Prácticas Responsables para Medios Sintéticos de la PAI

Un marco para la acción colectiva

NOTA
Esta traduccion en español incluye términos técnicos que posiblemente no tengan equivalentes directos en esos idiomas. En casos de ambigüedad, consulte con personas con conocimiento técnico o revise la versión del Marco en inglés.

27 DE FEBRERO DE 2023

PARTNERSHIP ON AI

Las Prácticas Responsables para Medios Sintéticos del Consorcio en IA (el nombre de la iniciativa es *Partnership on AI* y *PAI* por sus siglas también en inglés) son un conjunto de recomendaciones para apoyar el desarrollo responsable y la implementación de medios sintéticos.

Estas prácticas son el resultado de la retroalimentación de más de 100 actorxs interesadxs a nivel mundial. Se basa en el trabajo realizado por la *PAI* en los últimos cuatro años con representantes de la industria, la sociedad civil, los medios de comunicación, el periodismo y del mundo académico.

Con este marco, buscamos:

1. Avanzar en la comprensión sobre cómo aprovechar los beneficios de los medios sintéticos de manera responsable, construyendo consenso y comunidad en torno a las mejores prácticas para actorxs clave de la industria, de los medios de comunicación, el periodismo, del mundo académico y la sociedad civil.
2. Ofrecer orientación para actorxs emergentes y actorxs más grandes en el campo de los medios sintéticos.
3. Alinear normas/prácticas para reducir la redundancia y avanzar en la práctica responsable de forma amplia dentro de la industria y la sociedad, evitando una carrera cuesta abajo.
4. Asegurar que exista un documento y una comunidad asociada que sean útiles y puedan adaptarse a los desarrollos en un espacio incipiente que cambia rápidamente.
5. Servir como complemento de otros estándares y esfuerzos de políticas sobre los medios sintéticos, incluyendo a nivel internacional.

Gobernanza y Participación

Las audiencias objetivo son quienes construyen tecnología y herramientas de medios sintéticos, o quienes crean, comparten, y publican medios sintéticos.

Muchxs de estxs actorxs clave lanzarán formalmente las Prácticas Responsables para Medios Sintéticos del Consorcio en IA, uniéndose formalmente a este esfuerzo. Estas organizaciones:

1. Participarán en la comunidad de práctica de PAI.
2. Contribuirán anualmente con un ejemplo de caso o análisis que explore el marco en la práctica de una tecnología o un producto.

PAI no auditará o certificará organizaciones. Este marco incluye prácticas sugeridas que han sido desarrolladas para proporcionar orientación.

Las Prácticas Responsables para Medios Sintéticos del Consorcio en IA son un documento vivo. Mientras que está fundamentado en normas y prácticas existentes, evolucionará para reflejar nuevos desarrollos de tecnología, casos prácticos y actorxs clave. Los medios sintéticos responsables, el desarrollo, la creación y la distribución de infraestructura son áreas emergentes con cambios rápidos, que requieren flexibilidad y calibración en el tiempo. PAI planea realizar una revisión anual del marco, así como habilitar un activador de revisión en cualquier momento según lo solicite el Comité Directivo de Integridad de Inteligencia Artificial (IA) y Medios.

El Enfoque del Marco

Los medios sintéticos presentan oportunidades significativas para el uso responsable, incluyendo para propósitos creativos. Sin embargo, también pueden causar daño. A medida que la tecnología de medios sintéticos se vuelve más accesible y sofisticada, su impacto potencial también aumenta. Esto aplica tanto las posibilidades positivas como las negativas, de las cuales apenas comenzamos a explorar ejemplos en este marco. El marco se enfoca en cómo abordar mejor los riesgos que pueden presentar los medios sintéticos al tiempo que

garantiza que sus beneficios puedan materializarse de manera responsable.

Además, si bien las implicaciones éticas de los medios sintéticos son amplias e involucran elementos como los derechos de autor, el futuro del trabajo e incluso el significado del arte, el objetivo de este documento es dirigirse a un conjunto inicial de grupos de actorxs clave identificadxs por la comunidad de Integridad en los Medios e IA (*PAI AI and Media Integrity community*) que pueden desempeñar un rol significativo en: (a) reducir los daños potenciales asociados con los abusos de los medios sintéticos y promover usos responsables, (b) incrementar la transparencia y (c) permitir que las audiencias identifiquen y respondan mejor a los medios sintéticos.

Para más información sobre la creación, los objetivos y el desarrollo continuado de las Prácticas Responsables para Medios Sintéticos del Consorcio en IA, consulta estas preguntas frecuentes.

Prácticas Responsables para Medios Sintéticos del Consorcio en IA

Quienes construyen tecnología e infraestructura para medios sintéticos, crean medios sintéticos, y distribuyen o publican medios sintéticos buscarán promover un comportamiento ético y responsable.

Aquí, los medios sintéticos, también son denominados medios generativos, se definen como contenido visual, auditivo o multimodal que ha sido modificado o generado (comúnmente a través de inteligencia artificial). Sus resultados con frecuencia son muy realistas, no son identificables como sintéticos por la persona promedio, y pueden simular artefactos, personas o acontecimientos. Consulta el Apéndice A para más información sobre el alcance del marco.

La PAI ofrece **recomendaciones para diferentes categorías de actorxs interesadxs** sobre sus roles en el desarrollo, la creación, y la distribución de los medios sintéticos. **Estas categorías no son mutuamente excluyentes.** Unx actorx determinadx podría encajar en varias categorías, como es el caso de las plataformas de redes sociales. Estas categorías incluyen:

- Quienes construyen tecnología e infraestructura para medios sintéticos.
- Quienes crean medios sintéticos.
- Quienes distribuyen y publican medios sintéticos

SECCIÓN 1

Prácticas para habilitar el uso ético y responsable de los medios sintéticos

1. Colaborar para avanzar en las iniciativas de investigación, soluciones técnicas, alfabetización de medios, y en las propuestas de políticas que ayuden a contrarrestar los usos nocivos de los medios sintéticos. Además, hacemos notar que los medios sintéticos pueden utilizarse de forma responsable o pueden ser aprovechados para causar daño.

   Las categorías de uso responsable pueden incluir, entre otras, las siguientes:

   - Entretenimiento
   - Arte
   - Sátira
   - Educación
   - Investigación

2. Conducir investigaciones y compartir las mejores prácticas para seguir desarrollando categorías futuras de usos responsables y perjudiciales de los medios sintéticos.

   Estos usos a menudo implican áreas grises, y técnicas para navegar por estas áreas grises que son descritas en las secciones siguientes.

3. Cuando las técnicas descritas abajo se utilicen para crear y/o distribuir medios sintéticos con el objetivo de causar daños (ver ejemplos de daños en el Apéndice B), es necesario aplicar estrategias de mitigación razonables, coherentes con los métodos descritos en las Secciones 2, 3 y 4.

   Las siguientes técnicas pueden utilizarse responsablemente o para causar daño:

- Representar a cualquier persona o empresa, organización de medios de comunicación, organismo gubernamental o entidad.
- Creación de 'personajes falsos' realistas.
- Representar a una persona concreta que actuó, se comportó o realizó declaraciones de una manera en la que la persona real no lo hizo.
- Representar acontecimientos o interacciones que no se reprodujeron.
- Inserción de artefactos generados sintéticamente o eliminación de artefactos auténticos de los contenidos auténticos.
- Generación de escenas o paisajes sonoros totalmente sintéticos.

Para ver ejemplos de cómo estas técnicas pueden utilizarse para causar daños y conocer una lista explícita y no exhaustiva de impactos perjudiciales, consulta el Apéndice B.

SECCIÓN 2
Prácticas para **Creadorxs de Tecnología e Infraestructura**

Entre quienes construyen y proveen tecnología e infraestructura para los medios sintéticos se pueden incluir a: creadorxs de herramientas B2B (*Business-to-Business*, es decir de empresa a empresa) y B2C (*Business-to-Consumer*, es decir de empresa a consumidor); desarrolladorxs de código abierto; investigadorxs académicxs; *startups* de medios sintéticos, incluyendo quienes proporcionan la infraestructura para que las personas aficionadas creen medios sintéticos; plataformas de medios sociales; y tiendas de aplicaciones.

4. **Sé transparente con las personas usuarias sobre** las capacidades, funcionalidades, y limitaciones de las herramientas y las tecnologías, así como de los riesgos potenciales de los medios sintéticos.

5. La transparencia puede ser **directa y/o indirecta** dependiendo del caso práctico y el contexto:

   - La transparencia directa está orientada a la persona espectadora o a la oyente e incluye, pero no está limitada, a etiquetas de contenido, notas contextuales, marcas de agua y aviso de transparencia en fotogramas únicos de video.
   - La transparencia indirecta está incorporada e incluye, pero no está limitada a, la aplicación de la procedencia criptográfica de los resultados sintéticos (tal como el estándar C2PA), la aplicación de elementos trazables a los datos de entrenamiento y a los resultados, metadatos de archivos multimedia sintéticos, composición de píxeles multimedia sintéticos y declaraciones de transparencia con fotogramas únicos en videos.

6. **Cuando se desarrolle código y conjuntos de datos** (*data sets*), **modelos de entrenamiento, y se utilice *software*** para la producción de medios sintéticos, es necesario realizar todos los esfuerzos para aplicar elementos de transparencia indirecta (esteganográficos, de procedencia de medios, o de otro tipo) dentro de los activos respectivos y etapas de la producción de medios sintéticos.

   Intente divulgar de forma que se mitiguen las especulaciones sobre el contenido, los esfuerzos hacia buscar la resiliencia ante la manipulación y falsificación, se apliquen con precisión y, además, cuando sea necesario, se comunique la incertidumbre sin fomentar la especulación. (Nota: la habilidad de añadir transparencia duradera a los medios sintéticos es un reto abierto sobre el que se sigue investigando).

7. **Apoyar investigaciones adicionales para moldear futuras iniciativas de intercambio de datos** y determinar qué tipos de datos sería más apropiado y beneficioso recopilar y reportar, al tiempo que se equilibran consideraciones, tales como, la transparencia y la preservación de la privacidad.

8. Tomar los pasos para **investigar, desarrollar e implementar** tecnologías que:

    - Sean lo más detectables posible desde el punto de vista forense para la manipulación, sin ahogar la innovación en el fotorrealismo.
    - Conservar la transparencia duradera de la síntesis, como las marcas de agua o la procedencia vinculada criptográficamente que sean reconocibles, preserven la privacidad, y se pongan a disposición de la comunidad en general y se ofrezcan en código abierto.

9. **Proporcionar una política de forma pública y accesible** que describa el uso ético de tus tecnologías, así como de las restricciones de uso que se espera que cumplan las personas usuarias y que lxs proveedores hagan cumplir.

SECCIÓN 3
Prácticas para **Creadorxs**

Lxs creadorxs de medios sintéticos pueden ser desde productorxs a gran escala (como lxs productorxs de contenidos B2B) hasta productorxs a menor escala (como aficionadxs, artistas, *influencers* e integrantes de la sociedad civil, incluyendo activistas y satíricxs). Quienes se encargan y dirigen creativamente medios sintéticos también pueden entrar en esta categoría. Debido a la naturaleza cada vez más democratizada de las herramientas de creación de contenidos, cualquiera puede crear y tener la oportunidad de que sus contenidos lleguen a una audiencia amplia.

10. Sé transparente con lxs consumidorxs de contenidos sobre:

    - Cómo obtuviste el consentimiento informado de la(s) persona(s) que aparece en una pieza de contenido manipulado, adecuado al producto y al contexto, excepto cuando se utilice con fines artísticos, satíricos o expresivos razonables.
    - Cómo piensas sobre el uso ético de la tecnología y las restricciones de uso (por ejemplo, a través de una política pública y accesible, en tu sitio web o en una publicación sobre tu trabajo) y consulta estas directrices antes de crear medios sintéticos.
    - Las capacidades, limitaciones y riesgos potenciales de los contenidos sintéticos.

11. Divulgar cuando los medios que creaste o introdujiste incluyan elementos sintéticos, especialmente cuando el desconocimiento de la síntesis cambie la forma de percibir el contenido. Aprovecha todas las herramientas de transparencia que te proporcionen quienes crean tecnología e infraestructura para los medios sintéticos.

    La transparencia puede ser directa y/o indirecta dependiendo del caso práctico y el contexto:

    - La divulgación directa está orientada a la persona espectadora o a la oyente e incluye, pero no está limitada, a etiquetas de contenido, notas contextuales, marcas de agua y declaraciones de transparencia con fotogramas únicos en videos.
    - La transparencia indirecta está incorporada e incluye, pero no está limitada a, la aplicación de la procedencia criptográfica a los resultados sintéticos (tal como el estándar abierto C2PA), la aplicación trazable a los datos de entrenamiento y a resultados, metadatos de archivos multimedia sintéticos, composición de píxeles multimedia sintéticos y declaraciones de divulgación de fotogramas sencillos en los videos.

    Intente divulgar de forma que se mitiguen las especulaciones sobre el contenido, los esfuerzos hacia buscar la resiliencia ante la manipulación y falsificación, se apliquen con precisión y, además, cuando sea necesario, se comunique la incertidumbre sin fomentar la especulación.

SECCIÓN 4

Prácticas para **Distribuidorxs y Publicadorxs**

Quienes distribuyen medios sintéticos incluyen tanto a instituciones con un proceso activo de toma de decisiones editoriales que, en su mayoría, albergan contenidos de primera mano y pueden distribuir medios sintéticos creados editorialmente y/o informar sobre medios sintéticos creados por tercerxs (por ejemplo, instituciones de medios de comunicación, incluyendo a radiodifusoras). También incluyen plataformas en línea que tienen una presentación pasiva de medios sintéticos y albergan contenidos generados por usuarixs o por tercerxs (por ejemplo, plataformas de redes sociales).

Para canales de distribución activos y pasivos

12. **Divulgar** cuando con certeza detectes contenidos sintéticos generados o utilizados por tercerxs.

    La transparencia puede ser **directa y/o indirecta** dependiendo del caso práctico y el contexto:

    - La **tranparencia directa** está orientada a la persona espectadora o a la oyente e incluye, pero no está limitada, a etiquetas de contenido, notas contextuales, marcas de agua y declaraciones de transparencia con fotogramas únicos en videos.
    - La **transparencia indirecta** está incorporada e incluye, pero no está limitada a, la aplicación de la procedencia criptográfica de los resultados sintéticos (tal como el estándar abierto C2PA), la aplicación de elementos trazables a los datos de entrenamiento y a los resultados, metadatos de archivos multimedia sintéticos, composición de píxeles multimedia sintéticos y declaraciones de transparencia con fotogramas únicos en videos.

    Intente divulgar de forma que se mitiguen las especulaciones sobre el contenido, los esfuerzos hacia buscar la resiliencia ante la manipulación y falsificación, se apliquen con precisión y, además, cuando sea necesario, se comunique la incertidumbre sin fomentar la especulación.

13. **Proporcionar una política** de forma pública y accesible que describa el enfoque de la organización con respecto a los medios sintéticos, que cumplirán y tratarán de hacer cumplir.

Para canales de distribución activos

Canales (como las instituciones de medios de comunicación) que alojan principalmente contenidos propios y pueden distribuir medios sintéticos creados con fines editoriales y/o informar sobre medios sintéticos creados por otrxs.

14. **Realizar ajustes inmediatos** cuando te dés cuenta de que, sin saberlo, distribuiste y/o presentaste contenidos sintéticos nocivos.
15. **Evitar distribuir contenidos de medios sintéticos no atribuidos** o informar sobre medios sintéticos perjudiciales creados por otrxs sin un etiquetado y un contexto claros que garanticen que ninguna persona espectadora o lectora pueda pensar que no son sintéticos.
16. **Trabajar por** una infraestructura organizativa de procedencia de los contenidos, tanto para los medios no sintéticos como para los sintéticos, respetando la privacidad (por ejemplo, a través de un estándar abierto C2PA).
17. **Garantizar que el consentimiento transparente e informado** fue otorgado por la persona creadora y la(s) personas(s) representadas en el contenido sintético que será compartido y distribuido, aún si ya recibiste el consentimiento para la creación de contenido.

Para los canales de distribución pasivos

Canales (como plataformas) que alojan principalmente contenidos de tercerxs.

18. **Identificar** los medios sintéticos nocivos que se distribuyen en las plataformas aplicando métodos técnicos razonables, informes de usuarixs, y medidas del staff o equipo de trabajo para hacerlo.
19. **Realizar ajustes inmediatos** mediante el etiquetado, el descenso de rango (downranking), la eliminación u otras intervenciones como las descritas aquí, cuando se sepa que se distribuyen medios sintéticos novios en la plataforma.
20. **Comunicar** y **educar** claramente a lxs usuarixs de la plataforma sobre los medios sintéticos y acerca de qué tipos de contenido sintético está permitido crear y/o compartir en la plataforma.

Apéndices

APÉNDICE A
Prácticas Responsables de la PAI en el ámbito de los Medios Sintéticos

Aunque este marco se centra en formas altamente realistas de medios sintéticos, reconoce que el umbral de lo que se considera altamente realista puede variar en función de la alfabetización de medios de la audiencia y de los contextos globales. También reconocemos que los medios sintéticos que no son altamente realistas pueden causar daños, como en el contexto del abuso de imagen íntima.

Este marco fue creado con un enfoque centrado en los medios sintéticos audiovisuales, también conocidos como medios generativos más que en el texto sintético, que presenta otras ventajas y riesgos. No obstante, puede ser útil como guía para la creación y distribución de texto sintético.

Adicionalmente, este marco sólo cubre los medios generativos, no la categoría más amplia de IA generativa en su conjunto. Reconocemos que a veces estos términos se consideran intercambiables.

APÉNDICE B
Daños Potenciales de los Medios Sintéticos

Lista de daños potenciales de los medios sintéticos que hay que buscar mitigar:

- Hacerse pasar por una persona para obtener información o privilegios no autorizados.
- Realizar llamadas telefónicas no solicitadas, comunicaciones masivas, publicaciones o mensajes que engañen o acosen.
- Cometer fraude para obtener beneficios económicos
- Desinformación sobre una persona, grupo u organización.
- Explotación o manipulación de menores.
- *Bullying* y acoso.
- Espionaje.
- Manipulación de procesos democráticos y políticos, incluyendo engañar a la persona votante para que vote a favor o encontra de una persona candidata, dañando la reputación de la persona candidata mediante declaraciones o actos falsos, influyendo en el resultado de las elecciones mediante el engaño o la supresión de votantes.
- Manipulación del mercado y sabotaje empresarial.
- Creación o incitación al odio, la discriminación, la difamación, el terrorismo o actos de violencia
- Difamación y sabotaje de la reputación.
- Contenido íntimo o sexual no consentido.
- Extorsión y chantaje.
- Creación de nuevas identidades y cuentas a escala para representar a personas concretas con el fin de “fabricar la opinión pública”.