Reporte del encuentro:

Fortalecer la verdad en la era de los medios sintéticos

Bogotá, Colombia
2-3 de agosto de 2023

Reporte redactado por:
Juliana Guerra Rudas
(juliana@usuarix.net)

Entre el 2 y 3 de agosto de 2023, WITNESS convocó en Bogotá, Colombia, a personas activistas, artistas, desarrolladoras, creadoras y verificadoras de contenidos digitales, periodistas y defensoras de derechos humanos de diferentes países de América Latina, para identificar, discutir y priorizar las amenazas, soluciones y oportunidades de los medios sintéticos.

Este taller hace parte de una serie de consultas realizadas desde 2019 por WITNESS con las comunidades con quienes trabaja alrededor del mundo, para incidir tempranamente en el ecosistema de los medios sintéticos, reconociendo los diferentes contextos, necesidades y perspectivas situadas de estas comunidades.

En un contexto de cambios vertiginosos respecto de las capacidades técnicas, así como de la accesibilidad a herramientas para la generación y manipulación de contenidos con inteligencia artificial, este es el primer taller que se realiza en la región hispanohablante de América Latina, que se suma a los realizados antes en Sudáfrica, Brasil, Malasia (2019), Estados Unidos (2020) y Kenia (2023).

Con su trabajo, WITNESS busca garantizar que el desarrollo continuo de los medios sintéticos refleje la comprensión de las principales amenazas, soluciones y oportunidades priorizadas por las personas y comunidades de todo el mundo, que han sufrido violaciones de los derechos humanos similares a las que están surgiendo en la "era de los medios sintéticos". Estas personas son las que corren mayor riesgo y, a la vez, están marginadas de los espacios que dan forma a esta tecnología emergente.

Contenidos

Relatoría del día 1

Sesión 1. Bienvenida e introducción al trabajo de WITNESS

Indira Cornelio y Laura Salas del equipo de América Latina de WITNESS dieron la bienvenida al encuentro, presentaron los objetivos, la propuesta de agenda y de acuerdos generales de participación, que incluyeron seguir la regla de Chatham House, esto es que se puede usar la información escuchada durante el encuentro, pero no se puede revelar la identidad ni la afiliación de la persona oradora, ni la de ninguna otra persona participante, a menos que se haya acordado previamente. Luego se hizo una ronda de presentación.

Después presentaron el modelo holístico de trabajo de WITNESS alrededor del mundo, que consiste en empezar por la escucha y la anticipación, colaborar con quienes usan las tecnologías para amplificar su impacto, y crear herramientas y guías para expandir los conocimientos de las comunidades, con la intención de que esto ayude a un cambio sistémico y con repercusiones amplias.

También describieron los temas transversales que se trabajan en la región de América Latina: violencia estatal, junto a comunidades a Argentina, Perú, Ecuador, Chile y México; y defensa del territorio, apoyando la visión que viene desde los pueblos originarios en la Amazonía, México y Brasil.

Sesión 2. Introducción a los medios sintéticos y al marco “Fortalecer la verdad”

El director ejecutivo de WITNESS, Sam Gregory, presentó el trabajo pionero de WITNESS para anticiparse a los impactos de las tecnologías de manipulación de medios audiovisuales en el ejercicio de derechos humanos, y específicamente en el fenómeno de los *deepfakes* del que hemos tenido noticia desde hace poco más de 5 años, que implica una crisis de confianza e impone nuevos retos para las denuncias y demandas de los grupos activistas, y de defensores y defensoras de derechos humanos.

Explicó que los medios sintéticos incluyen un amplio rango de técnicas para alterar, crear y manipular videos, imágenes o audios, creando un sinnúmero de posibilidades emergentes: crear una voz, un rostro o un avatar realista, o una escena que nunca sucedió, o manipular una imagen para que parezca algo que nunca pasó.

Comentó que para los medios sintéticos se utilizan tres tipos de modelos de IA: Redes Generativas Adversariales (GAN, por sus siglas en inglés), modelos de difusión y Modelos de Lenguaje Extenso (LLM, por sus siglas en inglés). Y aclaró que el objetivo del taller no es entender cómo funcionan técnicamente estos sistemas sino apuntar al vacío de comprensión sobre lo que ocurre en el término medio. Esto es, qué posibilidades existen para la producción de medios sintéticos, cuáles ya están siendo utilizadas y de qué dependen.

Mencionó que el uso de herramientas para manipular imágenes con tono realista, tales como Photoshop o Google Pixel, no son nuevas, pero han tenido enormes avances desde 2017, y se han comercializado aceleradamente en los últimos meses. Estas y otras herramientas ya están disponibles en Bing, Google y Adobe. Tienen facilidad de uso y accesibilidad, y están siendo utilizadas a favor y en contra de los derechos humanos.

En resumen, son herramientas cada vez más sencillas de usar y que requieren menos

datos; son multimodales, y la calidad de audio y video va mejorando con el tiempo; se encuentran disponibles como servicio en teléfonos móviles; el abuso de medios sintéticos amplifica amenazas comunes que no son nuevas; los esfuerzos de detección no se adecúan a las características de este problema; nuestros cerebros son cada vez menos capaces de distinguir audios y videos que, con el avance de estas tecnologías, son más y más realistas.

Señaló que si bien los *deepfakes* y otros usos de IA generativa no son tan comunes en nuestro trabajo de hoy, su uso se va a seguir expandiendo y con esto los riesgos de abuso van a ir en aumento. Por lo que desde WITNESS vienen haciendo seguimiento continuo del problema, y haciendo consultas con personas activistas en diferentes partes del mundo para priorizar riesgos (Malasia, Sudáfrica, Brasil, EEUU y Kenia, y se espera continuar, ya que tanto las amenazas como las necesidades seguirán cambiando).

Por lo pronto, *shallowfakes* (ediciones o manipulaciones simples) y contenidos descontextualizados, siguen siendo el principal problema de la desinformación, mientras que a nivel global se identifican cuatro amenazas principales: Primero, violencia de género, ya que la mayoría de los medios sintéticos se utilizan para producir imágenes sexuales no consensuales; segundo, el uso de imágenes falsas, fabricadas para campañas de desinformación, gracias a la facilidad de acceso a herramientas para eso; tercero, el uso de medios sintéticos para la sátira y la parodia y lo que esto pueda implicar en términos de censura; y cuarto, el beneficio de la duda del que pueden gozar los actores poderosos, con capacidad para señalar de fabricado un contenido de denuncia que es auténtico.

Sesión 3. La IA generativa y los medios sintéticos en América Latina

¿Son reales las imágenes?

La presentación trataba sobre la investigación, realizada en el marco de una *fellowship* en el Berkman Klein Center de la Universidad de Harvard, sobre cómo durante 200 años de producción mecánica de imágenes, éstas han sido utilizadas para decir mentiras. A partir de algunos ejemplos, la autora mostró cómo, en este tiempo, ha cambiado la brecha de conocimiento sobre las imágenes, y también los medios de producción (desde el cuarto oscuro hasta la Inteligencia Artificial) y los medios de distribución de éstas. Lo que se ha mantenido es su función para ilustrar lo que socialmente se considera verdad, es decir, para reforzar lo que las personas saben o quieren creer. Así, concluye la autora, del mismo modo de que no ha habido en 200 años una única verdad, no existen imágenes reales o verdaderas. Siempre son manipuladas.

El proyecto de la autora se puede consultar en https://areimagesreal.com/

¿Inteligencia? ¿Artificial?

El autor presentó la perspectiva de trabajo con Inteligencia Artificial (IA) desde la productora Gusano Films, de la que hace parte. Comentó que con la proliferación de modelos computacionales en las últimas dos décadas, así como con las redes neuronales profundas y específicamente con las redes antagónicas generativas a partir de 2014, se han abierto nuevas posibilidades para la creación que rompen con el paradigma de la relación humano-máquina, y permite hablar de la IA como colega, co-creadora de contenidos.

Más allá de las promesas solucionistas de la IA, del hacer más rápido y más barato o de materializar las ideas que tenemos en la cabeza, propone hacer una pequeña taxonomía para ubicarse en la IA dentro de los medios audiovisuales: para la comprensión de contenido, y para la síntesis o generación de nuevo contenido. También mencionó que la investigación en tecnologías de IA no ha podido establecer una técnica para identificar contextos, por lo que solo es posible trabajar con objetos.

Habló de un proyecto desarrollado por ellos, para explorar entre millones de imágenes y videos, del Paro Nacional de Colombia en 2021. Generaron una serie de herramientas de código abierto, disponibles para que cualquier persona las pueda usar, para extraer los objetos y clasificarlos de manera automática. También mencionó como ejemplo el proyecto *VFRAME* (Visual Forensics and Metadata Extraction) de Adam Harvey.

Una vez entendemos cierto contenido, el ejercicio de generar nuevo contenido implica pasar de un medio a otro, por ejemplo de texto a imagen, un ejercicio semejante a la transducción porque hay una interpretación en el cambio de medio. Pero también es posible la generación de contenido a partir de los mismos medios, por ejemplo de vídeo a vídeo.

Sesión 4. Uso de IA generativa para incidencia de derechos humanos. Taller Texto-a-Imagen

Jacobo Castellanos de WITNESS presentó consideraciones de base que la organización ha definido como clave en el uso de medios sintéticos en contextos de derechos humanos:

1. No socavar la credibilidad de las organizaciones que trabajan en estos espacios;

2. Etiquetar (con metadatos o huellas invisibles) para comunicar que es contenido generado o manipulado con IA;

3. Cuidar del consentimiento, entendiendo que hay al respecto zonas grises, por ejemplo cuando se trata de personas muertas;

4. Entender que las expectativas sobre la veracidad y el alcance de la manipulación del contenido audiovisual dependerá del contexto en el cual se hace esta manipulación.

También presentó algunos ejemplos de casos de uso en que los medios sintéticos pueden ser utilizados para la defensa de los derechos humanos, por ejemplo con herramientas de pixelado, desenfoque, borrado, agregado o alteración de imagen y audio para la protección de identidad personal, recreación de espacios físicos, visualización de testimonios, y expresión artística y satírica.

Luego se habló de algunas posibilidades que ofrecen distintas herramientas como Dall-E-2 de OpenAI, Stable Diffusion de StabilityAI, Midjourney, así como algunas herramientas de Google y Meta para la creación y edición de contenido con Inteligencia Artificial.

También discutió sobre las dificultades que representa el lenguaje (cuando no es en inglés) en los resultados que pueden ofrecer, y se compartió un ejemplo práctico para reproducir una imágen.

Foto original

Fotos generadas con Dall-e 2

Descripción utilizada: *fila de personas de una comunidad de Guatemala, con vestidos tradicionales, esperando y aburridos y con hambre; cancha de fútbol abandonada; toma lateral; plano general; tres mujeres y dos hombres*.

A partir de éste y otros ejemplos, se explicó brevemente cómo funcionan las distintas herramientas, y las indicaciones que deben tenerse en cuenta para generar contenido con multimedia con Inteligencia Artificial, así como las funciones de *inpainting* (pintar dentro de la imagen) y *outpainting* (agregar elementos de contexto a una imagen) para después hacer un ejercicio práctico con el grupo.

Grupo 1

Intentaron hacer en Midjourney una recreación a partir de una imagen del ex presidente paraguayo Horacio Cartes. Tomaron una imagen de referencia donde aparece él con gente de su partido político en una celebración. Como descripción utilizaron: *Horacio Cartes; político paraguayo; con una multitud de personas festejando una elección; vestidos de rojo*. Notaron que la plataforma no reconoció a Cartes, ni en el primer ni en el segundo intento, donde agregaron más parámetros. En un tercer intento cambiaron el personaje de referencia por Donald Trump, a quien sí reconoció la plataforma.

Grupo 2

Utilizaron una fotografía de una fiesta de disfraces, disponible en internet, e intentaron reproducirla haciendo descripciones en español e inglés, en DreamStudio de Stability AI.

Imagen de referencia

Fotos generadas con DreamStudio: Izquierda: descripción en español. Derecha: descripción en inglés.

En los dos casos, los resultados cambiaron los rostros, pieles y otros atributos físicos de las personas presentes en las fotografías, para hacerles más estereotípicamente bellas. El grupo anotó que la herramienta arroja muchos errores, sobre todo en los dedos de las manos.

Grupo 3

Utilizando Dall-E-2, crearon una imagen falsa para que acompañe una historia anclada a una situación real del momento: no había agua en los baños del hotel, pero tampoco en las habitaciones, ni en la ciudad. Comentaron que ésto se debía a la supuesta instalación de una planta de gaseosas que requería unos conductos de agua muy grandes, los cuales se han tomado toda el agua destinada para Bogotá. Con la foto ficticia que crearon, el grupo aludía a que la empresa responsable era Coca-Cola.

Grupo 4

En el grupo cuatro, intentaron crear fotografías de los candidatos electorales en Guatemala, Bernardo Arévalo y Sandra Torres en una plaza pública, utilizando Stable Diffusion. En los primeros intentos, descubrieron que la herramienta no generaba los rostros de lxs candidatos, pero sí lo hacía, aunque no de manera ideal, cuando incluían en la descripción

que la imagen incluya globos de colores.

Hicieron un segundo intento con la descripción: *dos presidentes de Guatemala se abrazan con muchos globos alrededor*. De nuevo arrojó algunas imágenes con el estereotipo de un presidente, pero cuyos rostros no son reconocibles. En este segundo caso, la herramienta no ofreció resultados que reflejen el espacio de una plaza pública, lo cual suscitó preguntas sobre las restricciones de programa para no generar desinformación.

Grupo 5

Utilizando Midjourney, intentaron retratar algo en la mitad entre un mundo en que los humanos no existen y la naturaleza renace, y otro en que los humanos existen y la naturaleza muere. Le indicaron a la herramienta: *Monstruo verde. Espíritu del agua. Abel Rodríguez*. Notaron que la base de datos de la herramienta no incluye el estilo del artista colombiano Abel Rodríguez.

A partir del resultado, reflexionaron sobre las posibilidades de descolonizar la IA. Mencionaron que la idea de texto a imagen desconoce todas las tradiciones que no son textuales, por lo que se requiere pensar en otros formatos que no requieran pasar por el texto, y cómo sería posible integrar otras epistemes.

Grupo 6

Intentaron utilizar la barbie para representar las protestas de Irán, de la quema del hiyab. Notaron que en ninguna de las plataformas que utilizaron era posible generar una imagen de una quema de hiyab. Por otra parte, todo lo relacionado a la barbie contenía caras sonrientes.

Sesión 5. Miradas hacia la IA generativa y los medios sintéticos

Durante la tarde nos dividimos en cuatro grupos para discutir, por áreas temáticas, cómo se relacionan, así como qué riesgos y oportunidades representan los medios sintéticos para nuestro trabajo.

El grupo sobre género se concentró en la necesidad de apropiar las tecnologías de IA, de manera que sirvan a las necesidades de los diferentes contextos, y así contrarrestar los efectos de la violencia digital contra grupos vulnerables. El grupo sobre derechos digitales se concentró en el problema de la brecha digital y la necesidad de fortalecer el derecho al acceso en un sentido amplio. El grupo sobre desinformación centró su discusión en el reto regulatorio. Y el grupo sobre medios comunitarios y cine se centró en las posibilidades que ofrece la IA generativa para el trabajo creativo, así como los riesgos de limitar los espacios comunitarios y precarizar el trabajo.

Género

El grupo empezó señalando cómo la violencia virtual es una extensión de lo real, cuyos efectos sobre las mujeres pasan por poner en cuestión su participación en redes, y en general en la vida pública. Una mujer debe pensar dos veces si se somete a lo que implica asumir un cargo de visibilidad pública (por ejemplo en la política o el periodismo). Además, a veces para evitar la violencia digital una mujer debe elegir salir de las redes o no aparecer

en internet, es decir no subir contenidos como su nombre, fotografías y demás, pues lo digital siempre está conectado a una identidad.

Se mencionó el caso de una mujer de quien hicieron un montaje a partir de imágenes pornográficas, y en algún momento ella se sentía confundida respecto de si habría estado ahí, porque era muy real, aunque ella no lo recordara. También se mencionó que las personas adolescentes hoy utilizan las redes sociales pero son muy cuidadosas de las imágenes que comparten ahí, y casi no comparten nada por los riesgos que representa.

Se habló de cómo la IA reproduce el patriarcado, el clasismo, el racismo y el machismo. Esto se debe a quiénes desarrollan estas tecnologías (sobre todo hombres blancos), pero también a los insumos que usan para alimentar los modelos. No hay datos representativos de la diversidad poblacional sino que todo viene del norte global, y con esa perspectiva.

Por otro lado, reconociendo que muchos de los movimientos de búsqueda de personas desaparecidas en América Latina son liderados por mujeres, se mencionó que los medios sintéticos podrían servir a su lucha, sobre todo considerando que las comisiones de búsqueda no han sido muy exitosas en identificar a una persona desaparecida después de muchos años. Aunque ya hay algunos ejemplos del uso de medios sintéticos con este fin, como es el caso de IAbuelas, no se ha promovido mucho –de manera ética– su uso.

Durante la discusión se planteó como problema principal, ¿cómo aumentar la representación sin aumentar un riesgo de hipervigilancia para las poblaciones históricamente subrepresentadas?

Hay quienes consideran peligroso abogar por el aumento de la representatividad debido a que no hay estructuras y procesos para asegurar el consentimiento efectivo, no sólo en materia de género, sino desde una óptica interseccional. Alguien mencionó, por ejemplo, que el problema no es la existencia de datos, pues están allí y en muchos casos bajo políticas de datos abiertos de los gobiernos, sino cómo mantener actualizadas las bases de datos. Alguien más mencionó que hoy no podemos hablar de consentimiento efectivo pues no tenemos acceso a la información sobre lo que se va a hacer con las imágenes que llegan a internet, y las personas dueñas de esos datos no están incluidas en los debates sobre protección de datos.

Además, se mencionó que los modelos de negocio bajo los cuales se desarrollan y funcionan las tecnologías de IA no están tratando de trabajar con menos datos (minimización de datos), por lo que es difícil imaginar sistemas locales y adaptados a los contextos. Por eso es necesario promover mayor apropiación de las tecnologías por parte de comunidades a nivel local, tanto de los datos como de los modelos con los cuales funcionan los sistemas de IA. Y utilizar la imaginación radical, porque esto es todavía una utopía.

Ante esto, se planteó que los propios dueños de las imágenes tienen que decidir cómo se usan, y se habló de gobernanza de datos en entornos comunitarios como una decisión política. Se compartió un proyecto de IA para la gobernanza del agua y los recursos naturales del pueblo yaqui, en México, que busca desarrollar metodologías y técnicas para recabar datos generados por y para la comunidad, y que les ayuden en la gestión y gobernanza del agua. Se trata de datos cualitativos para una tecnología que requiere como insumo información cuantitativa.

Derechos digitales

En este grupo la conversación se centró en el problema de la brecha digital como un determinante a la hora de pensar en el ejercicio de derechos con relación a la IA y los medios sintéticos. Plantearon que el derecho al acceso, en un sentido amplio de acceso a dispositivos pero también de alfabetización y capacidades en el uso de herramientas digitales, no está garantizado en América Latina.

Debido a las políticas de *Zero Rating* en la que operadores de telefonía no cobran por los datos usados por aplicaciones específicas, para la mayor parte de la población en la región el acceso a internet se limita al acceso a Whatsapp, lo que pone a ésta y otras pocas plataformas en un lugar dominante, pues son quienes tienen mayor capacidad de desplegar tecnologías de IA en sus servicios, y además son el medio principal de distribución de información.

Esto conecta con otro tema prioritario: la protección de datos. ¿Qué posibilidades tienen las personas usuarias de decidir usar o no ciertas herramientas? ¿Qué control tienen sobre los datos que recaban? ¿Cómo se garantiza que esos datos están protegidos? Desde el punto de vista regulatorio, se planteó como necesario exigir a las plataformas ser más transparentes en la gestión de datos.

Sin embargo, según mencionaron, también son problemáticas las narrativas alrededor de la IA y los medios sintéticos, pues están llenas de tecnicismos, imponiendo más barreras para que las personas usuarias puedan tomar decisiones informadas sobre el uso de datos. Esto, de nuevo, conecta con el problema de la brecha digital.

Las personas que participaron de este grupo estuvieron de acuerdo en que es necesario trabajar en el desarrollo de capacidades críticas frente al uso de plataformas digitales y frente al consumo de contenidos a través de estas plataformas. Al respecto, se mencionó que también es necesario reconocer el derecho a la autodeterminación sobre cómo y para qué usar herramientas digitales, así como el derecho a la desconexión.

Por último, específicamente respecto a los contenidos digitales, se habló de la importancia de que existan mecanismos para identificar su procedencia, respetando el derecho al anonimato de las personas usuarias. También se mencionó el reto regulatorio en materia de libertad de expresión. Los medios sintéticos pueden facilitar la desinformación, pero también otras formas de expresión y por eso están protegidos. Intentar regular es un peligro gigante debido a quién define lo que se considera desinformación. En los casos de Perú, Brasil y Colombia, el poder lo tienen los gobiernos, pero las iniciativas de autorregulación por parte de las plataformas como Facebook o Twitter tampoco tienen un panorama mejor.

Desinformación

En el grupo se compartieron diferentes casos de desinformación y operaciones de influencia en América Latina, que utilizan herramientas de IA generativa en diferentes momentos: para la planificación, para la creación de perfiles y para la construcción de mensajes. El uso de IA en estas campañas se ha identificado desde 2020, principalmente en Venezuela pero también en otros países como Honduras, El Salvador y Ecuador.

Desde otra perspectiva, que analiza todo el trabajo detrás de la producción de desinformación a nivel local, se planteó la inquietud de cuánto tiempo llevará a los pequeños productores conocer y utilizar herramientas de IA generativa, que hasta ahora no se utilizan. Esto llevó a una conversación sobre cómo es posible identificar qué herramientas están siendo utilizadas para la generación de contenidos, y cómo es posible

combatir su divulgación.

Aclararon que las herramientas de IA están evolucionando permanentemente y a ritmos acelerados, por lo que se necesita mucha flexibilidad para actualizarse en los métodos y mecanismos para identificar contenidos. Además hablaron de arquitecturas y políticas de autenticación de contenidos dentro de las plataformas que no apuntan a verificar su veracidad sino los comportamientos automatizados que permiten su viralización.

También mencionaron que de fondo se encuentra el problema de la verdad, quién define qué es verdad, cómo la entendemos y cómo la identificamos en nuestros diferentes trabajos.

Debido a que no hay un acuerdo respecto de si debería o no existir una regulación para la desinformación, se habló de la importancia de implementar estándares para fortalecer la confianza de las audiencias, y exigir mayor transparencia a las plataformas. Este fue el tema central de discusión en el plenario, ya que en el contexto de América Latina hay mucha desconfianza sobre iniciativas de regulación.

Al respecto se habló del contraste entre la desconfianza por la regulación frente a la narrativa mediática que aboga por la regulación. Sin embargo, las campañas de desinformación son encubiertas, por lo que la regulación no serviría para contenerlas. Desde otra perspectiva, la regulación no debería hacerse sobre los contenidos sino sobre las plataformas que capitalizan la difusión de desinformación, y además tienen la capacidad de identificar las formas en que se desarrollan estas campañas.

También se mencionó que aunque son muy deficientes los sistemas de justicia, sí hay avances en libertad de expresión a nivel del sistema interamericano, y que la libertad de expresión no es un derecho absoluto. Lo difícil es cómo detectar la responsabilidad en internet.

Medios comunitarios y cine

El grupo planteó que las herramientas de IA generativa permiten facilitar, reducir costos y agilizar procesos creativos y de producción, y así pueden tener un potencial democratizador, por ejemplo de los efectos especiales y otros elementos de pre y posproducción. Además pueden ser útiles para la investigación, por ejemplo en grandes bases de datos de violaciones de derechos humanos y ambientales, y para la verificación de información.

La IA generativa supone un cambio de paradigma respecto de lo que implica un proceso creativo. En vez de considerar que se ve limitado por la introducción de una máquina, puede expandirse hacia la idea de la co-creación entre diferentes especies, y concebir que el ser humano no es el único ser inteligente sino también los hongos, las plantas, las máquinas.

Como ejemplo, compartieron un caso de la Amazonía ecuatoriana, donde se han ido registrando los sonidos del bosque con fines de conservación, y se interpreta el entorno acústico de los bosques con IA, utilizando grabadoras en las copas de los árboles, conectadas a máquinas que de inmediato están produciendo una evidencia bastante amplia que permite definir si un entorno está siendo afectado. Las comunidades decidieron hacer prácticas artísticas colaborativas para resistir a que todo esté mediado por IA, para tener una pertenencia cultural.

A propósito de la introducción de máquinas, mencionaron que es importante cuestionar quién tiene la custodia de esa máquina, si es Dall-E o si tenemos la infraestructura propia, donde podemos instalar software propio y tener todas las herramientas y la información bajo

nuestro control. Además, varias personas mencionaron que ven la IA generativa todavía muy lejana de las posibilidades dentro de los movimientos sociales, pues ha llegado la desinformación pero todavía no se conocen sus posibilidades creativas.

Señalaron también algunos riesgos, por ejemplo la desaparición de las prácticas comunitarias alrededor de la generación de contenidos, entendiendo los medios comunitarios como una manera propia de crear, un espacio de encuentro interpersonal que genera narrativas, genera cine y que se aleja de una lógica comercial. Ante esto, se planteaba la pregunta de cómo podría la IA apoyar a los medios comunitarios en vez de limitarlos. ¿Cómo hacer que la IA nos permita encontrarnos?

También hablaron de la posible precarización del trabajo creativo, pues si bien la IA generativa puede reducir costos, ésto puede ser fácilmente aprovechado por la industria, de manera que muchos trabajos (en música, graficación y demás) pueden ir desapareciendo. Al final, una participante propuso esta frase: frente a los medios sintéticos los medios orgánicos.

Cierre del día 1

Raquel Vásquez del equipo de WITNESS hizo un recuento del día, planteando algunas temáticas a abordar durante el día siguiente, en torno a los riesgos y amenazas de los medios sintéticos, y soluciones en distintas áreas.

Algunas de las temáticas abordadas incluyen:

La importancia de la comunidad en diferentes áreas, por ejemplo el consentimiento informado, no solo de los datos sino del producto que se genera. Es decir, tanto del *input* como del *output* que genera medios sintéticos. También, cómo cambian las relaciones dentro de las comunidades de acuerdo a los contextos.

Se puso sobre la mesa las posibles ventajas del software de código libre para las comunidades para el acceso a ciertas herramientas y el aumento de la representación y el ejercicio de derechos digitales, sobre todo cuando estas herramientas tienen sesgos del mundo del norte. Frente a las soluciones de accesibilidad, ¿qué otros riesgos o desventajas deberíamos pensar?

Se habló de marcas de agua, según contexto y género. Cuando se manda una señal a la audiencia, ¿cómo diferenciamos entre libertad de expresión y parodia o sátira? ¿Cómo marcamos la intencionalidad del contenido, o qué es producido a través de IA? Esto, desde el punto de vista técnico y en cuanto a las comunidades con quienes trabajamos.

¿Cuáles son las soluciones que queremos priorizar como sociedad civil? ¿Cuáles son los retos de la regulación? ¿Cuáles son las posiciones o ideas que queremos avanzar de acuerdo al contexto? ¿Cómo lo conectamos por ejemplo con derechos laborales de las personas que participan en el proceso de creación de medios sintéticos o quienes participan del etiquetado en las bases de datos?

Se habló también de Identidad, anonimato, y la posibilidad de reacción cuando no hay acceso a estas herramientas. ¿Qué hechos entendemos como comunes? ¿Cuál es la relación con la verdad y la “posverdad”?

Relatoría del día 2

Sesión 1. ‘Espectrograma’ de riesgos

En la mañana se analizaron las amenazas que trae la IA generativa a partir de tres categorías no estrictas propuestas por WITNESS: los desafíos de la verificación, las temáticas, y las áreas específicas de IA generativa. Raquel Vázquez del equipo de TTO hizo una presentación general de los vectores de riesgo.

Los vectores de riesgo se dividían en tres categorías: 1. Desafíos generales de verificación, 2. Derechos humanos/áreas temáticas, y 3. Panorama / ecosistema informativo. Dentro de cada categoría se incluían los siguientes riesgos:

Desafíos generales de verificación	**Derechos humanos / áreas temáticas**	**Panorama / ecosistema informativo**
Fecha, hora o ubicación descontextualizada o no existente Contenido editado o manipulado Contenido fabricado	Violencia de género Guerra, conflicto y violencia Elecciones Salud pública Derechos sobre la tierra u otros derechos ESC Propaganda Ataques a la sociedad civil y los medios de comunicación	Comercialización Facilidad de uso y accesibilidad Volumen y variación *Liar’s dividend* Alucinaciones Multimodalidad Personalización e interacción en vivo con deepfakes

Luego de exponer algunos ejemplos se repartieron matrices de riesgo donde, individualmente, cada participante respondió a la pregunta ¿dónde/cómo las nuevas formas de creación y manipulación de medios sintéticos **modifican**, **expanden** o **crean** nuevas amenazas? Esta es la sistematización completa de las respuestas.

De acuerdo con la información consignada en las matrices de amenazas, las áreas temáticas que más comentarios recibieron fueron elecciones y violencia de género, seguido de ataques a la sociedad civil y a los medios de comunicación, propaganda y derechos sobre la tierra u otros derechos económicos, sociales y culturales. Se considera que tanto la descontextualización como la manipulación y la fabricación son desafíos presentes para la verificación de información, y en la mayoría de los casos se trata de expansión de amenazas, mucho menos de alteración o aparición de nuevas amenazas.

Algunas de las amenazas identificadas fueron comentadas en plenario, dando lugar a algunos temas que no estaban contemplados dentro de la matriz de riesgos.

Violencia de género

En coherencia con lo consignado individualmente en las matrices, en la conversación se señaló cómo la accesibilidad y facilidad de uso de herramientas de IA contribuyen a expandir tanto las prácticas como los efectos de la violencia de género relacionada con la tecnología. Se comentó sobre los riesgos que vienen con la posibilidad de mejorar la edición de imágenes y videos de pornografía no consentida. También se resaltó el doxing, la comercialización de packs, o el ejercicio de violencia simulando voces, bien sea para acoso o para violencia vicaria, por ejemplo, o para estafar a madres buscadoras de personas desaparecidas.

Se señaló además cómo la brecha digital de género puede amplificar los efectos de estas violencias, cuando no hay una capacidad de respuesta inmediata frente a la difamación o la difusión no consentida de material íntimo manipulado o fabricado.

Elecciones y propaganda

Aunque en las matrices individuales el tema de elecciones fue el que recibió más comentarios, durante el plenario se amplió la conversación al tema de propaganda, segmentación de públicos y burbujas informativas—algo que puede ser utilizado en el contexto electoral, por ejemplo para persuadir e influenciar votantes, pero en general en la propaganda gubernamental y el posicionamiento de discursos y narrativas políticas.

De nuevo, la sensación predominante no es de un fenómeno nuevo sino de expansión de un problema anterior al uso y popularización de herramientas de IA generativa. El uso de IA se incluye dentro una serie de tácticas de desinformación, y potencia las campañas de posicionamiento de narrativas globales (por ejemplo asociar a los ucranianos con nazis en el contexto de la guerra).

Con relación a esto, se habló de la radicalización de burbujas de contenido en población que ha vivido un trauma como la búsqueda de refugio. De nuevo, no es algo que se haga con IA, pero la IA favorece y facilita este fenómeno. Así mismo, potencia las posibilidades de la propaganda personalizada en redes sociales.

Derechos sobre la tierra y otros DESC

Si bien en las matrices de riesgos se incluyeron pocos comentarios relativos a los derechos sobre la tierra u otros derechos económicos, sociales y culturales, durante el plenario se abordaron algunos temas que caben en esta área temática, y se señaló su ausencia dentro del mapeo de riesgos que adelanta WITNESS.

Uno de estos es el tema de derechos de autor cuando hay mediación de una herramienta de IA generativa, y cómo excluye los procesos de creación comunitaria. O cómo los beneficios en términos de facilidad y reducción de costos en la producción audiovisual pueden ser aprovechados mucho más por los grandes conglomerados de industrias creativas, y redundar en mayor precarización laboral para quienes participan de distintos momentos del proceso de producción.

Por otra parte, varias personas mencionaron lo que implica la IA para el calentamiento global, que se suma a una crisis ambiental que venía de antes. Al respecto, se señaló que el entrenamiento de IA requiere de servidores inmensos que necesitan cierto tipo de refrigeración y algunos están ubicados en el polo norte. Ante esto, un participante planteó el problema de la polución icónica y preguntó, ¿tiene sentido seguir produciendo imágenes?

Este punto se conecta con otro tema: la reducción de la variabilidad y homogeneización del contenido, lo cual en el mediano y largo plazo significa una reducción de los patrones culturales, algo que en palabras de otro participante evidencia un patrón colonial, donde las diversidades no pueden ser representadas, por lo que parece imposible utilizar la IA para generar futuros. Aunque quizás, decía, haya espacios de oportunidad en esa invisibilidad.

Y desde el punto de vista educativo, se planteó como problemática la narrativa generalizada de que la IA puede solucionar problemas sociales. Una participante mencionó que actualmente se están usando herramientas, sobre todo de texto generativo en los entornos académicos, pero hay demasiada confianza en estas herramientas, y se parte de la idea de que la tecnología siempre tiene la razón.

Sesión 2. Reflexiones sobre Prácticas Responsables para Medios Sintéticos.

En la segunda sesión, Jacobo Castellanos hizo una presentación breve sobre el marco Prácticas Responsables para Medios Sintéticos: Un marco para la acción colectiva de la Partnership on AI, y compartió versiones Marco traducido a español. La presentación completa está disponible aquí.

A partir de estas presentaciones y de los puntos propuestos por el PAI, por grupos se analizaron tres casos en América Latina.

Caso 1. Argentina

Un artista en Argentina agarró imágenes de archivos de los progenitores de los bebés que fueron desaparecidos por la dictadura, y que todavía hoy son buscados por las abuelas de Plaza de Mayo. Agarró una foto del padre y de la madre, y les agregó un filtro de envejecimiento en Midjourney para ver cómo sería el aspecto de una persona de cuarenta años, que es hijo o hija de esos dos progenitores, para intentar buscar la semblanza. El proyecto se puede consultar en la cuenta de Instagram @iabuelas.

Es un caso muy reciente, y por lo que se conoce online, las abuelas de Plaza de Mayo dijeron que les gustaba la idea para divulgar y reactivar la búsqueda, pero que no es algo científico. En muchos casos faltan datos sobre si el bebé era mujer u hombre.

Quién está involucrado / Cuáles serían las partes

La persona creadora del proyecto, el artista; las familias, las personas desaparecidas; y el gobierno, porque es un tema público y tiene una responsabilidad en el marco de garantía y protección de los derechos humanos; Instagram, porque la cuenta original se creó allí y es la plataforma principal donde se distribuye. Sin embargo es muy sencillo sacar el contenido de allí, haciendo una captura de pantalla de las fotos y poniéndola en cualquier otra parte.

Se discutió si cabrían aquí las organizaciones de derechos humanos, pero como se trata de establecer en quién debería recaer la responsabilidad se concluyó que no son una parte involucrada.

También se conversó sobre la posibilidad de hacer una prueba con nuestras fotografías. Podría pasar que no nos parezcamos, pues los datos de Midjourney están influenciados por

gente súper alta, súper guapa, musculosa. Esa plataforma también podría tener una responsabilidad.

Cuál es el tipo de uso: es un uso razonable, malicioso

Es un proyecto artístico, pero en la circulación cambia el tipo de uso. En términos de libertad de expresión, aunque sea un tema político está protegido como proyecto artístico. Es difícil hablar de un uso puramente artístico porque está hablando de algo social, y además es un tema sensible. Dónde está la línea entre lo artístico y lo político, lo social, los derechos humanos. También es un ejercicio de memoria

Aunque tanto el creador como las abuelas han dejado claro que no es una prueba científica sino un proyecto artístico, podría pasar que para las familias de las personas retratadas significase una oportunidad para encontrarles.

Consentimiento

Parece que el proyecto se hizo inicialmente sin consentimiento de las familias. Cuando lo presentó y salió público, las abuelas dijeron que era un uso artístico, no científico, y que daba visibilidad. El proyecto es muy claro en que “se propone colaborar” no dice “ en colaboración”.

Las imágenes son creadas a partir de fotos antiguas de los padres, que hacen parte de un archivo público, junto con los nombres y datos de las familias. En este caso hay distintos niveles de consentimiento y es necesario ver el código de conducta del archivo y cuáles usos están permitidos, porque puede tratarse de una base de datos completamente abierta, donde este tipo de uso no se ha considerado.

Se mencionan los casos de Colombia, donde se incentiva el trabajo con los archivos de memoria pero se autoriza y limita el uso previamente, y México donde también se están explorando los archivos de los años 70, y una herramienta de IA como esta sería útil para apoyar a las personas buscadoras, que continúan la búsqueda después de quince años o más.

No es claro si cambia algo con el hecho de que las abuelas sigan vivas, pero se menciona como posibilidad el derecho a la desconexión cuando una persona fallece, esto es, a que nadie utilice las fotografías de esta persona. En México al menos, si el objetivo de difusión va a traer un mayor bien que el mal que puede generar, es permitido.

Transparencia

El proyecto está en Instagram, y se debe solicitar entrada para verlo. En las publicaciones el artista utiliza los hashtag #midjourney y #ia, pero no tiene una marca de agua ni nada.

Los resultados generan inquietudes sobre qué pasaría con otras plataformas. El proyecto comunica muy poco sobre el proceso y los criterios de creación, más allá de la herramienta utilizada. Sabemos que agarra las dos fotos, las combina y les agrega un filtro, pero no sabemos si ha habido una alteración al juntar las dos fotos, ya que los criterios de envejecimiento también son un asunto de clase. Si es un chofer, va a tener la cara muy diferente después de veinte años.

Mecanismos para minimizar el mal uso

Es recomendable hacer mucha publicidad y difusión, para que todo el mundo sepa que es IA. Si desde el inicio el relato es muy claro, es más difícil que sea mal utilizado.

También es necesario hacer mucha pedagogía sobre cómo puede fallar fácilmente. Explicarle a una abuela todos los fallos que puede tener, pues este es un segmento de la población donde hay que hacer más alfabetización.

En el caso de personas desaparecidas, es importante mantener la ficha de localización de la persona y darle visibilidad. A nivel de incidencia e imaginario social, es parte de las funciones de esas fichas.

Caso 2. Colombia

La oficina de Amnistía Internacional en Noruega hizo una campaña en redes sociales para sensibilizar sobre el abuso policial en el segundo aniversario del inicio del Paro Nacional en Colombia en 2021, el 28 de abril. Para esta campaña utilizaron imágenes hechas con IA junto con mensajes sobre el uso de la fuerza por parte de la policía y la necesidad de una reforma policial. La campaña fue sumamente criticada por utilizar elementos visuales como una bandera errónea, y el equipo que trae la policía, que no está vigente. La campaña fue bajada luego de recibir muchas críticas, pero puede leerse el reportaje aquí.

Quién está involucrado / Cuáles serían las partes

La ONG que está haciendo una campaña, y las plataformas de redes sociales donde se difunde la campaña.

Cuál es el tipo de uso: es un uso razonable, malicioso

Una participante mencionó que se trata de una narrativa criminalizante, ya que ponen los ojos en los defensores y no en la policía. Otro participante expresó que ve cierto puritanismo en el movimiento social, respecto del uso de IA. Si Amnistía hubiera dicho que le habían encargado a un artista hacer las imágenes la respuesta habría sido otra.

Con la referencia de distintas corrientes en el teatro y el documental, se planteó que es necesario diferenciar entre algo que es mentira y algo que es ficción. Sin embargo, también se planteó que no es lo mismo algo con pretensión artística, o periodística y de denuncia, donde lo que más se tiene que proteger es la credibilidad.

Refiriéndose a la publicidad engañosa, alguien planteó la pregunta: ¿cuál es la intención de Amnistía al utilizar este tipo de imágenes? ¿Está demostrando lo que pasó en realidad o cuál es la agenda con esa imagen, que puede ser manipuladora?

Ante esto, algunas personas plantearon la posibilidad de que quienes diseñaron y publicaron la campaña, simplemente estuvieran evitando un problema de derechos de autor, o que fuera una simple expresión de ignorancia y poca importancia sobre el contexto de Colombia.

Consentimiento

Como se planteó en el debate en redes, hay mucho material disponible fotográfico que salió durante las protestas. La interpretación artística es válida, pero si es realista se pueden usar las fotos. Las caras de quienes salen en las imágenes generadas con IA responden a un modelo eurocéntrico.

Amnistía Internacional respondió que defendían el anonimato, pero había personas que no les importaba que las mostraran, y esto fue manifestado en el debate y las críticas. Es un caso muy diferente por ejemplo al trabajo de Forensics Architecture con los 43 de

Ayotzinapa, donde no hay imágenes de evidencia, y la recreación con imágenes digitales es un apoyo necesario.

Transparencia

Las imágenes traen una etiqueta minúscula, fácil de recortar. A propósito, se discutió sobre los sistemas de etiquetado de los alimentos, que al final no están cumpliendo su función porque en el mercado todo está etiquetado, pero no hay alternativas, pues no tiene sentido etiquetarlo todo. El sistema de etiquetado de alimentos falla en su objetivo de darle más claridad y herramientas de juicio al consumidor

Además, se mencionó que enfocarnos en las marcas de agua sobre los productos finales hace que se nos olvide todo lo que pasa atrás en el desarrollo de IA. Tenemos que entender que esto es un entramado mucho más complejo.

También se mencionó, como problemático, cuánta gente se sienta a analizar una imagen, si se están consumiendo cantidades enormes de datos. Se necesitan también programas de alfabetización digital, que aborden desde muy distintas dimensiones la formación de adultos.

Mecanismos para minimizar el mal uso

Tomando como referencia el teatro invisible, una técnica que hace parte del teatro de los oprimidos, se planteó la importancia de que la misma imagen señale que es falsa, exagerada, que es una expresión más artística. Jugar con este contexto en el que estamos donde no sabemos qué es real y qué no, pues usar medios sintéticos en denuncia sin hacerlo explícito es problemático.

Alguien mencionó que mínimamente debieron contactar a la oficina de Amnistía Colombia o América Latina para corroborar si las imágenes funcionaban

Caso 3. México

El medio de comunicación Milenio, en México, comparte un video creado con IA, al parecer realizado por simpatizantes de la candidata a la presidencia Xóchitl Gálvez, donde le responde a la entonces jefa de gobierno de la Ciudad de México, Claudia Sheinbaum. La nota se puede ver aquí.

Quién está involucrado / Cuáles serían las partes

Quien creó el video: Aunque no es posible establecerlo, se sugiere la posibilidad de que sea la misma campaña de la candidata, ya que la creación de cuentas falsas es algo que se acostumbra hacer en las operaciones de influencia. Desde una lectura cultural, entendemos que este video es un producto de marketing político, tenga o no IA.

Los medios de comunicación y específicamente Milenio que compartió el video completo y así amplifica el mensaje: Si bien debería regirse por normas de contenido electoral, están cubriendo la noticia de un video que se ha hecho público en redes sociales. Pretender que no lo comparta podría implicar un acto de censura, pero también se entiende que la lógica del medio es llamar la atención de su audiencia, competir por atención. En el caso de este medio, llega a muchísima gente que no le va a prestar mucha atención a la veracidad, técnicas u origen del video, porque funciona con burbujas informativas

También se mencionó que con este tipo de contenidos, es común que no se difundan

primero desde los medios afines porque van a ser criticados. Los comparten primero en redes sociales, donde comienzan a viralizarse y son captados por influencers. De ahí siguen movilizándose y son captados por medios digitales. En ese punto ya está suficientemente lavado el origen de los contenidos.

Cuál es el tipo de uso: es un uso razonable, malicioso

Se considera que el contenido no es malicioso sino que es propaganda y ese formato hace parte de una tendencia creciente. Está ubicado en una zona gris por varios motivos, pero en el marco de una campaña política hay derecho a persuadir a los votantes, eso no tiene de malo. El punto es que es una táctica que apunta a lo emocional, y que no es posible identificar el origen. Aunque puede ser un contenido orgánico, auténticamente creado por simpatizantes, lo que vemos en el grupo en el video es que es un producto de propaganda electoral.

Esta es una noticia (la creación de un video con IA, en apoyo a una candidata) y el medio hizo lo que podía para manifestarlo, pero si quisieran ser más responsables podrían haber puesto el video y un texto de descripción de lo que es el video, o no dejar el video completo, o silenciarlo.

Ante la pregunta por la responsabilidad de los medios, pero también de las herramientas con que se creó el video, se considera que ese no es el punto porque no se puede predecir cómo va a ser utilizadas una herramienta. El problema ahí es que si pretendes limitar la capacidad de la herramienta, implica una tecnología de vigilancia sobre el comportamiento con esa herramienta. Si no se pueden usar caras de políticos, la herramienta necesita tener una base de datos con esa información, que se va actualizando cada cierto tiempo. Se trata de bases de datos enormes (y el equipo de cómputo necesario para procesarlas), pero además se requiere que la herramienta esté analizando permanentemente cuándo se debe levantar una alerta, cuándo hay match entre la base de datos y el comportamiento de las personas usuarias.

Consentimiento

El primero problema con este contenido es que hay muchos supuestos que no manejamos, por eso se encuentra en una zona gris. Alguien más está poniendo sobre la candidata palabras que ella no dijo. En este caso no están poniendo ninguna palabra problemática, están apoyando su campaña, pero sería muy distinto si la estuvieran acusando de algo falso, porque podría apelar a su derecho a su reputación.

Alguien mencionó que el problema está en que el consentimiento es distinto en las figuras públicas y en las relaciones íntimas. No se puede considerar de la misma manera, pero además, en este caso es difícil creer que el contenido es de auténtico apoyo por parte de seguidores, por lo que tampoco es tan fácil suponer que están poniendo palabras en su boca, sin su consentimiento en el marco de una campaña electoral.

Transparencia

Se plantearon tres asuntos relevantes. El primero es que el medio de comunicación tiene la responsabilidad de decir que el video está hecho con IA, así sea o no evidente, por un tema de alfabetización porque hay mucha gente que no sabe lo que es IA. En este caso el medio dijo dos verdades: que es IA, y que no lo hizo la candidata, pero igual está amplificando una propaganda. Tampoco conocemos los intereses del medio, ni si ellos saben si se podría tratar de una campaña encubierta. Podría existir una regulación más clara frente a este tipo de imágenes, pero entonces estaríamos hablando de moderación de contenidos y no tanto

de transparencia.

Respecto de las variables específicas en materia de transparencia, se planteó que quien crea los contenidos tuviera la obligación de utilizar arquitecturas de autenticidad, sin caer en la regulación. Sin embargo, para este tipo de campañas encubiertas, no se podría utilizar esto, quien crea los contenidos perdería su trabajo, pero además el problema de alfabetización mediática persistiría. Y desde el punto de vista de una participante creadora, el uso de marcas visibles impuestas por las herramientas para la creación o edición de contenidos bajaría la calidad del producto. No quedaría tan chévere.

Por último, se habló de la obligatoriedad para las herramientas de utilizar marcas de agua y otras arquitecturas de autenticación en sus modelos, y cómo esto puede estar en contradicción con las políticas de open source, que permitirían fácilmente a alguien retirar las marcas para reentrenar una modelo o utilizarlo. Porque el modelo de OpenAI, por ejemplo, es completamente cerrado, mientras que Llama de Meta es totalmente abierto, con algunas cláusulas, pero permite hacer muchas cosas. Frente a esto, es importante cuestionar la posición dominante de algunas empresas que han desarrollado modelos. Quién tiene capacidad de implementación (por ejemplo Runway o Midjourvey). Entonces la marca de agua puede operar como una forma de control, como un sello de calidad similar a como se ha utilizado el estándar DRM, que protege los contenidos que tienen derechos de autor, limitando el acceso a la cultura.

Mecanismos para minimizar el mal uso

En este caso es fácil para el medio de comunicación no asumir ninguna responsabilidad. Amplifica pero atribuye a la IA y a los simpatizantes. Puede ser algo grave, pero lo vieron millones de personas. Podría pasar en otro caso hipotético, se mencionó, si el medio de comunicación está diciendo tener evidencia de una guerra mundial, y no dicen que el video es de inteligencia artificial. Esto también puede pasar.

Sesión 3. Transparencia, marcas de agua y procedencia

Para empezar el último bloque de trabajo, Raquel Vásquez del equipo de WITNESS habló sobre diferentes técnicas, visibles al ojo humano o solo reconocibles por máquinas, para marcar los contenidos creados o manipulados con IA. La presentación completa está disponible aquí. Se propuso a las personas participantes pensar desde el punto de vista sociotécnico cuáles son los beneficios o dificultades de las diferentes opciones.

Algunas personas preguntaron sobre la facilidad de retirar este tipo de marcas, por ejemplo cuando se suben a las redes sociales, alterando el código o simplemente tomando pantallazos, y la presentadora señaló que parte de la investigación que están haciendo las compañías en este momento busca identificar técnicas que hagan a este tipo de marcas más resilientes. También señaló que el objetivo de este encuentro no es adentrarse en el debate técnico, que es el que están teniendo las compañías, sino plantear un debate desde el punto de vista social.

Por último mencionó los desarrollos recientes en lo que, hasta ahora, se ha denominado marcas de agua invisibles dinámicas, que consisten en un set de metadatos adjuntos al contenido digital, donde se va registrando un histórico de la actividad que ha pasado por dicho contenido.

Con esta información a mano, siguió una conversación a partir de preguntas sobre las

implicaciones de este tipo de enfoques para el trabajo que realizan las personas participantes, atendiendo especialmente a temas de privacidad y anonimato, acceso, confianza, capacidad de elección, concentración de poder, usos indebidos y abusos por parte de reguladores, y confusión.

Inicialmente se planteó una preocupación por la posible inclusión de datos personales en los contenidos, por el efecto silencioso que esto puede tener en la confianza de las personas, sobre todo en regímenes represivos, para generar o publicar contenidos. Además se habló de otros elementos, independientes de las marcas de agua, que están presentes en los contenidos desinformativos, así como de las muchas técnicas que ya existen y seguramente seguirán existiendo, para saltar dichas marcas.

Una participante planteó como inquietud por qué estamos ahora hablando de marcas de agua en las imágenes si llevamos décadas editando contenidos con Photoshop, y la única diferencia es que ahora las imágenes se producen más rápido. A esto respondió la presentadora que quizás es un elemento de temporalidad, mientras nos adaptamos como sociedad a ver las imágenes producidas por IA. También hay un elemento contextual, dependiendo de las expectativas que hay en los distintos contextos, ¿es en ciertas plataformas o en ciertos géneros? ¿O es en todos los contenidos? ¿O quizás el trabajo debería ser al revés? Por ejemplo marcar los contenidos donde hay estas expectativas.

Se trajo a colación el ejemplo del teatro invisible, donde se crea una situación ficticia aunque el público no sabe que es ficticio, pero al final se explica al público que no era una situación real, para darle elementos de juicio. Así mismo, un participante comentó que cuando utilizan Midjourney en su trabajo, mucha gente piensa que es trabajo del diseñador, y entonces han decidido poner una etiqueta, pensando en hacer pedagogía con el público, mientras aprenden a identificar las imágenes producidas con IA, y señaló que las personas que acostumbran a trabajar con esta herramienta ya saben identificar su estilo.

Haciendo alusión al problema de los etiquetados alimentarios, alguien propuso no etiquetar las manipulaciones sobre los contenidos, sino la autenticidad. Frente a esto, alguien más señaló que el debate sobre etiquetado es reduccionista pues carga la responsabilidad en el usuario final, y deja por fuera del debate a las comunidades que desarrollan IA, limitándose a las grandes empresas que tienen la capacidad de etiquetar. Otra persona señaló que los sistemas de etiquetado sirven para fortalecer a las empresas mientras que la agencia y la autonomía de las personas usuarias se ve socavada. Y alguien más planteó que si bien las empresas que generan contenidos maliciosos no van a etiquetar. una parte del etiquetado serviría para entrenar el ojo y educar audiencias para cuestionar los contenidos.

La parte final de la conversación giró en torno a que este debate está teniendo lugar en la industria y en el Norte Global, y cómo desde los sures solo estamos participando una pequeña élite, pero seguimos sin encontrar solución sobre cómo escalar estas u otras soluciones. Entre las personas participantes hubo opiniones encontradas sobre la necesidad, pero sobre todo la eficacia de enfatizar en la alfabetización digital y otras políticas públicas educativas. Esto, si bien se considera importante, no tiene la capacidad de ser implementado a la velocidad que avanzan las tecnologías.

Sesión 4. Construyendo el futuro de los medios sintéticos en América Latina

Para la parte final se propusieron cuatro grupos temáticos para discutir sobre posibilidades a futuro en el contexto de América Latina, a partir de las conversaciones que se tuvieron durante todo el encuentro. Cada grupo compartió brevemente sus conclusiones antes de ir al plenario final.

Plataformas y legislación

El primer grupo empezó definiendo a qué tipo de plataformas se refieren con relación a los medios sintéticos. Por una parte las que permiten generar contenidos sintéticos, tales como OpenAI, Midjourney, Dall-E, y por otra parte las redes sociales y sistemas de mensajería que permiten difundirlos. Aclararon que su conversación se enfocó en las plataformas de redes sociales, y cómo los medios sintéticos aumentan problemas que ya existían de antes en estas plataformas, como la gestión de contenido parodia, suplantación, autenticidad, entre otros.

Hablaron de las diferencias en la manera como dos plataformas, Meta y Tik Tok, incluyen en sus políticas y términos de servicio el tema de los medios sintéticos. Meta se concentra en la intencionalidad de los contenidos, haciendo excepciones para sátira y parodia, lo cual no resuelve las dificultades para identificar esta intencionalidad. Tik Tok, en cambio, pide el uso de etiquetas para el contenido sintético, y distingue entre personas públicas y no públicas, protegiendo la imagen de estas últimas.

También comentaron algunas regulaciones existentes, como el Código de Práctica en línea en la Unión Europea, que incluye un pacto sobre contenido generado con IA, y varios intentos de regulación en América Latina. Uno de ellos es el proyecto de regulación de plataformas de Brasil, que se encuentra en pausa actualmente ya que ha sido criticado por crear un supuesto ministerio de la verdad estilo 1984.

También está la Comisión asesora contra la desinformación, creada por el Ministerio de Ciencia y Tecnología de Chile, cuyo objetivo es dar recomendaciones sobre cómo hacer regulación, aunque también se le ha señalado de poder convertirse en una comisión de la verdad. Y el Comité técnico de fake news en Perú, que funciona solo en periodos electorales pero tiene la capacidad de decir a las plataformas qué contenido bloquear. Frente a las discusiones sobre moderación de contenidos y autorregulación de las plataformas, el grupo planteó como una necesidad que las plataformas cuenten con protocolos claros sobre cómo realizan la verificación de contenido sintético, no como paredes legales sino para conectar sus esfuerzos con los de organizaciones de fact checking, y así abordar el problema de la intencionalidad, por ejemplo en casos de parodia o sátira.

Durante la discusión, se mencionó la importancia de considerar que la moderación de contenidos dentro de las plataformas no siempre es algorítmica sino que se terceriza. Se contratan trabajadores precarizados que después de un tiempo tienen síntomas de estrés postraumático. Es importante tener más información sobre cómo las empresas hacen ese tipo de trabajo.

Volviendo a los contenidos, se consideró necesario que mejore la accesibilidad de los informes de transparencia de las plataformas, pues muchas veces están en formato pdf. Los datos que incluyen deberían ser abiertos, claros y accionables; las metodologías también deberían transparentarse, y debería vincularse a la sociedad civil en la definición de estos informes. Además, es importante que se incluya un apartado sobre medios sintéticos.

Por otra parte, señalaron que las plataformas tienen una responsabilidad frente al daño que producen las campañas de desinformación. Reconocen que hicieron daños al público pero no hay propuestas de reparación. Propusieron que se creen fondos para trabajar estos temas desde la sociedad civil, de manera independiente.

Por último, señalaron que en América Latina se puede legislar pero no hay poder para

aplicar la ley, pero actualmente se están dando debates importantes en espacios globales, donde se requiere mayor participación desde el sur.

Expresión creativa de impacto para la defensa de los ddhh

Este grupo hizo un balance positivo y negativo sobre los medios sintéticos. Se reconoció que la IA permite hacer todo más rápido y por eso es necesario distinguir entre el activismo y la creación, pues por ejemplo en una protesta es importante la inmediatez, pero los procesos creativos llevan más tiempo.

Hubo posiciones encontradas respecto de cómo la IA puede ayudar a aumentar la creatividad. Por una parte, los procesos pueden desarrollarse más rápido y a menores costos, pero ésto es un valor capitalista que no necesariamente responde a los tiempos y ritmos necesarios en los procesos creativos. Sin embargo, la IA permite organizar material, es decir que hay un cambio de paradigma respecto de la creatividad ilimitada, pues ahora consiste en decidir sobre posibilidades. Con esto se corre el riesgo de que todo sea muy genérico y que las personas creadoras se frustren porque no tienen los resultados que esperaban, pero también se pueden explorar inmensas posibilidades.

Por otra parte, se planteó que la introducción de nuevas tecnologías suele ser capitalizada por quienes tienen más poder y así aumentan las brechas. De ahí que sea necesario preguntarse por cómo la IA puede ayudarnos a ser más resilientes y tener mayor impacto como personas activistas.

Al respecto, se reconoció que aunque la IA se esté desarrollando hace décadas, llega de una manera muy invasiva a nuestros territorios, y no ha habido oportunidad de educarnos para ser parte o tener una postura crítica. En lugares donde las personas ni siquiera saben leer y escribir es necesario pensar en estrategias para adaptarse a estas tecnologías que llegarán en un tiempo cercano. E incluso para las personas que sí escriben, hay un problema en cuanto a la redacción, y es necesario también aprender a escribir una descripción (*prompt*), a relacionarse con la IA y darle indicaciones bien.

En el caso de los medios comunitarios, también se planteó como problemático que estas tecnologías limiten los espacios de encuentro humano. Sin embargo, también es posible producir y trabajar con imágenes propias dentro de las comunidades, y concebir la imaginación como un proceso compartido a través de procesos de visualización de grandes cantidades de imágenes.

En general, el grupo consideró que la educación es fundamental para aprender a cuestionar los contenidos y también para crearlos, pero los sistemas educativos han colapsado. Al respecto se propuso, como una solución utópica, trabajar en políticas públicas para empoderar a las comunidades directamente, y romper con el ciclo tradicional de que las tecnologías pasen primero por la academia y luego lleguen a las comunidades. Esto, considerando que tecnologías como la IA pueden utilizarse de manera creativa para la defensa de los derechos humanos, más allá del campo artístico.

Se propuso entonces entender las soluciones reconociendo las diferencias:

Primero, no satanizar las plataformas y las opciones que ofrecen las herramientas de IA generativa. Las instituciones universitarias, por ejemplo, enfocan sus esfuerzos en identificar si algo fue plagio o en liberarse de responsabilidad por el uso de contenidos protegidos por derechos de autor, en vez de preguntarse por los cambios de paradigma en la enseñanza, o conocer las metodologías emergentes para la construcción de

conocimiento aprovechando estas herramientas.

Segundo, cuestionar el uso. Es importante entender que la relación con la IA en realidad es con quien entrenó la IA, y esto tiene problemas asociados a los datos, a la privacidad, a la representación de las personas. También es necesario criticar el impacto que tienen estas tecnologías en lo laboral. Se trata de máquinas que reemplazan, de manera gratuita o a bajo costo, ciertos trabajos, contribuyendo a su precarización.

Y tercero, explorar qué existe por fuera de la tecnología.

Habilidades y herramientas periodísticas y de investigación

El tercer grupo conversó sobre las herramientas que ya eran útiles para el trabajo periodístico antes de la IA generativa, tales como los subtítulos de Youtube. Y sobre las oportunidades que ofrecen herramientas de IA, tanto para la clasificación y análisis de información, como para la generación de contenido. Herramientas como Midjourney han sido muy útiles para la expresión gráfica, y ChatGPT para la edición y revisión de textos.

Sobre esto, señalaron que es importante saber usar estas herramientas como asistentes, nunca delegando las revisiones de contenido finales a éstas. Es necesaria una alfabetización para que efectivamente se democratice su uso, entendiendo cuál es su utilidad. Comentaron que hay muchos mitos alrededor de la IA generativa, pero al usar cualquiera de estas herramientas un periodista se dará cuenta de todas las limitaciones que tiene. Propusieron hacer entrenamientos con organizaciones de la sociedad civil y derechos humanos, ya que estas herramientas también tienen un enorme potencial para la desinformación y eso plantea enormes desafíos.

Debido a la velocidad con que es posible producir y divulgar grandes volúmenes de contenido, no hay capacidad de respuesta en momentos críticos como eventos electorales. Es difícil establecer dónde enfocar los esfuerzos de detección, y las técnicas utilizadas deben actualizarse permanentemente ya que estas tecnologías avanzan también muy rápido.

Sin embargo, aunque parezca incontenible, el volumen de este tipo de contenido puede jugar en contra de las campañas de desinformación porque aparecen errores evidentes, por ejemplo distintos medios que comparten desinformación, pero todos tienen la misma IP. Por eso no es posible hablar de una solución única frente a la desinformación. Las marcas de agua pueden ser una alternativa y pueden funcionar en algunos casos, pero durante la difusión en redes es posible retirarlas, así que es importante contar con otros medios y técnicas de verificación.

Además, es necesario fortalecer las redes de apoyo y respuesta entre comunidades, medios de comunicación, verificadores y organizaciones de la sociedad civil. Esto implica bajar las tecnologías del norte, apropiarlas, desarrollar capacidades para generar contenido nuevo, mantener bases de datos locales (aunque esto es bastante complicado), y también verificar la información que reciben.

En el grupo **se planteó que la polución informativa es también una oportunidad para que las voces realmente confiables resurjan**, así que una labor importante para los periodistas es desarrollar estrategias para fortalecer su credibilidad, por ejemplo siendo más cuidadosos en la manera como se utilizan herramientas emergentes como la IA generativa, pero también guiando a las audiencias hacia fuentes más confiables.

El trabajo en red se ve como una necesidad más urgente que la incidencia legislativa o con

plataformas, ya que los criterios de desinformación varían y, en muchos casos en América Latina, lo que se considera desinformación para los gobiernos es contrario a lo que se considera desde el periodismo.

Alfabetización pública y de medios

Este grupo empezó por discutir qué se entiende por alfabetización digital en los contextos latinoamericanos, pues la idea de *media literacy*, enfocado en la lectura de imágenes, no es suficiente para abordar otras ausencias que están presentes, desde comer o saber leer y escribir. Incluso entendiendo internet como un proceso social, se requieren bases técnicas mínimas no solo para usar una herramienta, sino para habitar ese espacio, y en la región existen enormes brechas en ese sentido.

Por otra parte, el problema de la desinformación no llegó con la IA sino que viene de mucho antes, así que la alfabetización tiene que venir de mucho más abajo, desde lo que se conoce y a lo que se tiene acceso inicialmente, que es la falta de credibilidad en los medios tradicionales y la falta de educación.

Ese es un problema que podría abordarse desde políticas públicas en educación, pero solo en el largo plazo. En el corto y mediano plazo, el grupo planteó como urgente priorizar a los a las redes comunitarias y a los medios independientes, alternativos y comunitarios, para que se aproximen y utilicen de manera crítica las herramientas de IA generativa. Estos grupos han sido muy importantes durante los levantamientos sociales recientes en los países de América Latina, y son la fuente de información más confiable porque están cerca de las comunidades, que también tienen necesidades muy diferentes. Los medios comunitarios son un actor con capacidad de movilizar transformaciones.

Se propone entonces hacer manuales de uso, en distintas lenguas, con un enfoque de alcances y limitaciones, utilidad y cuestionamientos a las herramientas digitales y de IA. Esto, sumado a planes para la instalación de capacidades, de manera que sean aprendizajes replicables, muy cercanos al modelo con el cual funcionan las redes comunitarias, aunque este modelo tiene un problema de escala.

Por eso, a largo plazo se considera necesario enfocar el trabajo en políticas y planes educativos, colaboración institucional entre sociedad civil y gobiernos para que las perspectivas críticas, pero también actualizadas, estén presentes también en las aulas. Sobre esto, mencionaron que es importante superar las miradas punitivistas, sobre todo en los entornos educativos, frente al uso de herramientas de IA generativa como ChatGPT, y que esto hace parte de un proyecto de alfabetización crítica.

Plenario final

El equipo de WITNESS preparó un listado de temas que surgieron durante el encuentro, y propuso una ronda de observaciones para detallar o ajustar este listado, que podría servir como un plan de acción. El listado de temas y recomendaciones está aquí y algunos de los comentarios incluyen:

Considerar las implicaciones ambientales y las infraestructuras necesarias para el sostenimiento de los sistemas de IA.

Cómo considerar la explicabilidad (*explainability*) para fortalecer los sistemas de IA basados en la comunidad.

Aplicabilidad para usos positivos es limitada si sólo se consideran medios sintéticos. Si bien el enfoque ha sido estratégicamente en usos y abusos, se abrirían más posibilidades considerando otras técnicas.

Cómo se manifiesta la IA basada en la comunidad. Por una parte, puede ser riesgoso porque pareciera que nos estamos refiriendo a comunidades autónomas, independientes, resilientes, pero también se podría utilizar por comunidades que están haciendo micro segmentaciones en bases de datos para utilizar esta información en campañas durante procesos electorales. Al respecto, se planteó que es muy importante trabajar y fortalecer a los periodistas, pero sobre todo a los medios y redes comunitarias, porque su trabajo ya se basa en la confianza social, no solo técnica.

Por otra parte, cuando hablamos de comunidades indígenas, por ejemplo, se trata de comunidades que ya tienen cierta experiencia en el manejo de sus datos, y que proteger sus datos con cierto recelo, es decir que sí tienen experiencia y podrían beneficiarse de estas herramientas, pues quizás en unos años no se requieran equipos de cómputo tan grandes para procesar los datos locales. De hecho existen ya LLM que pueden ser procesados en computadores personales, y también hay muchos desarrollos en IA distribuida, que podrían ser utilizadas por las comunidades para gestionar sus datos por sí mismas, y no para el beneficio de otros entes con poder, ni tener tanto impacto ambiental.

La frase que fue mencionada durante el taller, “frente a los medios sintéticos, medios orgánicos” es aplicable a estos modelos de gobernanza de IA desde las comunidades, pero también al valor central que tienen los derechos laborales y la representatividad, que también se comentó durante el encuentro. Se propone agregar a la frase “medios cyborg”.

Es necesario involucrar a más personas tecnólogas e ingenieras en esta conversación, personas que tengan la capacidad de ejecutar estas propuestas. Pero también es necesario que distintos grupos sociales se insertan en los espacios de desarrollo de tecnologías, y también en cuerpos de estandarización. Hay actualmente muchos espacios donde se están debatiendo temas de IA y ética, también sería relevante llevar lo que se ha identificado aquí a esos espacios, pero también alimentar estos espacios con lo que se están discutiendo allá.

En general, es importante acercar a las comunidades técnicas con las no técnicas, y promover que se entiendan mejor las herramientas de IA para poder adaptarla a nuestros contextos, pero también porque hay mucha mitificación y desinformación sobre las capacidades que tiene la IA:

Respecto al uso de medios sintéticos durante procesos electorales, aunque se habló todo el tiempo de volumen, se plantearon dudas sobre la efectividad en generar narrativas convincentes gracias a la sistematicidad en la generación de contenido, o volumen de este.

Y respecto a los procesos electorales, que no son los únicos momentos donde está presente este problema, pero sí son críticos, se habló de la necesidad de apoyar y fortalecer a las autoridades electorales, que muchas veces son también objetivo de campañas de desinformación. Una herramienta útil puede ser la legislación especial para estos contextos, y exigir a las campañas registros sobre gasto en propaganda digital, marketing, influencers, y otros elementos que también hacen parte de la desinformación.

Por último se mencionó la ausencia de temas de género dentro de las posibles soluciones. Si bien la violencia de género fue un tema central durante todo el encuentro como riesgo, es necesario reconocer que tecnologías como la IA no impactan de la misma manera a

diferentes grupos, ni las herramientas serán adoptadas de la misma manera. Es necesario priorizar a los grupos vulnerables en torno a la IA. Y también pensar cómo pueden estas herramientas servir para la investigación en otros temas, por ejemplo la trata de personas, que en el Perú es el segundo negocio más lucrativo y ocurre principalmente en la selva fronteriza con Colombia. La IA podría ayudar también a combatir ese problema.