MY CHAVARRO GUTIERREZ LUIS
AULA R

1. ¿Cuál es la diferencia fundamental, según el texto, entre "misinformation" y "disinformation"?

La diferencia fundamental es la intencionalidad. "Disinformation" es información falsa creada y difundida deliberadamente para causar daño o manipular, mientras que "misinformation" es información falsa que se comparte sin intención de engañar, es decir, el emisor cree que es verdadera.

2. Según el Reuters Institute Digital News Report 2023, ¿qué tendencia preocupante se observa en España con respecto al interés por las noticias?

Se observa una fuerte caída en el interés por las noticias: del 85% de personas con interés alto o muy alto en 2015, se pasó al 51% en 2023. Además, la desconfianza en los medios ha alcanzado un récord del 40%, especialmente entre los menores de 45 años

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

Las noticias falsas se difunden más rápido y llegan a más personas que las verdaderas. El 1% de las noticias falsas más difundidas alcanzó entre 1.000 y 100.000 personas, mientras que el 1% de las verdaderas rara vez superó las 1.000 personas.

4. ¿Qué ventaja clave ofrecen las redes latentes de difusión sobre los modelos epidemiológicos para el estudio de la desinformación?

Las redes latentes de difusión permiten identificar quién propaga la información y cómo lo hace, es decir, permiten conocer la estructura de la red y los actores clave, mientras que los modelos epidemiológicos solo muestran patrones anónimos y no identifican a los individuos responsables.

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

Son modelos de inteligencia artificial capaces de generar texto de alta calidad de manera automática (como GPT-3/4). Su principal riesgo es que pueden crear grandes volúmenes de desinformación de forma rápida, barata y en múltiples idiomas, sin distinguir entre verdad y mentira.

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

La accesibilidad (fácil uso, disponibilidad pública, bajo requerimiento técnico) permite que cualquier persona, incluso sin conocimientos avanzados, pueda generar y difundir desinformación usando IA, reduciendo la barrera de entrada para campañas maliciosas.

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

Las "cajas negras" son sistemas de IA cuyos procesos internos no son comprensibles para los usuarios. El desafío es que resulta difícil explicar cómo llegan a sus conclusiones, lo que dificulta la confianza y la transparencia en la detección de desinformación.

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

La AGI podría suponer un salto en la capacidad de generar y detectar desinformación, pero también plantea riesgos éticos y de control, ya que una IA con capacidades generales podría ser utilizada tanto para crear desinformación sofisticada como para combatirla de manera más efectiva.

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

Se mencionan el Reglamento General de Protección de Datos (GDPR) y la propuesta de Reglamento de Inteligencia Artificial de la Unión Europea, que buscan garantizar la privacidad y el uso ético de la IA.

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

FacTeR-Check utiliza únicamente datos públicos y anonimizados, y se asegura de no almacenar información personal identificable, cumpliendo así con la normativa de protección de datos como el GDPR.

PREGUNTAS DE FORMATO ENSAYO

1. 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 IA puede ser utilizada para generar desinformación mediante grandes modelos de lenguaje que crean textos falsos, bots que automatizan la difusión, y modelos generativos de imágenes, audio y video (deepfakes). Estas herramientas permiten crear contenido falso de alta calidad y difícil de distinguir del real, facilitando campañas de manipulación a gran escala. Por otro lado, la IA también es clave para combatir la desinformación: sistemas

como FacTeR-Check emplean modelos de procesamiento de lenguaje natural para verificar hechos, analizar similitud semántica y monitorizar redes sociales, ayudando a identificar y frenar la propagación de bulos. Así, la IA es un arma de doble filo: puede amplificar el problema, pero también es esencial para la solución.

La IA actúa como una "espada de doble filo" (Lazer et al., 2018).

1. Generación:

    - Los grandes modelos de lenguaje (LLM) pueden crear textos persuasivos en segundos, imitando estilos periodísticos o testimonios personales (Bender et al., 2021).
    - Los modelos generativos adversariales (GAN) producen imágenes y videos hiper-realistas ("deepfakes") que erosionan la evidencia visual (Chesney & Citron, 2019).
    - Los sistemas de "bot-farms" con IA gestionan miles de cuentas automatizadas que amplifican narrativas falsas (Chakraborty et al., 2021).

2. Mitigación:

    - Detección algorítmica de patrones lingüísticos y difusivos. Ej.: modelos híbridos que combinan BERT con redes GCN para detectar noticias falsas con >90 % de precisión (Shu et al., 2017).
    - Verificación automática de hechos (fact-checking) mediante recuperación de evidencias y razonamiento semántico (Thorne et al., 2018).
    - Modelado de redes sociales para identificar "superspreaders" y cortar cascadas tempranas (Vosoughi, Roy & Aral, 2018).

La clave es un enfoque sociotécnico: combinar IA con periodistas, educadores y reguladores para contrarrestar la misma potencia generativa que habilita la desinformación (Floridi & Chiriatti, 2020).

2. 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 XAI busca que los sistemas de IA sean comprensibles y transparentes, lo que es fundamental para que el público confíe en las herramientas de detección de desinformación. Si los usuarios entienden cómo y por qué una IA clasifica una noticia como falsa, estarán más dispuestos a aceptar sus resultados y aprenderán a identificar patrones de desinformación. Sin embargo, el principal obstáculo es la complejidad de los modelos actuales (cajas negras), que dificulta explicar sus decisiones de manera sencilla. Además,

existe el reto de equilibrar la precisión con la interpretabilidad y de adaptar las explicaciones a diferentes niveles de conocimiento del usuario.

La Inteligencia Artificial Explicable (XAI) busca "hacer visible lo invisible" (Doshi-Velez & Kim, 2017).
• Confianza: explicaciones comprensibles aumentan la aceptación de las alertas de contenido falso, reduciendo el "descreimiento reactivo" (Barredo Arrieta et al., 2020).
• Educación: mostrar al usuario por qué un post se marca como engañoso refuerza su alfabetización mediática (Guess et al., 2020).

Obstáculos principales:

1. Complejidad de los LLM, cuyas representaciones de billones de parámetros son opacas ("cajas negras") (Bender et al., 2021).
2. Trade-off entre explicabilidad y precisión: los modelos lineales son más comprensibles pero menos eficaces (Doshi-Velez & Kim, 2017).
3. Riesgo de revelar información sensible del modelo o de los datos al ofrecer explicaciones detalladas (Goodman & Flaxman, 2017).

Las líneas de trabajo incluyen técnicas post-hoc (SHAP, LIME) y arquitecturas intrínsecamente interpretable basadas en reglas o grafos semánticos (Barredo Arrieta et al., 2020).

3. 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 permiten analizar la propagación de la desinformación de forma anónima, identificando patrones generales y simulando escenarios de contagio, pero sin identificar a los individuos responsables. Son útiles para detectar flujos anómalos y predecir la evolución global de la desinformación. En cambio, las redes latentes de difusión permiten identificar quiénes son los actores clave, cómo se relacionan y quién influye en quién, proporcionando información detallada sobre la estructura de la red y los nodos más influyentes. Así, los modelos epidemiológicos ofrecen una visión macro, mientras que las redes latentes permiten intervenciones más precisas y dirigidas.

Aspecto	Modelos epidemiológicos	Redes latentes de difusión
Metáfora	Infectados, susceptibles, recuperados (SIR)	Vértices ocultos y pesos que se infieren
Ventajas	Simplicidad, cálculo de R_0, umbral crítico (Pastor-Satorras & Vespignani, 2001)	Captura caminos reales de propagación sin suponer estructura previa (Yang & Leskovec, 2010)
Limitaciones	Necesitan parámetros globales; poco detalle individual	Requieren datos granulares y cómputo intensivo

Aspecto	Modelos epidemiológicos	Redes latentes de difusión
Información obtenible	Velocidad promedio de contagio, tamaño final de la cascada	Quién influye a quién, rutas específicas, detección de "brotes" locales

En desinformación, los modelos epidemiológicos ayudan a estimar la rapidez de la ola y la efectividad de "cuarentenas" (p. ej. demorar retuits). Las redes latentes descubren nodos clave —influencers o comunidades— para intervenciones quirúrgicas (Vosoughi et al., 2018).

4. 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 accesibilidad de las herramientas de IA generativa (fácil uso, disponibilidad pública) facilita que más personas puedan crear y difundir desinformación, aumentando el riesgo de campañas maliciosas. Para mitigar este riesgo, el texto sugiere desarrollar métodos de detección más ágiles y escalables, establecer controles y restricciones en el acceso a modelos avanzados, y fomentar la educación y la alfabetización mediática para que los usuarios sean más críticos y capaces de identificar contenido falso.

La reducción de barreras (modelos open-source, interfaces "one-click") democratiza la creatividad, pero también el engaño (Weidinger et al., 2021). Evidencia empírica muestra picos de desinformación coincidentes con lanzamientos de herramientas como Stable Diffusion 2.0 (Kreps & McCain, 2024).

Estrategias de mitigación:

- Watermarking y firmas criptográficas incrustadas en outputs generados (OpenAI, 2023).
- Políticas de plataforma: IA para IA ("AI-in-the-loop") que detecta estilo sintético antes de la publicación (Schuster et al., 2023).
- Alfabetización mediática orientada a IA: currículos que enseñan reconocimiento de patrones de generación automática (Guess et al., 2020).
- Gobernanza multilateral: la propuesta Ley de IA de la UE impone requisitos de transparencia y gestión de riesgos a proveedores de sistemas fundacionales (European Commission, 2024).

5. 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 identificando si existen normativas en nuestro país similares.

El uso de IA para combatir la desinformación plantea retos éticos y de privacidad, como el riesgo de vigilancia masiva, el uso indebido de datos personales y la posible censura. Las normativas europeas, como el GDPR y la propuesta de Reglamento de IA, buscan garantizar la protección de datos y el uso ético de la tecnología. En España, la Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales (LOPDGDD) complementa el GDPR y establece requisitos similares. Es fundamental que las herramientas de IA respeten la privacidad, utilicen solo datos públicos o anonimizados y sean transparentes en su funcionamiento para mantener la confianza social.

El reglamento europeo establece un "triángulo normativo":
• GDPR (2016/679) → principios de minimización y finalidad de datos.
• DSA (2022/2065) → obligaciones de diligencia y transparencia de plataformas.
• AI Act (2024) → clasificación de riesgo y supervisión de sistemas IA.

Proyectos como FacTeR-Check aplican "privacy-by-design": anonimización de identificadores y almacenamiento descentralizado conforme al GDPR y a la LOPDGDD española 3/2018 (European Data Protection Board, 2022).

Riesgos éticos: vigilancia masiva, sesgos algorítmicos, chilling effect sobre la libertad de expresión (Floridi & Chiriatti, 2020). Para mitigarlos se proponen auditorías externas independientes y mecanismos de derecho a explicación (Goodman & Flaxman, 2017). España carece aún de una ley específica de IA, pero el anteproyecto de "Ley de IA y Derechos Digitales" (BOE, 2024) pretende alinear el marco nacional con la AI Act.

REFERENCIAS

Barredo Arrieta, A., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of ACM FAccT 2021*, 610-623. https://doi.org/10.1145/3442188.3445922

Chakraborty, A. J., et al. (2021). A survey on misinformation detection techniques. *Social Network Analysis and Mining, 11*(98). https://doi.org/10.1007/s13278-021-00766-4

Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review, 107*, 1753-1820.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*:1702.08608.

European Commission. (2024). *Regulation (EU) 2024/... on artificial intelligence (AI Act)*.

European Parliament & Council. (2016). *General Data Protection Regulation (EU) 2016/679*.

European Parliament & Council. (2022). *Digital Services Act (EU) 2022/2065*.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines, 30*, 681-694. https://doi.org/10.1007/s11023-020-09548-1

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine, 38*(3), 50-57. https://doi.org/10.1609/aimag.v38i3.2741

Guess, A. M., et al. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *PNAS, 117*(27), 15536-15545. https://doi.org/10.1073/pnas.1920498117

Kreps, S., & McCain, R. (2024). From novelty to normal: Public reactions to generative AI releases. *Journal of Online Trust, 2*(1), 45-62.

Lazer, D. M. J., et al. (2018). The science of fake news. *Science, 359*(6380), 1094-1096. https://doi.org/10.1126/science.aao2998

Pastor-Satorras, R., & Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Physical Review Letters, 86*(14), 3200-3203. https://doi.org/10.1103/PhysRevLett.86.3200

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *SIGKDD Explorations, 19*(1), 22-36. https://doi.org/10.1145/3137597.3137600

Thorne, J., et al. (2018). FEVER: a large-scale dataset for fact extraction and verification. *NAACL-HLT 2018*, 809-819.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146-1151. https://doi.org/10.1126/science.aap9559

Weidinger, L., et al. (2021). Ethical and social risks of harm from language models. *arXiv*:2112.04359.

Yang, J., & Leskovec, J. (2010). Modeling information diffusion in implicit networks. *IEEE ICDM 2010*, 599-608. https://doi.org/10.1109/ICDM.2010.70

España. Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales. *BOE* núm. 294, 06-12-2018.