct. 2015), 12–12. DOI: https://doi.org/10.7146/aahcc.v1i1.21197 [23] Alan F. Blackwell. 2019. Objective Functions: (In)Humanity and Inequity in Artificial Intelligence. HAU: Journal of Ethnographic Theory 9, 1 (March 2019), 137–146. DOI: https://doi.org/10.1086/703871 [24] David M. Blei. 2012. Probabilistic Topic Models. Communications of the ACM 55, 4 (April 2012), 77–84. DOI: https://doi.org/10.1145/2133806.2133826 [25] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research 3, Jan (2003), 993–1022. [26] Gemma Boleda. 2020. Distributional Semantics and Linguistic Theory. Annual Review of Linguistics 6, 1 (2020), 213–234. DOI: https://doi.org/10.1146/annurev-linguistics-011619-030303

[27] Gemma Boleda and Aurélie Herbelot. 2016. Formal Distributional Semantics: Introduction to the Special Issue. Computational Linguistics 42, 4 (Dec. 2016), 619–635. DOI: https://doi.org/10.1162/COLI_a_00261 [28] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings. In Advances in Neural Information Processing Systems (NeurIPS). D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.), Curran Associates, Inc., Barcelona, Spain, 4349–4357. [29] Jacob T. Borodovsky, Lisa A. Marsch, and Alan J. Budney. 2018. Studying Cannabis Use Behaviors With Face- book and Web Surveys: Methods and Insights. JMIR Public Health and Surveillance 4, 2 (May 2018), e9408. DOI: https://doi.org/10.2196/publichealth.9408 [30] Geoffrey C. Bowker and Susan Leigh Star. 1999. Sorting Things Out: Classification and Its Consequences. MIT Press, Cambridge, MA. [31] Rosi Braidotti. 2019. Posthuman Knowledge. Polity Press, Cambridge. [32] Stacy M. Branham, Steve H. Harrison, and Tad Hirsch. 2012. Expanding the Design Space for Intimacy: Supporting Mutual Reflection for Local Partners. In Proceedings of the ACM Conference on Designing Interactive Systems (DIS). ACM, New York, NY, 220–223. DOI: https://doi.org/10.1145/2317956.2317990 [33] Sarah Brayne. 2020. Predict and Surveil: Data, Discretion, and the Future of Policing. Oxford University Press, Oxford, NY. [34] Simone Browne. 2015. Dark Matters: On the Surveillance of Blackness. Duke University Press, Durham, NC. [35] Hrönn Brynjarsdóttir, Maria Håkansson, James Pierce, Eric P. S. Baumer, Carl DiSalvo, and Phoebe Sengers. 2012. Sustainably Unpersuaded: How Persuasion Narrows Our Vision of Sustainability. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). ACM, New York, NY, 947–956. [36] Taina Bucher. 2017. The Algorithmic Imaginary: Exploring the Ordinary Affects of Facebook Algorithms. Information, Communication & Society 20, 1 (2017), 30–44. DOI:

https://doi.org/10.1080/1369118X.2016.1154086 [37] Jenna Burrell. 2016. How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms. Big Data & Society 3, 1 (Jan. 2016), 2053951715622512. DOI: https://doi.org/10.1177/2053951715622512 [38] Jenna Burrell and Marion Fourcade. 2021. The Society of Algorithms. Annual Review of Sociology 47, 1 (2021), 213–237. DOI: https://doi.org/10.1146/annurev-soc-090820-020800 [39] Judith Butler. 2001. Doing Justice to Someone: Sex Reassignment and Allegories of Transsexuality. GLQ: A Journal of Lesbian and Gay Studies 7, 4 (Sept. 2001), 621–636. [40] Robyn Caplan. 2018. Content or Context Moderation? Technical Report. Data & Society. [41] Sarah J. Carrington, Rachel G. Kent, Jarymke Maljaars, Ann Le Couteur, Judith Gould, Lorna Wing, Ilse Noens, Ina Van Berckelaer-Onnes, and Susan R. Leekam. 2014. DSM-5 Autism Spectrum Disorder: In Search of Essential Behaviours for Diagnosis. Research in Autism Spectrum Disorders 8, 6 (June 2014), 701–715. DOI: https://doi.org/10.1016/j.rasd.2014.03.017 [42] John M. Carroll. 2000. Making Use: Scenario-Based Design of Human-Computer Interactions. The MIT Press, Cambridge, MA. [43] Simon Caton and Christian Haas. 2023. Fairness in Machine Learning: A Survey. ACM Computing Surveys 56, 7 (Aug. 2023), 1–38. DOI: https://doi.org/10.1145/3616865 [44] Stevie Chancellor, Eric P. S. Baumer, and Munmun De Choudhury. 2019. Who Is the “human” in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media. In Proceedings of the ACM on Human- Computer Interaction 3, CSCW (Nov. 2019), 147: 1–147:32. DOI: https://doi.org/10.1145/3359249 [45] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. #Thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW). ACM, New York, NY, 1201–1213. DOI: https://doi.org/10.1145/2818048.2819963 [46] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2022. Quarantined! Examining the 29, 4 (March 2022), 29: 1–29:26. DOI: https://doi.org/10.1145/3490499 [47] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW). ACM, New York, NY, 1217–1230. DOI: https://doi.org/10. 1145/2998181.2998213 [48] Angele Christin. 2017. Algorithms in Practice: Comparing Web Journalism and Criminal Justice. Big Data & Society 4, 2 (Dec. 2017), 2053951717718855. DOI: https://doi.org/10.1177/2053951717718855 [49] Josh Constine. 2011. Facebook Adds Keyword Moderation and Profanity Blocklists to Pages. Retrieved from https: //www.adweek.com/digital/keyword-moderation-profanity-blocklist/

[50] Geoff Cooper and John Bowers. 1995. Representing the User: Notes on the Disciplinary Rhetoric of HCI. In The Social and Interactional Dimensions of Human-Computer Interfaces. Peter J. Thomas (Ed.). Cambridge University Press, Cambridge, 48–66. [51] Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. 2023. The Measure and Mismeasure of Fairness. Journal of Machine Learning Research 24, 312 (2023), 1–117. [52] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. arXiv:1808.00023 (July 2018). [53] Nick Couldry and Ulises A. Mejias. 2021. The Costs of Connection – How Data Is Colonizing Human Life and Appropri- ating It for Capitalism. Stanford University Press, Redwood City, CA. [54] Constantinos K. Coursaris and Ming Liu. 2009. An Analysis of Social Support Exchanges in Online HIV/AIDS Self- Help Groups. Computers in Human Behavior 25, 4 (July 2009), 911–918.

DOI: https://doi.org/10.1016/j.chb.2009.03.006 [55] Kate Crawford and Tarleton Gillespie. 2016. What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint. New Media & Society 18, 3 (March 2016), 410–428. DOI: https://doi.org/10.1177/1461444814543163 [56] Mark Davies. 2008. Corpus of Contemporary American English (COCA). Retrieved from https://www.english- corpora.org/coca/. [57] Tisha Dejmanee. 2013. Bodies of Technology: Performative Flesh, Pleasure and Subversion in Cyberspace. Gender Questions 1, 1 (Jan. 2013), 3–17. [58] Gilles Deleuze and Félix Guattari. 1987. A Thousand Plateaus: Capitalism and Schizophrenia. University of Minnesota Press, Minneapolis, MN. Translation of: Mille plateaux, v. 2 of Capitalisme et schizophrénie A companion volume to Anti-Oedipus : Capitalism and schizophrenia Includes index. [59] Vinciane Despret. 2008. The Becomings of Subjectivity in Animal Worlds. Subjectivity 23, 1 (July 2008), 123–139. DOI: https://doi.org/10.1057/sub.2008.15 [60] Vinciane Despret. 2013. From Secret Agents to Interagency. History and Theory 52, 4 (2013), 29–44. DOI: https: //doi.org/10.1111/hith.10686 [61] Michael Ann DeVito, Jeremy Birnholtz, Jeffery T. Hancock, Megan French, and Sunny Liu. 2018. How People Form Folk Theories of Social Media Feeds and What It Means for How We Study Self-Presentation. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). ACM, New York, NY, 120: 1–120:12. DOI: https://doi.org/10.1145/3173574.3173694 [62] Michael Ann DeVito, Darren Gergle, and Jeremy Birnholtz. 2017. ”Algorithms Ruin Everything”: #RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). ACM, New York, NY, 3163–3174. DOI: https://doi.org/10.1145/3025453.3025659 [63] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-Training of Deep Bidi- rectional Transformers for Language Understanding. In Proceedings of the Annual Meeting of the North Ameri- can Chapter of the Association for Computational Linguistics (NAACL). ACL, Stroudsburg, PA, 4171–4186. DOI: https://doi.org/10.18653/v1/N19-1423 [64] Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. Sexuality & Culture 25, 2 (April 2021), 700–732. DOI: https://doi.org/10.1007/s12119-020-09790-w [65] Catherine D’Ignazio and Lauren F. Klein. 2020. Data Feminism. MIT Press, Cambridge, MA. [66] Karthik Dinakar, Jackie Chen, Henry Lieberman, Rosalind Picard, and Robert Filbin. 2015. Mixed-Initiative Real-Time Topic Modeling & Visualization for Crisis Counseling. In Proceedings of the ACM Conference on Intelligent User Interfaces (IUI). ACM, New York, NY, 417–426. DOI: https://doi.org/10.1145/2678025.2701395 [67] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. 2018. Explainable Artificial Intelligence: A Survey. In 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). 0210–0215. DOI: https://doi.org/10.23919/MIPRO.2018.8400040 [68] Paul Dourish. 2006. Implications for Design. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). ACM, New York, NY, 541–550. DOI: https://doi.org/10.1145/1124772.1124855 [69] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). ACM, New York, NY, 278–288. DOI: https://doi.org/10.1145/3025453.3025739 [70] Dmitry Epstein, Merrill C. Roth, and Eric P. S. Baumer. 2014. It’s the Definition, Stupid! Framing of Online Privacy in the Internet Governance Forum Debates. Journal of Information Policy 4 (2014), 144. DOI:

https://doi.org/10.5325/jin- fopoli.4.2014.0144 [71] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I “like” It, Then I Hide It: Folk Theories of Social Feeds. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). ACM, New York, NY, 2371–2382. DOI: https://doi.org/10.1145/2858036.2858494

[72] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. “I Always Assumed That I Wasn’t Really That Close to [Her]”: Reasoning about Invisible Algorithms in News Feeds. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). ACM, New York, NY, 153–162. DOI: https://doi.org/10.1145/2702123.2702556 [73] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User Attitudes Towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. In Pro- ceedings of the ACM Conference on Human Factors in Computing Systems (CHI). ACM, New York, NY, 1–14. DOI: https://doi.org/10.1145/3290605.3300724 [74] Eran Fisher. 2022. Algorithms and Subjectivity: The Subversion of Critical Knowledge. Routledge, London. DOI: https://doi.org/10.4324/9781003196563 [75] Jodi Forlizzi. 2018. Moving Beyond User-centered Design. Interactions 25, 5 (Aug. 2018), 22–23. DOI: https://doi.org/ 10.1145/3239558 [76] Michel Foucault. 1965. Madness and Civilization (1988 ed.). Vintage Books, New York, NY. [77] Michel Foucault. 1977. Discipline and Punish: The Birth of the Prison. Vintage Books, New York, NY. [78] Michel Foucault. 1988. Technologies of the Self. Lectures at the University of Vermont, October 1982. In Technologies of the Self . Luther H. Martin, Huck Gutman, and Patrick H. Hutton (Eds.), University of Massachusetts Press, 16–49. [79] Christopher Frauenberger. 2019. Entanglement HCI the Next Wave? ACM Transactions on Computer-Human Interac- tion 27, 1 (Nov. 2019), 2: 1–2:27. DOI: https://doi.org/10.1145/3364998 [80] Tarleton Gillespie. 2013. The Relevance of Algorithms. In Media Technologies. Tarleton Gillespie, Pablo Bockzkowski, and Kirsten Foot (Eds.). MIT Press, Cambridge, MA, 167–193. [81] Tarleton Gillespie. 2020. Content Moderation, AI, and the Question of Scale. Big Data & Society 7, 2 (July 2020), 2053951720943234. DOI: https://doi.org/10.1177/2053951720943234 [82] Temple Grandin and Richard Panek. 2013. The Autistic Brain: Thinking Across the Spetrum. Houghton Mifflin Harcourt. [83] Mary L. Gray and Siddharth Suri. 2019. Ghost Work. Houghton Mifflin Harcourt, Boston. [84] Wei Guo and Aylin Caliskan. 2021. Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-Like Biases. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES). ACM, New York, NY, 122–133. DOI: https://doi.org/10.1145/3461702.3462536 [85] Weiwei Guo and Mona Diab. 2011. Semantic Topic Models: Combining Word Distributional Statistics and Dictionary Definitions. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). ACL, Stroudsburg, PA, 552–561. [86] Rahul Kumar Gupta, Ritu Agarwalla, Bukya Hemanth Naik, Joythish Reddy Evuri, Apil Thapa, and Thoudam Doren Singh. 2022. Prediction of Research Trends Using LDA Based Topic Modeling. Global Transitions Proceedings 3, 1 (June 2022), 298–304. DOI: https://doi.org/10.1016/j.gltp.2022.03.015 [87] Kevin D. Haggerty and Richard V. Ericson. 2000. The Surveillant Assemblage. The British Journal of Sociology 51, 4 (2000), 605–622. DOI: https://doi.org/10.1080/00071310020015280 [88] Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social

Media Users: Marginalization and Moderation Gray Areas. In Proceedings of the ACM on Human-Computer Interaction 5, CSCW2 (Oct. 2021), 466: 1–466:35. DOI: https://doi.org/10.1145/3479610 [89] Blake Hallinan and Jed R. Brubaker. 2021. Living With Everyday Evaluations on Social Media Platforms. International Journal of Communication 15 (2021), 19. [90] Donna Haraway. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspec- tive. Feminist Studies 14, 3 (1988), 575–599. DOI: https://doi.org/10.2307/3178066 [jstor]3178066 [91] Donna J. Haraway. 1991a. A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century. In Simians, Cybors, and Women: The Reinvention of Nature. Routledge, New York, NY, 149–182. [92] Donna J. Haraway. 1991b. Simians, Cyborgs, and, Women: The Reinvention of Nature. Routledge, New York, NY. [93] Donna J. Haraway. 2013. SF: Science Fiction, Speculative Fabulation, String Figures, So Far. Ada: A Journal of Gender, New Media, and Technology 3 (2013). [94] Graham Harman. 2018. Object-Oriented Ontology: A New Theory of Everything. Pelican Books, London. [95] Camille Harris, Matan Halevy, Ayanna Howard, Amy Bruckman, and Diyi Yang. 2022. Exploring the Role of Grammar and Word Choice in Bias Toward African American English (AAE) in Hate Speech Classification. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT). ACM, Seoul, Republic of Korea, 789–798. DOI: https://doi.org/10.1145/3531146.3533144 [96] Denise L. Haynie, Tonja Nansel, Patricia Eitel, Aria Davis Crump, Keith Saylor, Kai Yu, and Bruce Simons-Morton. 2001. Bullies, Victims, and Bully/Victims:: Distinct Groups of At-Risk Youth. The Journal of Early Adolescence 21, 1 (Feb. 2001), 29–49. DOI: https://doi.org/10.1177/0272431601021001002

[97] Joanne Hinds and Adam N. Joinson. 2018. What Demographic Attributes Do Our Digital Footprints Reveal? A Systematic Review. PLOS ONE 13, 11 (Nov. 2018), e0207112. DOI: https://doi.org/10.1371/journal.pone.0207112 [98] Anna Lauren Hoffmann. 2019. Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse. Information, Communication & Society 22, 7 (June 2019), 900–915. DOI: https://doi.org/10.1080/1369118X.2019.1573912 [99] Gregory Hollin, Isla Forsyth, Eva Giraud, and Tracey Potts. 2017. (Dis)Entangling Barad: Materialisms and Ethics. Social Studies of Science 47, 6 (2017), 918–941. [100] Carla Hustak and Natasha Myers. 2012. Involutionary Momentum: Affective Ecologies and the Sciences of Plant/Insect Encounters. Differences 23, 3 (Dec. 2012), 74–118. DOI: https://doi.org/10.1215/10407391-1892907 [101] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit. In Proceedings of the ACM on Human-Computer Interaction 3, CSCW (Nov. 2019), 150: 1–150:27. DOI: https://doi.org/10.1145/3359252 [102] Jialun Aaron Jiang, Charles Kiene, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. 2019. Moderation Challenges in Voice-Based Online Communities on Discord. In Proceedings of the ACM on Human-Computer Interaction 3, CSCW (Nov. 2019), 55: 1–55:23. DOI: https://doi.org/10.1145/3359157 [103] Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R. Brubaker. 2021. Understanding International Perceptions of the Severity of Harmful Content Online. PLOS ONE 16, 8 (Aug. 2021), e0256762. DOI: https://doi.org/10.1371/journal.pone.0256762 [104] Matthew L. Jockers. 2013. Macroanalysis: Digital Methods & Literary History. University of Illinois Press, Chicago. [105] Christ Plauché Johnson and Scott M. Myers. 2008. CHAPTER 15 - Autism spectrum disorders. In Developmental- Behavioral Pediatrics. Mark L. Wolraich, Dennis D. Drotar, Paul H. Dworkin, and Ellen C. Perrin (Eds.), Mosby, PA, 519–577. DOI: https://doi.org/10.1016/B978-0-

323-04025-9.50018-0 [106] Os Keyes. 2020. Automating Autism: Disability, Discourse, and Artificial Intelligence. The Journal of Sociotechnical Critique 1, 1 (Dec. 2020), 8. DOI: https://doi.org/10.25779/89bj-j396 [107] Charles Kiene, Andres Monroy-Hernández, and Benjamin Mako Hill. 2016. Surviving an ”Eternal September”: How an online community managed a surge of newcomers. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). ACM, New York, NY, 1152–1156. DOI: https://doi.org/10.1145/2858036.2858356 [108] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. 2014. Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks. Proceedings of the National Academy of Sciences (PNAS) 111, 24 (2014), 8788–8790. DOI: https://doi.org/10.1073/pnas.1320040111 [109] Robert E. Kraut and Paul Resnick. 2011. Building Successful Online Communities: Evidence-Based Social Design. Massachusetts Institute of Technology, Cambridge, MA. [110] Ray Kurzweil. 1990. The Age of Intelligent Machines. MIT Press, Cambridge, MA. [111] u/landoflobsters. 2018. Revamping the Quarantine Function. Retrieved March 1, 2021 from https://www.reddit.com/ r/announcements/comments/9jf8nh/revamping_the_quarantine_function [112] Bruno Latour. 1994. On Technical Mediation: Philosophy, Sociology, Genealogy. Common Knowledge 3, 2 (1994), 29–64. [113] Bruno Latour. 1999. On Recalling Ant. The Sociological Review 47, 1_suppl (May 1999), 15–25. DOI: https://doi.org/10.1111/j.1467-954X.1999.tb03480.x [114] Min Kyung Lee, Alexandros Psomas, Ariel D. Procaccia, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, and Siheon Lee. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. In Proceedings of the ACM on Human-Computer Interaction 3, CSCW (Nov. 2019), 1–35. DOI: https://doi.org/10.1145/3359283 [115] Lawrence Lessig. 2000. Code is Law. Retrieved from https://harvardmagazine.com/2000/01/code-is-law-html. [116] David Lewis. 1978. Truth in Fiction. American Philosophical Quarterly 15, 1 (1978), 37–46. [117] Renkai Ma and Yubo Kou. 2022. “I’m Not Sure What Difference Is Between Their Content and Mine, Other Than the Person Itself”: A Study of Fairness Perception of Content Moderation on YouTube. In Proceedings of the ACM on Human-Computer Interaction (Oct. 2022). DOI: https://doi.org/10.1145/3555150 [118] Julie Maitland and Matthew Chalmers. 2011. Designing for Peer Involvement in Weight Management. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI). ACM, New York, NY, 315–324. DOI: https://doi.org/10.1145/1978942.1978988 [119] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to Information Retrieval. Cambridge University Press, Cambridge. [120] Lucas Matney. 2019. Reddit Quarantines Its Biggest Headache. Retrieved from https://social.techcrunch.com/2019/ 06/26/reddit-quarantines-its-biggest-headache/ [121] Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. Retrieved from http: //mallet.cs.umass.edu/

[122] Jie Mei, Christian Desrosiers, and Johannes Frasnelli. 2021. Machine Learning for the Diagnosis of Parkinson’s Disease: A Review of Literature. Frontiers in Aging Neuroscience 13 (2021), 633752. DOI: https://doi.org/10.3389/fnagi.2021.633752 [123] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In Advances in Neural Information Processing Systems (NeurIPS), C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates, Inc., Red Hook, NY, 3111–3119. [124] Marvin Minsky (Ed.). 1968. Semantic Information Processing. MIT Press, Cambridge, MA. [125] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016.

The Ethics of Algorithms: Mapping the Debate. Big Data & Society 3, 2 (Dec. 2016), 1–21. DOI: https://doi.org/10.1177/2053951716 679679 [126] Torin Monahan. 2017. Regulating Belonging: Surveillance, Inequality, and the Cultural Production of Abjection. Journal of Cultural Economy 10, 2 (March 2017), 191–206. DOI: https://doi.org/10.1080/17530350.2016.1273843 [127] Michael Muller, Shion Guha, Eric P. S. Baumer, David Mimno, and N. Sadat Shami. 2016. Machine Learning and Grounded Theory Method: Convergence, Divergence, and Combination. In Proceedings of the ACM Conference on Supporting Group Work (GROUP). ACM, New York, NY. 3–6. [128] Margi Murphy. 2017. Artificial Intelligence Will Detect Child Abuse Images to Save Police from Trauma. The Telegraph (Dec. 2017). [129] Sarah Myers West. 2018. Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms. New Media & Society 20, 11 (Nov. 2018), 4366–4383. DOI: https://doi.org/10.1177/1461444818773059 [130] Roberto Navigli. 2009. Word Sense Disambiguation: A Survey. ACM Computing Surveys 41, 2 (Feb. 2009), 10: 1–10:69. DOI: https://doi.org/10.1145/1459352.1459355 [131] Laura K. Nelson. 2017. Computational Grounded Theory: A Methodological Framework. Sociological Methods & Research (Nov. 2017), 3–42. DOI: https://doi.org/10.1177/0049124117729703 [132] Mark W. Newman, Debra Lauterbach, Sean A. Munson, Paul Resnick, and Margaret E. Morris. 2011. It's Not That I Don't Have Problems, I'm Just Not Putting Them on Facebook: Challenges and Opportunities in Using Online Social Networks for Health. In Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW). ACM, New York, NY, 341–350. DOI: https://doi.org/10.1145/1958824.1958876 [133] Casey Newton. 2019. The Secret Lives of Facebook Moderators in America. Retrieved from https://www.theverge. com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona [134] Lene Nielsen. 2004. Engaging Personas and Narrative Scenarios. Ph.D. Dissertation. Department of Informatics, Copenhagen Business School, Copenhagen, Denmark. [135] Safiya Umoja Noble. 2018. Algorithms of Oppression: How Search Engines Reinforce Racism. NYU Press, New York, NY. [136] Walter J. Ong. 1982. Orality and Literay: The Technologizing of the Word (2002 ed.). Routledge, New York, NY. [137] Cathy O'Niel. 2016. Weapons of Math Destruction. Crown, New York, NY. [138] Yotam Ophir, Dror Walter, and Eleanor R. Marchant. 2020. A Collaborative Way of Knowing: Bridging Computational Communication Research and Grounded Theory Ethnography. Journal of Communication 70, 3 (June 2020), 447–472. DOI: https://doi.org/10.1093/joc/jqaa013 [139] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training Language Models to Follow Instructions with Human Feedback. In Advances in Neural Information Processing Systems (NeurIPS). New Orleans, LA. [140] Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in Word Embeddings. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*). ACM, New York, NY, 446–457. DOI: https://doi.org/10.1145/3351095.3372843 [141] Samir Passi and Steven Jackson. 2017. Data Vision: Learning to See Through Algorithmic Abstraction. In Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW). ACM, New York, NY, 2436–2447. DOI: https://doi.org/10.1145/2998181.2998331 [142] Dana Pessach and Erez Shmueli. 2022. A Review on Fairness in Machine Learning. Comput. Surveys 55, 3 (Feb. 2022), 51:1–51:44. DOI: https://doi.org/10.1145/3494672 [143] Davor Petreski and Ibrahim

C. Hashim. 2022. Word Embeddings Are Biased. But Whose Bias Are They Reflecting? AI & Society 38, 2 (May 2022), 1–8. DOI: https://doi.org/10.1007/s00146-022-01443-w [144] Steven Piantadosi, David P. Byar, and Sylvan B. Green. 1988. The Ecological Fallacy. American Journal of Epidemiology 127, 5 (May 1988), 893–904. DOI: https://doi.org/10.1093/oxfordjournals.aje.a114892 [145] Angelisa C. Plane, Elissa M. Redmiles, Michelle L. Mazurek, and Michael Carl Tschantz. 2017. Exploring User Perceptions of Discrimination in Online Targeted Advertising. In Proceedings of the USENIX Security Symposium. USENIX Association, 935–951.

[146] John Pruitt and Jonathan Grudin. 2003. Personas: Practice and Theory. In Proceedings of the ACM Conference on Designing for User Experiences (DUX). ACM, San Francisco, CA, 1–15. DOI: https://doi.org/10.1145/997078.997089 [147] Emilee Rader and Rebecca Gray. 2015. Understanding User Beliefs about Algorithmic Curation in the Facebook News Feed. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). ACM, Seoul, 173–182. DOI: https://doi.org/10.1145/2702123.2702174 [148] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. Technical Report. OpenAI, San Francisco, CA. [149] Anand Rajaraman and Jeffrey David Ullman. 2011. Data Mining. In Mining of Massive Datasets. Cambridge University Press, Cambridge, 1–17. DOI: https://doi.org/10.1017/CBO9781139058452.002 [150] Reddit. [n. d.]. Quarantined Subreddits. Retrieved from https://reddit.zendesk.com/hc/en-us/articles/360043069012- Quarantined-Subreddits. [151] Michael J. Reddy. 1979. The Conduit Metaphor: A case of Frame Conflict in Our Language about Language. In Metaphor and Thought. Andrew Ortony (Ed.), Cambridge University Press, Cambridge, 164–201. [152] Johan Redström. 2006. Towards User Design? On the Shift from Object to User as the Subject of Design. Design Studies 27, 2 (March 2006), 123–139. DOI: https://doi.org/10.1016/j.destud.2005.06.001 [153] Jason Ren, Russell Kunes, and Finale Doshi-Velez. 2020. Prediction Focused Topic Models via Feature Selection. In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS). PMLR, 4420–4429. [154] Philip Resnik. 1997. Selectional Preference and Sense Disambiguation. In Tagging Text with Lexical Semantics: Why, What, and How? ACL. Retrieved from https://aclanthology.org/W97-0209.pdf [155] Lisa Rhody. 2014. The Story of Stopwords: Topic Modeling an Ekphrastic Tradition. In Proceedings of the Digital Humanities Conference. ADHO, Lausanne. [156] Lisa Marie Rhody. 2013. Topic Modeling and Figurative Language. Journal of Digital Humanities 2, 1 (April 2013), 19–35. [157] Lisa Marie Rhody. 2017. Beyond Darwinian Distance: Situating Distant Reading in a Feminist Ut Pictura Poesis Tradition. PMLA 132, 3 (May 2017), 659–667. DOI: https://doi.org/10.1632/pmla.2017.132.3.659 [158] Elaine Rich and Kevin Knight. 1991. Artificial Intelligence. McGraw Hill, New York, NY. [159] George Ritzer. 1993. The McDonaldization of American Society: An Investigation into the Changing Character of Contemporary Social Life. Pine Forge Press, Newbury Park, CA. [160] Sarah T. Roberts. 2019. Behind the Screen: Content Moderation in the Shadows of Social Media. Yale University Press, New Haven, CT. DOI: https://doi.org/10.2307/j.ctvhrcz0v [161] Nikolas Rose. 2010. Inventing Our Selves: Psychology, Power, and Personhood. Cambridge University Press, Cambridge. [162] Marie-Laure Ryan. 2013. Possible worlds. In The Living Handbook of Narratology. Peter Hühn, John Pier, Wolf Schmid, and Jörg Schönert (Eds.). Hamburg University, Hamburg, Germany. Retrieved from http://lhn.sub.uni-hamburg.de/index.php/Possible_Worlds.html [163] Nithya Sambasivan, Ed Cutrell, Kentaro

Toyama, and Bonnie Nardi. 2010. Intermediated Technology Use in Develop- ing Communities. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). Atlanta, GA, 2583–2592. DOI: https://doi.org/10.1145/1753326.1753718 [164] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. ITU Journal: ICT Discoveries 1, Special Issue 1 (Oct. 2017), 39–48. [165] Ari Schlesinger, W. Keith Edwards, and Rebecca E. Grinter. 2017. Intersectional HCI: Engaging Identity Through Gender, Race, and Class. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). ACM, Denver, CO, 5412–5427. DOI: https://doi.org/10.1145/3025453.3025766 [166] Ari Schlesinger, Kenton P. O'Hara, and Alex S. Taylor. 2018. Let's Talk about Race: Identity, Chatbots, and AI. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). ACM, Montréal, QC, 315:1–315:14. DOI: https://doi.org/10.1145/3173574.3173889 [167] Alexandra Schofield and David Mimno. 2016. Comparing Apples to Apple: The Effects of Stemmers on Topic Models. Transactions of the Association for Computational Linguistics 4 (July 2016), 287–300. [168] Nick Seaver. 2022. Computing Taste: Algorithms and the Makers of Music Recommendation. University of Chicago Press, Chicago, IL. [169] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator Engagement and Community Develop- ment in the Age of Algorithms. New Media & Society 21, 7 (July 2019), 1417–1443. DOI: https://doi.org/10.1177/1461 444818821316 [170] Patrick Skeba and Eric P. S. Baumer. 2020. Informational Friction as a Lens for Studying Algorithmic Aspects of Privacy. In Proceedings of the ACM on Human-Computer Interaction 4, CSCW2 (Oct. 2020), 101:1–101:22. DOI: https://doi.org/10.1145/3415172

[171] Susan Leigh Star and Geoffrey C. Bowker. 2002. How to Infrastructure. In Handbook of New Media: Social Shaping and Consequences of ICTs, L. Lievrouw and S. Livingstone (Eds.). SAGE, 151–162. DOI: https://doi.org/10.4135/978184 8608245 [172] Luke Stark. 2018. Algorithmic Psychometrics and the Scalable Subject. Social Studies of Science 48, 2 (April 2018), 204–231. DOI: https://doi.org/10.1177/0306312718772094 [173] Luke Stark. 2019. Facial Recognition Is the Plutonium of AI. XRDS: Crossroads, The ACM Magazine for Students 25, 3 (April 2019), 50–55. DOI: https://doi.org/10.1145/3313129 [174] Isabelle Stengers. 2013. Introductory Notes on an Ecology of Practices. Cultural Studies Review 11, 1 (Aug. 2013), 183–196. DOI: https://doi.org/10.5130/csr.v11i1.3459 [175] Isabelle Stengers. 2017. Thinking Life: The Problem Has Changed. In Posthumous Life: Theorizing Beyond the Posthuman. Columbia University Press, New York, NY, 325–338. DOI: https://doi.org/10.7312/wein17214-015 [176] Alex S. Taylor. 2009. Machine Intelligence. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). 2109–2118. [177] Alex S. Taylor. 2015. After Interaction. Interactions 22, 5 (Aug. 2015), 48–53. DOI: https://doi.org/10.1145/2809888 [178] J. K. Trotter. 2016. Facebook Admits Pulitzer-Winning Photograph Is Not Child Pornography. Retrieved from https://gizmodo.com/facebook-admits-pulitzer-winning-photograph-is-not-chil-1786441732. [179] Ted Underwood. 2012. What Kinds of "Topics" Does Topic Modeling Actually Produce? The Stone and the Shell. Retrieved from https://tedunderwood.com/2012/04/01/what-kinds-of-topics-does-topic-modeling-actually-produce/ [180] Ted Underwood. 2014. Theorizing Research Practices We Forgot to Theorize Twenty Years Ago. Representations 127, 1 (2014), 64–72. [181] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "At the End of the Day Facebook Does What It Wants": How Users Experience Contesting Algorithmic Content Moderation. In Proceedings of the

ACM on Human-Computer Interaction 4, CSCW2 (Oct. 2020), 167:1–167:22. DOI: https://doi.org/10.1145/3415238 [182] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability for Content Mod- eration. In Proceedings of the ACM on Human-Computer Interaction 5, CSCW2 (Oct. 2021), 318:1–318:28. DOI: https://doi.org/10.1145/3476059 [183] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In Advances in Neural Information Processing Systems (NeurIPS), Vol. 30. Curran Associates, Inc. [184] Ron Wakkary. 2020. A Posthuman Theory for Knowing Design. International Journal of Design 14, 3 (2020), 12. [185] Ron Wakkary. 2021. Things We Could Design: For More Than Human-Centered Worlds. The MIT Press, Cambridge, MA. [186] Ron Wakkary, William Odom, Sabrina Hauser, Garnet Hertz, and Henry Lin. 2015. Material Speculation: Actual Artifacts for Critical Inquiry. In Decennial Aarhus Conference on Human Centered Computing. DOI: https://doi.org/10.7146/aahcc.v1i1.21299 [187] Joseph B. Walther and Shawn Boyd. 2022. Attraction to Computer-Mediated Social Support. In Communication Technology and Society: Audience Adoptino and Uses. C. A. Lin and D. Atkins (Eds.), Hampton Press, Cresskill, NJ, 153–188. [188] Chi Wang, Rajat Raina, David Fong, Ding Zhou, Jiawei Han, and Greg Badros. 2011. Learning Relevance from Heterogeneous Social Network and Its Application in Online Targeting. In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). ACM, Beijing, China, 655–664. DOI: https://doi.org/10.1145/2009916.2010004 [189] Max Weber. 1920 (2003). The Protestant Ethic and the Spirit of Capitalism. Dover, Mineola, NY. [190] Martin Wiener, W. Alec Cram, and Alexander Benlian. 2023. Algorithmic Control and Gig Workers: A Legit- imacy Perspective of Uber Drivers. European Journal of Information Systems 32, 3 (May 2023), 485–507. DOI: https://doi.org/10.1080/0960085X.2021.1977729 [191] Lorna Wing, Judith Gould, and Christopher Gillberg. 2011. Autism Spectrum Disorders in the DSM-V: Bet- ter or Worse Than the DSM-IV? Research in Developmental Disabilities 32, 2 (March 2011), 768–773. DOI: https://doi.org/10.1016/j.ridd.2010.11.003 [192] Andrew Winzelberg. 1997. The Analysis of an Electronic Support Group for Individuals with Eating Disorders. Computers in Human Behavior 13, 3 (Aug. 1997), 393–407. DOI: https://doi.org/10.1016/S0747-5632(97)00016-2 [193] Ludwig Wittgenstein. 1922. Tractatus Logico-Philosophicus. Routledge & Kegan Paul, London. [194] Ludwig Wittgenstein. 1953. Philosophical Investigations. Blackwell, Oxford. [195] Marisol Wong-Villacres, Arkadeep Kumar, Aditya Vishwanath, Naveena Karusala, Betsy DiSalvo, and Neha Kumar. 2018. Designing for Intersections. In Proceedings of the ACM Conference on Designing Interactive Systems (DIS). ACM, Hong Kong, 45–58. DOI: https://doi.org/10.1145/3196709.3196794

[196] Alex J. Wood, Mark Graham, Vili Lehdonvirta, and Isis Hjorth. 2019. Good Gig, Bad Gig: Autonomy and Algorithmic Control in the Global Gig Economy. Work, Employment and Society 33, 1 (Feb. 2019), 56–75. DOI: https://doi.org/ 10.1177/0950017018785616 [197] Qian Yang, Justin Cranshaw, Saleema Amershi, Shamsi T. Iqbal, and Jaime Teevan. 2019. Sketching NLP: A Case Study of Exploring the Right Things to Design with Language Intelligence. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). ACM, Glasgow, 185:1–185:12. DOI: https://doi.org/10.1145/3290605.3300415 [198] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating How Experienced UX Designers Effectively Work with Machine Learning. In Proceedings of the ACM Conference on Designing Interactive Systems (DIS). ACM, Hong Kong, 585–596. DOI:

https://doi.org/10.1145/3196709.3196730 [199] Koji Yatani, Michael Novati, Andrew Trusty, and Khai N. Truong. 2011. Review Spotlight: A User Interface for Summarizing User-Generated Reviews Using Adjective-Noun Word Pairs. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). ACM, Vancouver, BC, 1541–1550.

Received 7 June 2023; revised 6 March 2024; accepted 7 March 2024