Cuestionario Lectura

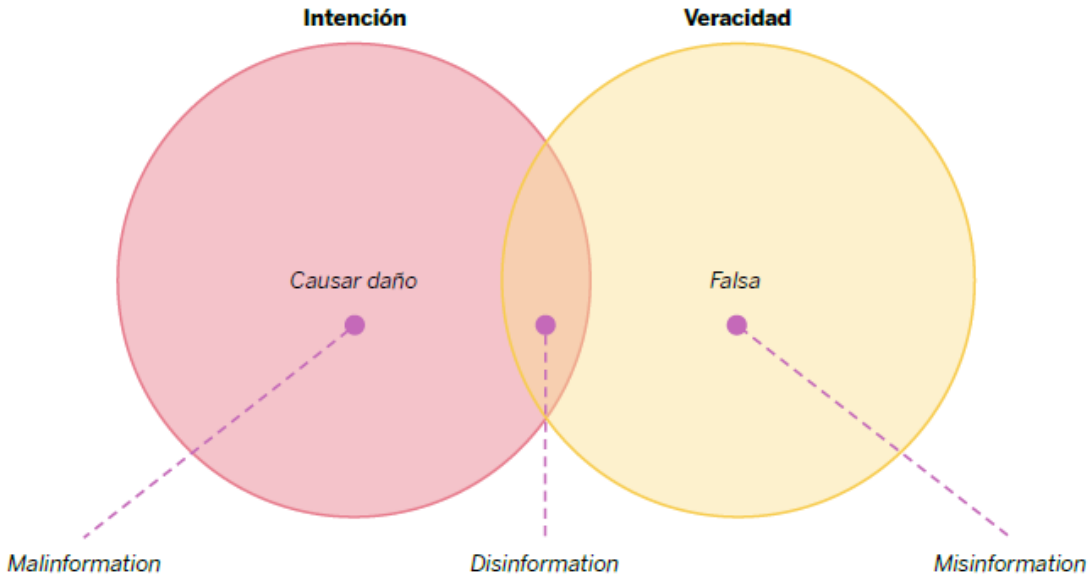
Luchando contra la Desinformación mediante la inteligencia artificial
Fundación BBVA
2024

Asignatura: Habilidades prácticas en el ciberespacio

Estudiante: CC Diego Edison Cabuya Padilla

1. ¿Cuál es la diferencia fundamental entre 'misinformation' y 'disinformation'?

La diferencia fundamental entre *misinformation* y *disinformation* radica en la **intencionalidad**.
Según el informe, ***misinformation*** se refiere a información falsa difundida sin intención de causar daño; es decir, el emisor presume que lo que comunica es cierto, aunque no lo sea. En cambio, ***disinformation*** es información falsa creada y difundida deliberadamente con el objetivo de manipular, engañar o generar perjuicio.
Ambos conceptos se diferencian también de ***malinformation***, que hace referencia a información verídica que se comparte de forma maliciosa, fuera de contexto o con sesgo, con el fin de causar daño.



2. **¿Qué tendencia preocupante se observa en España según el Reuters Institute Digital News Report 2023?**

Según el Reuters Institute Digital News Report 2023, en España se observa una tendencia preocupante de **disminución del interés y confianza en las noticias**.

Entre 2015 y 2023, el porcentaje de personas con un interés alto o muy alto por las noticias descendió del 85 % al 51 %, lo que representa una caída de 34 puntos porcentuales, una de las más pronunciadas a nivel internacional.

Paralelamente, la desconfianza hacia los medios de comunicación alcanzó un máximo histórico del 40 %, especialmente entre los menores de 45 años. Esta combinación de desinterés informativo y creciente desconfianza ha contribuido a generar un entorno fértil para la proliferación masiva de desinformación en redes sociales y plataformas digitales.

3. **¿Cómo se comparan, según Vosoughi, Roy y Aral (2018), la velocidad y facilidad de difusión de noticias falsas frente a las verdaderas?**

Según el estudio de Vosoughi, Roy y Aral (2018), **las noticias falsas se difunden en redes sociales más rápidamente y con mayor alcance que las verdaderas**.

En concreto, el 1 % de las noticias falsas más viralizadas alcanzaron entre 1.000 y 100.000 personas, mientras que el 1 % de las verdaderas rara vez superaron las 1.000 personas. Estas diferencias significativas en los patrones de propagación permiten rastrear las "trazas" que dejan al circular por la red, lo cual ha sido aprovechado para desarrollar filtros automáticos de detección temprana de desinformación en redes sociales

4. **¿Qué ventaja clave ofrecen las redes latentes de difusión sobre los modelos epidemiológicos?**

La ventaja clave que ofrecen las redes latentes de difusión frente a los modelos epidemiológicos es **su capacidad para identificar de forma precisa quién propaga la desinformación y cómo lo hace**. Mientras que los modelos epidemiológicos operan de forma anónima y solo permiten detectar flujos anómalos sin conocer los agentes responsables ni sus interacciones, las redes latentes reconstruyen las dinámicas de influencia individual entre usuarios.

Estas redes permiten:

- Modelar la evolución temporal de la información dentro de una red social.
- Determinar qué usuarios influyen en otros, y con qué intensidad.

- Clasificar a los nodos según su rol (controladores, difusores o receptores).
- Aplicar análisis para identificar comunidades, nodos y patrones de propagación.
- Granularidad analítica que permite actuar con precisión frente a la desinformación.

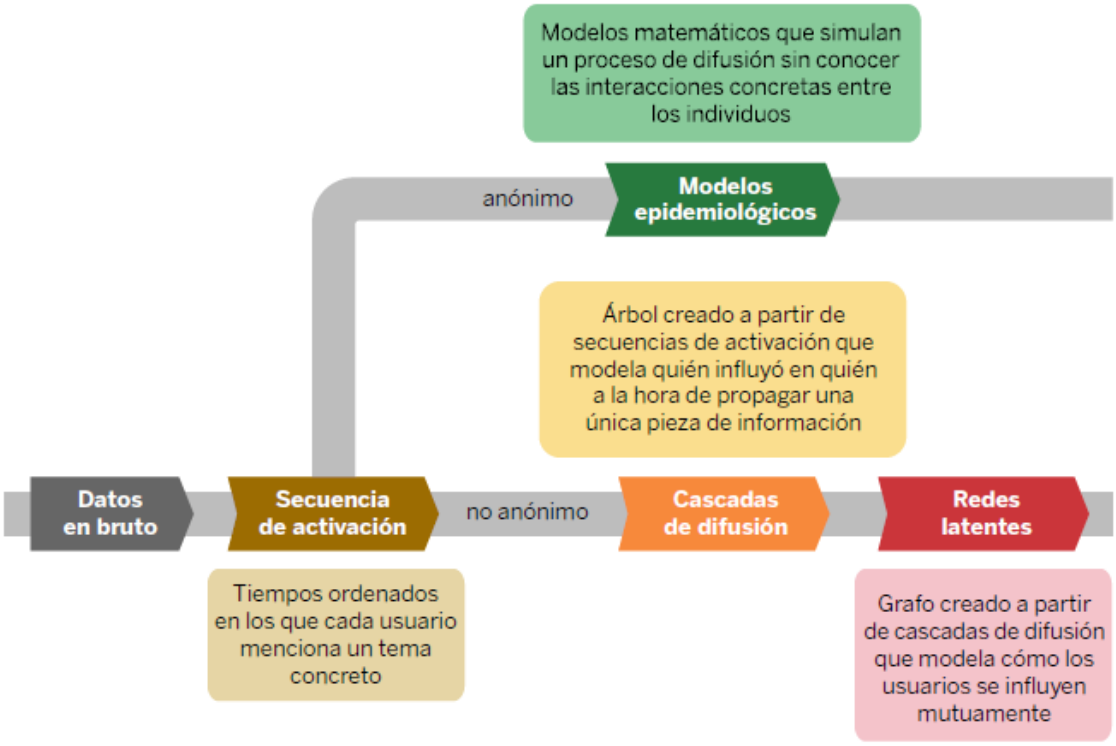


Figura 1.1. Distintos enfoques para analizar y modelizar la propagación de la desinformación en redes

5. ¿Qué son los 'grandes modelos de lenguaje' y cuál es su principal riesgo en el contexto de la desinformación?

Los grandes modelos de lenguaje (*large language models*) **son sistemas de inteligencia artificial basados en la arquitectura Transformer, capaces de generar texto en lenguaje natural con alta coherencia y calidad**. Han revolucionado el procesamiento del lenguaje natural al permitir capturar relaciones complejas entre palabras sin depender del orden secuencial. Entrenados con enormes volúmenes de datos textuales, estos modelos pueden redactar artículos, responder preguntas o simular conversaciones de forma fluida.

En el contexto de la desinformación, **su principal riesgo radica en que facilitan de forma masiva, rápida y económica la creación de contenido falso.** Entre sus amenazas más relevantes destacan:

- Capacidad para generar textos engañosos y persuasivos en segundos, incluso a partir de información completamente falsa.

- Ausencia de comprensión de la veracidad, lo que los hace susceptibles de producir afirmaciones incorrectas sin distinguir entre lo verdadero y lo falso.
- Vulnerabilidad ante manipulaciones externas, ya que pueden ser inducidos a generar desinformación incluso cuando inicialmente muestran resistencia.
- Reducción drástica de barreras técnicas, permitiendo que actores sin experiencia en redacción o conocimiento del idioma produzcan desinformación creíble.
- Utilización en *bots* avanzados, que simulan interacciones humanas de forma convincente, dificultando su detección.

6. ¿Cómo facilita la accesibilidad de los modelos de IA la generación de desinformación?

La accesibilidad de los modelos de inteligencia artificial facilita significativamente la generación de desinformación, ya que **reduce los conocimientos técnicos, recursos económicos y barreras operativas necesarias para producir contenido falso convincente**. Esta facilidad permite democratizar la capacidad de generar desinformación, convirtiéndola en una amenaza global fácil de activar a gran escala.

La facilidad descrita se expresa en varios aspectos:

- **Baja barrera de entrada:** modelos potentes como los LLM pueden ejecutarse en computadores domésticos sin necesidad de hardware especializado ni conocimientos avanzados en programación o IA, lo que permite que cualquier actor, incluso sin dominar el idioma o técnicas de redacción, pueda generar desinformación con facilidad.
- **Velocidad y calidad de generación:** los sistemas permiten crear, en minutos, artículos falsos de apariencia profesional, con una narrativa atractiva y estructurada, listos para su difusión masiva.
- **Imprecisión irrelevante para el desinformador:** la veracidad no es una prioridad para quien busca manipular. Lo importante es la capacidad de generar relatos creíbles y atractivos, y los modelos actuales permiten precisamente eso, incluso si contienen errores fácticos.
- ***Bots* automatizados más sofisticados:** gracias a estos modelos, es posible generar *bots* que simulan interacciones humanas creíbles, publicando mensajes adaptativos y personalizables. Estos *bots* son más difíciles de detectar y se utilizan eficazmente en campañas de propaganda o manipulación informativa.
- **Tendencia hacia modelos más ligeros y móviles:** la evolución hacia modelos más compactos y funcionales desde dispositivos móviles incrementa aún más su disponibilidad, amplificando el riesgo de desinformación por parte de usuarios sin formación técnica.

7. **¿Qué son las 'cajas negras' en el contexto de la IA explicativa y cuál es el desafío asociado?**

En el contexto de la inteligencia artificial explicativa, las llamadas “cajas negras” **son modelos altamente complejos —como los basados en redes neuronales profundas— cuyo funcionamiento interno es opaco y difícil de interpretar, incluso para sus propios desarrolladores**. Estos sistemas operan con millones de parámetros interconectados, lo que dificulta rastrear cómo una entrada específica se transforma en una decisión o resultado.

El principal desafío asociado es la falta de transparencia y explicabilidad, especialmente en aplicaciones críticas como la detección de desinformación, donde es fundamental que los usuarios comprendan por qué una afirmación es clasificada como verdadera o falsa. Esta opacidad afecta la confianza, la trazabilidad y la posibilidad de identificar errores o sesgos en el sistema.

La IA explicativa busca justamente enfrentar este reto, desarrollando métodos que permitan abrir esas “cajas negras” para ofrecer justificaciones comprensibles, auditables y verificables. Lograrlo es un objetivo complejo y aún en construcción, pero clave para asegurar el uso ético, responsable y confiable de la inteligencia artificial en ámbitos sensibles como la lucha contra la desinformación.

8. **¿Qué implicaciones tiene el concepto de 'Inteligencia Artificial General (AGI)' para la lucha contra la desinformación?**

La Inteligencia Artificial General (AGI) representa un **escenario profundamente transformador —y ambivalente**— para la lucha contra la desinformación. Definida como un sistema capaz de razonar, aprender y actuar con una flexibilidad similar a la inteligencia humana, **la AGI podría potenciar tanto los mecanismos de defensa como los riesgos asociados a la manipulación informativa**. En términos positivos, puede utilizarse para detectar, contextualizar, rastrear y contrarrestar campañas de desinformación con mayor eficacia, gracias a su capacidad para comprender lenguaje, imágenes y audio de forma integrada y adaptativa.

No obstante, desde una perspectiva de riesgo, las implicaciones son críticas y abarcan:

- Generación de desinformación altamente sofisticada y personalizada, con contenidos que superan el umbral de credibilidad humana y se adaptan a perfiles psicológicos o socioculturales específicos.
- Automatización avanzada de campañas de manipulación, mediante decisiones autónomas sobre cuándo, cómo y a quién desinformar, optimizando la estrategia de difusión en función del entorno político, digital o emocional.

- Aumento de la asimetría en la dinámica de verificación, ya que el desinformador —apoyado por AGI— tendría ventaja temporal y tecnológica, dificultando una respuesta efectiva por parte de los sistemas de fact-checking.
- Escalabilidad global del problema, al permitir la ejecución simultánea de múltiples campañas desde dispositivos accesibles, sin requerir infraestructura compleja ni intervención humana directa.

En suma**, la AGI podría convertirse en una herramienta poderosa tanto para combatir como para intensificar la desinformación, dependiendo del marco ético, regulatorio y técnico que rija su implementación**.

9. ¿Qué normativas europeas importantes se mencionan en relación con la IA y la privacidad?

La lectura menciona tres normativas clave en el contexto europeo que abordan aspectos fundamentales relacionados con la inteligencia artificial y la protección de la privacidad:

- **Reglamento General de Protección de Datos (RGPD o GDPR):** adoptado el 27 de abril de 2016, reconoce que "toda persona tiene derecho a la protección de los datos de carácter personal que le conciernan", un principio consagrado también en el Tratado de Funcionamiento de la Unión Europea y en la Carta de los Derechos Fundamentales. El RGPD establece principios rectores para el tratamiento de datos personales, derechos del interesado y obligaciones para los responsables. Se enfatiza que cualquier sistema de IA diseñado para combatir la desinformación debe respetar este marco legal, asegurando, por ejemplo, que se utilicen únicamente datos de perfiles públicos o previamente autorizados.

- **Libro Blanco sobre la Inteligencia Artificial:** enmarcado en la Estrategia Digital de la Unión Europea, este documento establece las bases para el desarrollo de una IA centrada en el ser humano y que respete los derechos fundamentales, incluida la protección de datos. Sirve como marco general para las iniciativas regulatorias y de desarrollo tecnológico de la IA en Europa.

- **Ley de Inteligencia Artificial (AI Act):** su objetivo es abordar no solo las preocupaciones en torno a la privacidad, sino también proteger otros derechos fundamentales como la libertad de expresión, el pensamiento y la no discriminación en el contexto del uso de sistemas de IA.

10. ¿Cómo garantiza FacTeR-Check el cumplimiento de la normativa de protección de datos al analizar redes sociales?

FacTeR-Check garantiza el cumplimiento del Reglamento General de Protección de Datos (RGPD) y otras directrices europeas mediante un **enfoque de *privacy-by-design*, incorporando la protección de datos desde su diseño y funcionamiento**. Este enfoque se materializa en tres pilares fundamentales:

- **Diseño centrado en la privacidad desde el inicio**: el sistema asegura que la protección de datos no sea un añadido posterior, sino un componente estructural del sistema.
- **Uso de fuentes verificadas y perfiles públicos**: en el análisis de redes sociales, FacTeR-Check solo emplea datos provenientes de perfiles públicos. En caso de requerir información de perfiles privados, se exige el consentimiento explícito del usuario, en línea con los principios de licitud, lealtad y transparencia de datos.
- **Adopción de principios de IA responsable**: el sistema considera dimensiones como la explicabilidad, la seguridad, la supervisión humana y la transparencia. Esto es especialmente relevante en contextos donde las decisiones automatizadas pueden afectar derechos fundamentales, como el perfilado y la atribución de credibilidad informativa.

11. Analice las diferentes formas en que la Inteligencia Artificial puede ser utilizada tanto para generar como para combatir la desinformación, basándose en los ejemplos y conceptos presentados en el texto.

La Inteligencia Artificial (IA) cumple una **función dual y ambivalente** en el ecosistema informativo contemporáneo: puede ser utilizada para amplificar y sofisticar la desinformación, pero también para detectar, rastrear y contrarrestarla de manera eficaz.

a) Generación de desinformación asistida por IA

- **Texto:** los LLMs permiten crear textos coherentes, persuasivos y difíciles de distinguir de los producidos por humanos. Su accesibilidad pública, velocidad y capacidad de personalización permiten generar campañas masivas de desinformación a bajo costo, incluso por usuarios sin conocimientos técnicos.
- **Imágenes y vídeo:** tecnologías como deepfakes, inpainting, style transfer y modelos de difusión posibilitan la creación de contenido visual falsificado, que refuerza narrativas engañosas. Estas imágenes, generadas con herramientas como DALL·E o Stable Diffusion, pueden simular rostros, alterar contextos o fabricar escenas completas. Aunque el vídeo presenta más dificultades técnicas, también puede ser manipulado para modificar expresiones o sincronizar voces falsas con movimiento facial.

- **Audio:** técnicas como la clonación de voz, la manipulación auditiva y la síntesis de habla permiten crear grabaciones falsas con alta fidelidad. La combinación de texto, voz e imagen genera contenido multimedia integral que puede suplantar a figuras públicas y manipular emocionalmente al receptor.

b) Combatir a la desinformación con IA: caso FacTeR-Check

Como contrapeso, la IA también se posiciona como una herramienta crítica para la verificación automatizada. Un ejemplo es FacTeR-Check, una plataforma basada en IA que emplea múltiples módulos para detectar y contrarrestar narrativas falsas, que se caracteriza por utilizar:

- **Similitud semántica:** identifica correspondencias entre frases sospechosas y una base de datos de hechos verificados, empleando *ensembles* de modelos Transformer multilingües.
- **Inferencia del lenguaje natural (NLI):** determina si una afirmación implica, contradice o es neutral respecto a un hecho verificado, evaluando relaciones lógicas y semánticas entre enunciados.
- **Monitorización de redes sociales:** integra procesamiento de contenido en plataformas como Twitter, extracción de entidades (NER), análisis de palabras clave (KeyBERT), y clasificaciones automatizadas del discurso como apoyo, negación o neutralidad respecto a los bulos.
- **Visualización de la diseminación:** representa la difusión de narrativas en grafos, permitiendo identificar usuarios influyentes, puntos de quiebre informativos y evolución del bulo frente a los desmentidos.

c) Desafíos y perspectivas

- **Carrera asimétrica:** el desinformador siempre tiene la ventaja temporal, pues actúa antes de que se verifique el contenido. La mejora simultánea de los modelos ofensivos y defensivos perpetúa un ciclo de adaptación constante.
- **Explicabilidad (XAI):** para confiar en los sistemas de detección, es esencial que sus decisiones sean comprensibles. Sin embargo, la mayoría de los modelos actuales son "cajas negras", lo que dificulta la transparencia.
- **Privacidad y regulación:** FacTeR-Check aplica principios de privacy-by-design, utilizando datos de perfiles públicos o con consentimiento explícito, conforme al RGPD y al marco normativo europeo en desarrollo.

En conclusión, la IA es una herramienta de doble filo. Si bien potencia como nunca la creación de desinformación, también proporciona los medios más avanzados para combatirla. El equilibrio entre ambos usos dependerá del desarrollo responsable de la tecnología, de marcos regulatorios adecuados y de la capacidad social para adaptarse.

12. Discuta el papel de la Inteligencia Artificial Explicativa (XAI) en la mejora de la confianza pública en los sistemas de detección de desinformación y en la educación de los usuarios. ¿Cuáles son los principales obstáculos para su desarrollo?

La Inteligencia Artificial Explicativa (XAI) desempeña un papel crucial en la legitimación del uso de sistemas automatizados de detección de desinformación, al **proporcionar explicaciones claras y comprensibles sobre las decisiones tomadas por modelos complejos**. En un contexto en el que la desinformación puede tener implicaciones sociales y políticas significativas, la confianza del público depende de la transparencia y trazabilidad de las herramientas empleadas.

a) Contribuciones principales de la XAI.

- **Confianza y transparencia:** la XAI permite que los usuarios comprendan por qué un contenido ha sido clasificado como falso, verdadero o ambiguo. Esta transparencia es esencial para reforzar la legitimidad del sistema, especialmente cuando se procesan afirmaciones delicadas o polémicas.
- **Alfabetización y autonomía digital:** ofrecer explicaciones comprensibles promueve la alfabetización mediática, ayudando a los usuarios a identificar patrones engañosos por sí mismos. Según el texto, la XAI podría incluso destacar fragmentos contrastables de los mensajes, favoreciendo la formación crítica del usuario.
- **Apoyo a expertos y científicos:** más allá del usuario común, la XAI también facilita que desarrolladores y analistas comprendan el comportamiento de sus propios modelos, lo cual es clave para corregir sesgos y mejorar su rendimiento.

b) Obstáculos clave para su desarrollo.

- **Modelos opacos o "cajas negras":** la mayoría de los sistemas con alto rendimiento, son inherentemente difíciles de interpretar. Operan con millones de parámetros que no permiten una comprensión intuitiva de sus decisiones.
- **Trade-off entre explicabilidad y rendimiento:** los modelos más explicables suelen ser menos potentes, mientras que los más potentes sacrifican claridad por exactitud. Encontrar un equilibrio entre ambos es un desafío técnico persistente.
- **Alfabetización digital del usuario final:** incluso cuando se generan explicaciones, su utilidad depende de la capacidad del usuario para comprenderlas. Esto plantea un reto sociotécnico adicional: adaptar el nivel de explicabilidad al perfil del receptor.

En síntesis, la XAI no solo mejora la confianza pública en los sistemas de detección de desinformación, sino que también contribuye activamente a la educación digital. Sin embargo, su desarrollo requiere resolver complejidades técnicas y humanas, desde la comprensión de modelos hasta la presentación adecuada de las explicaciones generadas.

13. Compare los modelos epidemiológicos y las redes latentes de difusión como enfoques para estudiar la propagación de la desinformación en las redes sociales. ¿Qué información específica puede obtenerse de cada tipo de modelo?

Los modelos epidemiológicos y las redes latentes de difusión representan enfoques complementarios para analizar cómo se propaga la desinformación en entornos digitales, particularmente en redes sociales.

a) Modelos epidemiológicos

Estos modelos, inspirados en la propagación de enfermedades infecciosas, simplifican la dinámica de difusión al agrupar a los usuarios en categorías como Susceptibles (S), Infectados (I) y Recuperados (R), con parámetros que definen probabilidades de transición entre estos estados (por ejemplo, β para contagio y γ para recuperación).

Aunque permiten simular el crecimiento o decrecimiento de una narrativa falsa a escala poblacional y prever la evolución de la difusión bajo diferentes escenarios, su principal limitación, cuando no son usados con otros modelos matemáticos, radica en su carácter anónimo:

- No permiten identificar quién propaga la desinformación.
- No consideran las estructuras reales de interacción entre individuos.
- No explican los mecanismos específicos que impulsan o frenan la difusión.

Sin embargo, estos modelos permiten generar información específica, como:

- Predicción agregada de la evolución temporal del fenómeno.
- Impacto potencial de cambios en parámetros (p. ej., campañas de intervención).
- Identificación de anomalías a gran escala.

b) Redes latentes de difusión

Este enfoque modela explícitamente las relaciones entre usuarios en redes sociales, partiendo de cascadas de activación (cuando un usuario publica o replica un contenido). A partir de estas observaciones, infiere redes de influencia latente, incluso cuando la estructura real de la red no es completamente observable.

A diferencia del enfoque epidemiológico, las redes latentes no son anónimas: permiten rastrear el origen, trayectoria y actores involucrados en la difusión. Adicionalmente, permiten generar el siguiente tipo de información:

- Identificación de autores, replicadores y nodos clave ("influencers").
- Reconstrucción de comunidades y patrones de difusión.

- Análisis de interacciones entre narrativas (p. ej., entre bulo y desmentido).
- Visualización temporal y topológica del fenómeno.

c) Comparación.

- Los modelos epidemiológicos ofrecen una visión agregada, útil para caracterizar la dinámica general de la difusión.
- Las redes latentes de difusión permiten una comprensión granular, enfocada en la interacción entre individuos y comunidades.
- Ambos requieren datos empíricos para su ajuste, pero las redes latentes exigen mayor precisión y volumen de información sobre los usuarios, lo cual representa un desafío en entornos donde las cuentas son volátiles y las interacciones no siempre son públicas.

En conclusión, mientras los modelos epidemiológicos permiten anticipar tendencias generales, las redes latentes son esenciales para diseñar respuestas específicas e identificar puntos de intervención efectivos. Su complementariedad resulta clave frente al crecimiento exponencial de campañas de desinformación potenciadas por la IA.

Adicionalmente, los modelos epidemiológicos complementados con otras técnicas de modelamiento y simulación como la modelación basada en agentes, permite individualizar el agente sujeto de observación y así poder realizar análisis de comportamientos individuales dentro de una población.

14. Examine la relación entre la accesibilidad de las herramientas de IA generativa y el aumento potencial de la desinformación. ¿Qué estrategias se sugieren para mitigar este riesgo?

La creciente accesibilidad de las herramientas de inteligencia artificial generativa ha **reducido drásticamente las barreras técnicas y económicas para producir desinformación de alta calidad.**

Modelos como GPT, DALL·E o Stable Diffusion, disponibles de forma pública o en código abierto, permiten a actores maliciosos generar textos, imágenes, audios o videos manipulados sin necesidad de conocimientos especializados.

Esta **democratización tecnológica** ha intensificado el volumen, la velocidad y la personalización de la desinformación, contribuyendo a la expansión de campañas de "*astroturfing", deepfakes* y *bots* sofisticados que simulan comportamiento humano.

El texto sugiere diversas estrategias de carácter técnico, educativo y regulatorio:

- **Filtros y sistemas de control integrados en los modelos**: incorporar salvaguardas para impedir la generación de contenido malicioso o engañoso desde el diseño mismo del sistema.
- **Aplicación de IA explicativa (XAI):** facilita la comprensión del funcionamiento de los modelos por parte de usuarios y verificadores, fortaleciendo la confianza en los sistemas de detección y promoviendo la alfabetización digital.
- **Limitación de acceso a modelos altamente potentes:** restringir el uso irrestricto de herramientas con alto potencial de abuso, especialmente aquellas que permiten manipulación multimedia realista.
- **Verificación semiautomática basada en hechos:** herramientas como FacTeR-Check contrastan afirmaciones con bases de datos verificadas, manteniéndose actualizadas ante la "plasticidad" de la desinformación.
- **Análisis de estilo y contexto:** más allá del contenido, se propone examinar el estilo lingüístico del mensaje y la red de conexiones del emisor, identificando patrones típicos de desinformadores reincidentes.
- **Combinación de técnicas a priori y a posteriori:**
    - *A priori:* evalúan estilo y contexto sin contrastar con bases de datos.
    - *A posteriori:* verifican el contenido comparándolo con hechos comprobados, aunque con la desventaja temporal de reaccionar después de publicada la desinformación.
- **Alfabetización mediática y campañas de sensibilización:** formar a los usuarios para identificar tácticas engañosas y desarrollar pensamiento crítico ante los contenidos digitales.
- **Marco normativo y colaboración multisectorial:** la regulación es clave para exigir transparencia, explicabilidad y uso ético. Se requiere, además, una articulación entre academia, sector privado y entes reguladores para desarrollar respuestas coordinadas.

En conclusión, la relación entre accesibilidad y riesgo de desinformación es directa: a mayor disponibilidad de herramientas generativas, mayor es la amenaza. Sin embargo, mediante una combinación de salvaguardas técnicas, enfoque educativo, verificación inteligente y regulaciones robustas, es posible mitigar este riesgo de forma proactiva.

15. Analice las consideraciones éticas y de privacidad asociadas con el uso de la Inteligencia Artificial para combatir la desinformación, haciendo referencia a las normativas europeas mencionadas e identificado si existen normativas similares en nuestro país.

El uso de sistemas de Inteligencia Artificial (IA) para combatir la desinformación plantea desafíos éticos y de privacidad significativos. Aunque estas tecnologías ofrecen herramientas poderosas para detectar y contrarrestar contenidos engañosos, su aplicación debe estar alineada con el respeto a derechos fundamentales como la privacidad, la libertad de expresión y la no discriminación.

a) Consideraciones éticas y de privacidad destacadas en el documento.

- **Protección de datos personales:** la recolección y análisis de información en redes sociales implica tratamiento de datos personales. El documento insiste en que cualquier herramienta de IA —como FacTeR-Check— debe diseñarse conforme a un enfoque privacy-by-design, aplicando técnicas como anonimización, seudonimización y análisis local. Además, se establece que los datos deben provenir de perfiles públicos o requerir consentimiento explícito si son privados.
- **Transparencia y explicabilidad:** la IA explicativa (XAI) es clave para garantizar que los usuarios comprendan por qué un sistema clasifica un contenido como desinformación. Esta comprensión es esencial para fortalecer la confianza pública y reducir la opacidad de los modelos actuales, muchos de los cuales operan como "cajas negras".
- **Sesgos y errores algorítmicos:** la posibilidad de generar errores o reproducir sesgos en los modelos puede afectar injustamente a ciertos grupos o limitar indebidamente la libertad de expresión. La XAI también es fundamental para identificar y mitigar estos riesgos.
- **Supervisión humana y derechos fundamentales**: las decisiones automatizadas deben estar sujetas a supervisión humana, especialmente cuando pueden impactar derechos fundamentales. La clasificación automatizada de contenidos debe ser proporcional y estar claramente justificada.
- **Fiabilidad de las fuentes de verificación:** las herramientas deben basarse en bases de datos validadas por entidades de fact-checking y aplicar criterios rigurosos para asegurar su objetividad.

b) Normativas europeas mencionadas en el documento.

- **Reglamento General de Protección de Datos (RGPD / GDPR):** principal referente legal en la UE sobre protección de datos personales, establece principios y derechos fundamentales para el tratamiento de datos.

- **Ley de Inteligencia Artificial (AI Act):** en desarrollo al momento de la publicación del informe, esta normativa busca regular los sistemas de IA según su nivel de riesgo, estableciendo requisitos de transparencia, seguridad y derechos del usuario.
- **Libro Blanco de la IA:** define la estrategia europea para el desarrollo confiable y seguro de la IA, con atención a la privacidad y la ética.

c) Normativas similares en Colombia.

En Colombia, si bien no existe una ley específica que regule integralmente el uso de la Inteligencia Artificial (IA), el país ha avanzado en la formulación de marcos normativos y estratégicos relevantes.

En materia de protección de datos personales, la **Ley 1581 de 2012** establece el régimen general aplicable, con la Superintendencia de Industria y Comercio (SIC) como entidad de control. Esta ley reconoce los derechos de los titulares de la información y define los principios que deben regir el tratamiento de datos personales, tales como la finalidad, libertad, veracidad, seguridad y confidencialidad.

En cuanto a la IA, el **Documento CONPES 4144 de 2023**, titulado **Política Nacional para el Desarrollo y la Adopción de la Inteligencia Artificial,** representa el esfuerzo más reciente y estructurado del Estado colombiano para promover el uso ético, responsable e inclusivo de estas tecnologías. El CONPES 4144 incluye, entre otros aspectos:

- Principios éticos para el desarrollo y uso de la IA, como la transparencia, la explicabilidad, la equidad, la privacidad y la no discriminación.
- Lineamientos para la protección de datos en sistemas de IA, en coherencia con la Ley 1581 y estándares internacionales como el RGPD europeo.
- Orientaciones para la gobernanza algorítmica, enfocadas en la rendición de cuentas, la trazabilidad de decisiones automatizadas y la supervisión humana.

Aunque este documento no tiene fuerza de ley ni establece sanciones, sí define una hoja de ruta interinstitucional que busca alinear el desarrollo de la IA en Colombia con estándares globales de derechos fundamentales y protección de datos personales.

En conclusión, el desarrollo y uso de IA contra la desinformación exige un equilibrio delicado entre eficacia tecnológica y respeto por los derechos fundamentales. La normativa europea establece un marco robusto en este sentido, mientras que Colombia ha avanzado en políticas, pero aún requiere marcos regulatorios específicos para IA y desinformación. Adaptar las buenas prácticas europeas, como el enfoque privacy-by-design, será esencial para garantizar una implementación ética y efectiva en contextos locales.