Escuela Superior de Guerra “General Rafael Reyes Prieto”

Cuestionario

Autores:

MY. Lopez Zapata Dolman Andrés

Electiva Habilidades prácticas en el ciberespacio

Docente: Jaider Ospina Navas

Curso de Estado Mayor

Bogotá D.C 25 de abril 2025

Cuestionario

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

Misinformation se refiere a información falsa o incorrecta que se comparte sin la intención de engañar, es decir, por error o desconocimiento.

Disinformation implica información falsa que se crea y difunde deliberadamente para manipular, engañar o causar daño.

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

Solo el 46% de los españoles muestra un interés alto o muy alto por las noticias**, lo que supone una caída significativa en comparación con años anteriores.

Esta cifra sitúa a España por debajo de la media global del estudio (48%) y refleja un creciente distanciamiento de parte del público hacia el consumo de información periodística.

Causas principales mencionadas en el informe:

Fatiga informativa (especialmente por temas repetitivos o negativos, como crisis políticas, económicas o conflictos).

Desconfianza en los medios (percepción de sesgo o sensacionalismo).

Cambio en los hábitos de consumo (mayor uso de redes sociales como fuente de información fragmentada).

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?

   Velocidad:

Las falsedades alcanzan a 1,500 personas 6 veces más rápido que las noticias verdaderas.

Los contenidos falsos tienen un 70% más de probabilidades de ser retuiteados.

Alcance: Las noticias falsas llegan a cascadas de difusión más grandes (hasta 100 veces más extensas en algunos casos). La profundidad (niveles de reenvíos) es también mayor en las falsedades.

Facilidad de viralización:

Las falsedades políticas se viralizan más que las de otros temas (como terrorismo, desastres o ciencia).

Los bots no fueron la causa principal: la difusión se debe sobre todo al comportamiento humano (sesgos como la novedad, el impacto emocional o la confirmación de prejuicios).

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

   Modelos epidemiológicos (SIR, SEIR, etc.):

Enfoque: Tratan la desinformación como un "virus" que se contagia de persona a persona mediante interacciones directas.

Limitaciones:

Asumen una transmisión homogénea (ignoran diferencias en credibilidad, afinidad ideológica o estructura de redes sociales).

No consideran variables latentes como sesgos cognitivos, emociones o algoritmos de plataformas.

Redes latentes de difusión:

Ventaja principal: Modelan factores no observables que influyen en la propagación, como:

Homofilia (tendencia a interactuar con usuarios similares).

Estructuras de comunidades en redes sociales (ej.: burbujas ideológicas).

Influencia de algoritmos (priorización de contenido polarizante).

Contexto sociocultural (ej.: desconfianza en instituciones).

Métodos: Usan técnicas como embedding de nodos (ej.: Node2Vec) o modelos basados en aprendizaje profundo para inferir patrones ocultos en los datos de difusión.

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

Riesgo principal en la desinformación:

Su capacidad para generar contenido persuasivo, coherente y a escala —incluyendo texto, imágenes o vídeos (*deepfakes*)— sin mecanismos inherentes para garantizar veracidad. Esto amplifica:

Producción masiva de desinformación:

Crean noticias falsas, discursos manipulativos o respuestas plausibles pero erróneas en segundos.

Ejemplo: Generar miles de artículos o comentarios con narrativas conspirativas.

Personalización de la manipulación:

Adaptan el mensaje al perfil del usuario (basado en datos de comportamiento).

Ejemplo: Un bot que diseña argumentos antivacunas según la ideología de quien interactúa.

Dificultad para detectarlo:

El contenido es lingüísticamente sofisticado, sin errores típicos de bots simples.

Herramientas de detección actuales (como *fact-checkers*) no escalan al ritmo de generación.

Erosión de la confianza:Al saturar el ecosistema informativo, socavan la capacidad de distinguir lo real de lo falso (*efecto "liar's dividend"*).

Agravantes:

Sesgos en los datos de entrenamiento: Pueden reforzar estereotipos o narrativas falsas preexistentes.

Accesibilidad: Herramientas como ChatGPT permiten que actores malintencionados sin expertise técnico creen desinformación

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

Bajo costo y facilidad de usoHerramientas gratuitas o baratas: Plataformas como ChatGPT, Gemini o modelos open-source (Llama, Mistral) permiten generar texto convincente sin necesidad de programación.

Interfaces intuitivas: Cualquier persona puede crear contenido falso con solo ingresar un *prompt* (instrucción). Ejemplo: *"Escribe un artículo científico falso que vincule vacunas con autismo"*.

Escalabilidad masiva

Producción en minutos: Un solo usuario puede generar miles de tweets, artículos o comentarios falsos con scripts simples.

Multiformato: No solo texto, también imágenes (DALL-E), audio (ElevenLabs) o vídeos (*deepfakes*).

Personalización de la desinformación

Adaptación a audiencias: Los LLMs ajustan el mensaje según el idioma, cultura o sesgos del objetivo.

Ejemplo: Crear versiones de una conspiración para grupos políticos opuestos.

Sofisticación persuasiva

Lenguaje verosímil: Los modelos imitan estilos periodísticos, académicos o "oficiales", dificultando la detección.

Apariencia de autoridad: Citando fuentes inventadas pero plausibles (estudios falsos, expertos inexistentes).

Automatización de tácticas de influencia

Granjas de bots: IA puede gestionar cuentas falsas que interactúan entre sí para simular tendencias (*astroturfing*).

Ataques coordinados: Ejemplo: Inundar redes con hashtags o memes manipulativos durante elecciones.

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

Características de las "cajas negras" en IA:

Falta de transparencia:
El modelo genera resultados (ej.: una predicción o un texto) sin revelar cómo llegó a ellos.

Ejemplo: ChatGPT no explica por qué eligió ciertas palabras o fuentes en una respuesta.

Complejidad intrínseca:
Millones de parámetros interactúan de formas no lineales, imposibles de rastrear manualmente.

No determinismo:Mismos inputs pueden producir outputs ligeramente distintos, dificultando la replicabilidad.

Desafíos principales asociados:Responsabilidad y ética:

Si un sistema de IA comete un error (ej.: difamar a alguien o negar un crédito injustamente), es difícil asignar responsabilidades o corregir sesgos.

Sesgos ocultos:
Los modelos pueden aprender y replicar prejuicios presentes en sus datos de entrenamiento (racismo, sexismo, etc.), pero al ser opacos, estos sesgos son difíciles de detectar.Desconfianza pública y adopción:

Sectores como la medicina, la justicia o el periodismo exigen explicaciones para confiar en las decisiones de la IA.

Regulación y cumplimiento legal:
Leyes como el GDPR (Art. 22) en la UE exigen "derecho a explicación" en decisiones automatizadas, algo imposible con cajas negras puras.

Desinformación y manipulación:
Si no se entiende cómo un LLM genera contenido falso, es más difícil prevenir su propagación.

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

Detección y verificación hipereficiente:Un AGI podría analizar masas de datos en tiempo real (noticias, redes sociales, discursos) para identificar patrones de desinformación con precisión superior a los humanos.

Ejemplo: Cruzar fuentes, contextos e historiales de credibilidad en segundos.

Contextualización profunda: Entendería el subtexto cultural, histórico o político detrás de un mensaje, detectando manipulaciones sutiles (ej.: ironía, dogwhistles políticos).

Generación de contra-narrativas: Crearía contenido educativo o correctivo personalizado para audiencias específicas, adaptándose a idiomas, niveles educativos o sesgos cognitivos.

Autocorrección ética: En teoría, un AGI alineado con valores humanos podría rechazar colaborar en la creación de desinformación, incluso si se le ordena.

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

   Aplicación: Desde 2018.

Relevancia para IA:

Exige transparencia en decisiones automatizadas (Art. 22): Los usuarios tienen derecho a una explicación si una IA afecta sus derechos (ej.: denegar un crédito).

Requiere consentimiento explícito para procesar datos personales usados en entrenamiento de modelos.

Limita el perfilado automatizado sin supervisión humana.

Ley de Inteligencia Artificial (AI Act)

Estado: Aprobada en marzo de 2024 (primera ley integral de IA a nivel mundial).

Enfoque clave:

Clasificación de riesgos: Prohíbe IA de "riesgo inaceptable" (ej.: manipulación conductual o *social scoring*).

Obligaciones para IA de alto riesgo (ej.: sistemas de reclutamiento o vigilancia):

Evaluaciones de cumplimiento, transparencia y supervisión humana.Requisitos para modelos fundacionales (como GPT): Documentar datos de entrenamiento y cumplir con derechos de autor.Ley de Servicios Digitales (DSA)Aplicación: Desde 2024.

Impacto en IA:

Plataformas que usen IA (ej.: redes sociales) deben:

Mitigar riesgos sistémicos (como desinformación generada por IA).

Permitir auditorías externas de algoritmos recomendadores.

Ley de Mercados Digitales (DMA)Objetivo: Regular a los "guardianes de acceso" (*gatekeepers*, como Google o Meta).

Conexión con IA:Exige interoperabilidad y evita preferencias algorítmicas injustas que afecten competencia.

Directiva de Ciberseguridad (NIS2)Relación con IA:

Obliga a empresas críticas (incluidas las de IA) a reportar brechas de seguridad quepuedan afectar privacidad.

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

    En la era de la desinformación masiva, herramientas como FacTeR-Check (un sistema hipotético de verificación de hechos basado en IA) deben equilibrar su misión de detectar falsedades con el estricto cumplimiento de las normativas de protección de datos, especialmente en Europa, donde el Reglamento General de Protección de Datos (GDPR) y la Ley de Inteligencia Artificial (AI Act) establecen requisitos rigurosos. Este ensayo explora los mecanismos clave que permitirían a FacTeR-Check operar de manera ética y legal al analizar contenido en redes sociales.

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.

    IA como Arma de Desinformación

Generación Masiva de Contenido Falso

Grandes Modelos de Lenguaje (LLMs): Herramientas como GPT-4 o Gemini pueden crear noticias falsas persuasivas en segundos.

Ejemplo: Artículos pseudocientíficos que vinculan vacunas con efectos adversos, escritos con jerga médica realista.

Multimodalidad: IA como DALL-E o Sora generan imágenes, audios y vídeos falsos (*deepfakes*).*Ejemplo*: El *deepfake* de Volodímir Zelensky pidiendo la rendición de Ucrania (2022). Personalización de la Manipulación

Algoritmos de microtargeting: Analizan datos de usuarios para adaptar mensajes falsos a sus sesgos.

Ejemplo: Campañas políticas que difunden narrativas diferentes a progresistas y conservadores.

Automatización de Redes de Difusión

Granjas de bots: Cuentas falsas gestionadas por IA simulan engagement orgánico.

Ejemplo: Cuentas automatizadas que amplifican teorías conspirativas en Twitter.

Sofisticación de la Evasión

IA contra IA: Modelos como *WormGPT* (variante maliciosa de ChatGPT) eluden filtros de contenido.*Ejemplo*: Correos de phishing con redacción impecable, evitando detección.

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?

Transparencia en la Detección de Desinformación

Explicaciones comprensibles: La XAI permite que los sistemas de fact-checking muestren cómo y por qué clasifican un contenido como falso (ej.: señalar incoherencias en imágenes o contradicciones en fuentes).

Ejemplo: Un modelo podría destacar que un vídeo es un *deepfake* al mostrar anomalías en el parpadeo o la sincronía labial.

Auditoría algorítmica: Facilita la supervisión humana al revelar los criterios usados por la IA, evitando decisiones arbitrarias.

Educación Mediática

Herramientas pedagógicas: Sistemas basados en XAI pueden enseñar a los usuarios a identificar señales de desinformación.

Ejemplo: Un chatbot que explica por qué un titular es sensacionalista o cómo verificar una imagen con herramientas como *Reverse Image Search*.Visualización de sesgos: Muestra cómo algoritmos de redes sociales priorizan contenido polarizante, fomentando pensamiento crítico.

Legitimidad InstitucionalCumplimiento normativo: La XAI ayuda a satisfacer requisitos como el "derecho a explicación" del GDPR o la transparencia exigida por la AI Act de la UE.

Ejemplo: Plataformas como Meta o Twitter podrían justificar por qué eliminan cierto contenido "fake".

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?

Características PrincipalesInspiración: Basados en modelos de propagación de enfermedades (ej.: SIR – Susceptible, Infectado, Recuperado).Supuestos:

La desinformación se "contagia" de persona a persona mediante interacciones directas.La población se divide en categorías fijas (*susceptibles*, *infectados*, *recuperados*).

Enfoque matemático: Ecuaciones diferenciales o simulaciones de agentes.

Información Específica que Proporcionan

Tasas de propagación:

Velocidad a la que la desinformación se expande (*R0*, tasa básica de reproducción).

Ejemplo: Un estudio podría calcular que un bulo político tiene un *R0 = 2.5*, indicando que cada usuario infectado lo comparte con 2.5 personas en promedio.

Umbrales de viralización:

Punto crítico donde la desinformación se vuelve epidémica.

Impacto de intervenciones:Efectividad de estrategias como *"inmunización"* (ej.: fact-checking temprano) para reducir la difusión.

Limitaciones

Simplificación excesiva: Ignora factores sociales como homofilia o influencia de algoritmos.

No captura contextos complejos: Asume que todos los usuarios son igualmente susceptibles.

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?**

Producción Masiva y AutomatizadaEjemplo: Un solo usuario puede generar miles de artículos falsos, imágenes o *deepfakes* en minutos usando IA.

Caso real: En las elecciones argentinas de 2023, se usó IA para crear audios falsos de candidatos.

Sofisticación Persuasiva

Lenguaje natural convincente: Los LLMs (como GPT-4) redactan discursos falsos con estilo periodístico o académico, dificultando la detección.

Multimodalidad: Herramientas como Sora (IA de vídeo de OpenAI) generan *deepfakes* casi indistinguibles de contenido real.

Personalización de la Manipulación

Microtargeting algorítmico: La IA adapta mensajes falsos a los sesgos ideológicos de audiencias específicas.

Ejemplo: Una misma campaña de desinformación puede generar versiones diferentes para progresistas y conservadores.

Escalabilidad de las Granjas de BotAutomatización de cuentas falsas: IA gestiona redes de bots que simulan engagement orgánico en redes sociales.

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 identificadas si existen normativas en nuestro país similares.

Consideraciones Éticas Clave

1. Privacidad y Protección de Datos

- Problema: Los sistemas de IA para detectar desinformación suelen analizar grandes volúmenes de datos de redes sociales, lo que puede incluir información personal (ej.: mensajes privados, ubicaciones, preferencias políticas).
- Normativa relevante:
  - GDPR (UE, 2018): Exige minimización de datos, consentimiento explícito y anonimización.
    - *Ejemplo*: Un sistema como FacTeR-Check debe evitar almacenar datos personales innecesarios.
  - AI Act (UE, 2024): Clasifica los sistemas de moderación de contenido como de riesgo limitado, pero exige transparencia en el uso de datos.

2. Sesgos Algorítmicos y Discriminación

- Problema: Los modelos de IA pueden replicar sesgos presentes en sus datos de entrenamiento, llevando a falsos positivos (ej.: etiquetar como "falsa" una noticia legítima de grupos minoritarios).
- Normativa relevante:
  - AI Act (Art. 10): Obliga a evaluar y mitigar riesgos de discriminación en sistemas de alto riesgo.

3. Libertad de Expresión y Censura

- Problema: La IA puede sobrecensurar contenido legítimo al confundirlo con desinformación.
- Normativa relevante:
  - DSA (Ley de Servicios Digitales, UE, 2024): Exige que las plataformas justifiquen la eliminación de contenido y permitan apelaciones humanas.

4. Transparencia y Explicabilidad

- Problema: Los usuarios afectados por decisiones automatizadas tienen derecho a saber por qué su contenido fue marcado como falso.
- Normativa relevante:
  - GDPR (Art. 22): Derecho a no ser sujeto de decisiones basadas únicamente en IA sin supervisión humana.
  - AI Act (Transparency Risk Mitigation): Los modelos generativos (como GPT-4) deben revelar que su contenido es IA-generated.