OdiseIA. Group 1.3

Assesing the impact of AI externalities on society and vulnerable groups

Content

Nov, 30 2024

Apunte metodológico

Para tratar los efectos de las externalidades de la inteligencia artificial sobre los grupos vulnerables a estas se ha partido de una categorización inicial de ambos conceptos: externalidades y grupos vulnerables. En tanto estos han sido definidos por otro de los grupos de trabajo, las externalidades tienen que aplicarse a cada uno de ellos con sus peculiaridades.

Se considera, no obstante, que resulta más apropiado hablar de grupos o individuos en situación de vulnerabilidad en lugar de utilizar el término, más conciso, de grupos vulnerables.

Aunque a continuación se emplearán de forma indistinta ambos términos, debe entenderse siempre que se pretende discutir la combinación de las características de un individuo o grupo en un entorno concreto, y no de forma genérica.

La misma persona o grupo puede o no ser vulnerable dependiendo del contexto en el que se desarrolle su actividad. Esta vulnerabilidad también se verá afectada por el rol que desempeñe en un momento concreto. Se podría, incluso, describir una gradación de las vulnerabilidades en función de estos y otros parámetros que queda fuera del alcance y pretensiones del presente trabajo.

Las manifestaciones concretas de las externalidades en cada uno de los grupos o individuos son muy diversas gracias a la capacidad algorítmica de personalización de los contenidos. Cada grupo o individuo en situación de vulnerabilidad reacciona de un modo distinto.

Incluso la diferenciación entre grupo e individuo puede resultar relevante en la definición de las vulnerabilidades a que se enfrenta cada uno de ellos. La componente gregaria de los primeros genera fortalezas, pero también sesgos, en los individuos que los componen.

El origen de las externalidades

Una primera aproximación al origen de las externalidades podría partir de la componente que se explota en cada caso. Esto es, se propone comenzar la taxonomía considerando si la vulnerabilidad se establece en el medio o en el sujeto de la relación.

En base a ese criterio, los elementos diferenciales fundamentales que determinan las externalidades pueden dividirse en:

- Externalidades de carácter técnico y tecnológico.
- Externalidades de carácter social y psicológico.

Obviamente, ambas externalidades tienen zonas de intersección o de influencia mutua.

Los sesgos cognitivos humanos, por ejemplo, determinan muchas veces las características de los datos para el entrenamiento de los algoritmos y, aunque el proceso de discriminación se lleve a cabo de un modo técnicamente impecable, el resultado queda adulterado por una inadecuada técnica de selección de los datos. En este caso concreto, la externalidad se deriva de factores psicológicos o sociológicos, pero volcados en el proceso algorítmico con carácter independiente del grupo o individuo que vaya a resultar afectado.

Externalidades de carácter técnico y tecnológico

Son las derivadas de los procesos informáticos (incluyendo el entrenamiento de los algoritmos) y cuyos efectos se circunscriben al interior del proceso algorítmico. Estas externalidades resultan, por lo tanto, relativamente independientes de las características del sujeto afectado, aunque no de sus circunstancias específicas en relación con el medio.

Externalidades de carácter social y psicológico

Son las que se producen en el proceso cognitivo humano que se deriva de los resultados proporcionados por el proceso algorítmico. Por lo tanto, se trata de externalidades que explotan vulnerabilidades intrínsecas al sujeto en cuestión, aunque siempre teniendo en cuenta el contexto en el que se encuentra.

Estudio de la revisión bibliográfica

Los estudios previos utilizados en la revisión bibliográfica llevada a cabo para la realización de este trabajo presentan un exceso de granularidad en las taxonomías -tanto la de las externalidades como la de los grupos vulnerables-derivado dos factores:

- por un lado, una aproximación bottom-up a la clasificación que pretende llegar al establecimiento de las diferentes categorías construyéndolas desde la agrupación de casos concretos, muchas veces sin ánimo de identificar factores comunes y, por otro,
- un intento de politización de los resultados centrado en resultados concretos.

El énfasis que se ha puesto en los últimos años en la identificación de factores diferenciales que particularicen las características de cada grupo o individuo vulnerable puede resultar útil en el acometimiento de medidas concretas para la inclusión o protección de estos grupos o individuos a nivel académico, pero presenta desafíos importantes a la hora de adoptar políticas prácticas. Parece más adecuada la búsqueda de comunalidades en las

características del mayor número de grupos posible que facilite el diseño de medidas mitigadoras desde una perspectiva de escalas pragmática.

Propuesta de taxonomía de las externalidades

En lo que respecta a los grupos e individuos vulnerables, el acceso a -y capacidad de uso de- las herramientas de IA determina en muy buena medida las posibilidades de obtener beneficios de su empleo o de ser víctimas de este. Hay un consenso generalizado entre los académicos respecto de que la falta de acceso a las herramientas digitales marca la diferencia fundamental entre grupos o individuos en lo que respecta a las oportunidades de realización social y profesional. Frases como “la IA no te quitará el trabajo, sino alguien que la utilice mejor que tú” enfatizan que el carácter híbrido de las actividades humanas contemporáneas requiere de la interacción con la IA.

Partiendo de la disponibilidad de dicho acceso, el proceso de interacción humano-máquina y los sesgos implícitos en estas últimas son los elementos más significativos en la generación de desigualdades.

Así, podríamos clasificar las externalidades por:

- Limitaciones en el acceso mismo a la IA.
- Limitaciones en el uso de la IA.
- Desequilibrios en la representación en la sociedad o en las bases de datos de entrenamiento de los algoritmos.
- Colectivos infrarrepresentados en los modelos de entrenamiento de las bases de datos por alguna otra razón.
- El conjunto de la sociedad humana con sus diferentes sesgos, características y limitaciones.

Limitaciones de acceso a la tecnología

Esta categoría comprende distintos colectivos:

- Las poblaciones de regiones remotas o desconectadas sin acceso a los servicios digitales, temporal o permanentemente. Por supuesto, aquí se incluyen los más de 800 millones de personas que no tienen acceso a la Internet, pero también aquellos que, por circunstancias concretas, se encuentren privados de este con carácter más o menos permanente.
- Las personas y poblaciones desplazadas que, por distintas circunstancias, ven limitada su capacidad para utilizar las redes digitales.
- Los grupos e individuos cuyo poder adquisitivo limita su acceso a la IA, aún cuando este esté disponible en la región en la que desarrollen su actividad.

Limitaciones en el uso de la tecnología

Esto es, en las habilidades digitales requeridas para el acceso o la explotación de sus posibilidades:

- Poblaciones sin formación en aspectos digitales (brecha digital) o cuya formación está desactualizada o resulta muy insuficiente.
- Población de edad avanzada carente de suficiente formación básica en los aspectos digitales y, por lo tanto, de contexto. Es decir, no toda la población de avanzada edad sería vulnerable, sino aquella que no ha tenido acceso a la digitalización en suficiente medida. Se diferencia del grupo anterior precisamente en esta falta de contexto por una tardía o insuficiente exposición al entorno digital.
- Población infantil con insuficiente contexto social para entender las implicaciones del uso de la IA. En este caso, la falta de contexto no lo es con respecto al ámbito digital, sino al social. El tratamiento de estas tres externalidades, por lo tanto, requerirá de una aproximación distinta para cubrir las carencias específicas de cada uno de ellos.

Desequilibrios en la representación en la sociedad o en las bases de datos de entrenamiento de los algoritmos.

En este caso, se pretenden cubrir los sesgos psicológicos, sociológicos o de entrenamiento algorítmico que aplica a las minorías y que, según se ha demostrado, tiende a marginarlas. En algunos casos, se puede dar lugar a intentos de homogeneización de estos grupos para conformarlos con la norma estadística o a que sean considerados distorsiones indeseables:

- Minorías étnicas, culturales, religiosas o nacionales dentro de sus respectivos países. Aquí cabe incluir las manifestaciones de estas culturas, como los lenguajes minoritarios, insuficientemente representados y la rentabilidad de cuya inclusión es muy limitada. Un escenario que estudiar será la identificación de consecuencias derivadas de la exclusión de lenguajes minoritarios de los modelos de lenguaje (LLM).
- Minorías diversas o diversificadas, grupos o individuos no suficientemente representados. Incluye grupos de personas con capacidades diferentes o con discapacidades.
- Grupos socialmente subordinados en determinadas colectividades por razón de género.

Colectivos infrarrepresentados en los modelos de entrenamiento de las bases de datos por alguna otra razón.

La falta de representación algorítmica *per se* invisibiliza (o reduce la visibilidad) de personas o colectivos, independientemente del aspecto concreto que permanezca oculto o mitigado.

El conjunto de la sociedad humana con sus diferentes sesgos, características y limitaciones

- Sesgos cognitivos humanos en general y los específicos de grupos concretos, sean estos culturales, filosóficos o ambientales.
- Influencia de la industria del sector y de un entorno centrado en el desarrollo de la IA que favorece sus procesos frente a los de la cognición humana. Esto es, sesgos de producción que privilegian los formatos accesibles a las máquinas sobre los más comprensibles para el ser humano.
- Ritmo de batalla. Establecimiento de la necesidad de adaptación del humano a los ritmos decisorios de los algoritmos mecánicos. Se favorece de este modo el pensamiento instintivo y la disminución de capacidades analíticas humanas incrementando el desequilibrio de la interacción humano-máquina en favor de esta última.

Otras aproximaciones metodológicas

Otros estudios proponen aproximaciones similares o complementarias con la que se propone en este trabajo. Vemos, en ambos casos, el exceso de granularidad y la posible aparición de sesgos derivados del interés en poner el foco en determinados factores.

Esta última circunstancia es similar a la que se presenta en el entrenamiento mismo de los sistemas algorítmicos. La elaboración de una metodología y una taxonomía lo más neutra posible, como la que se pretende exponer en este trabajo, resulta un paso previo indispensable para asegurar la imparcialidad de los sistemas dotados de IA. La pérdida de esta imparcialidad generará o incrementará las externalidades sufridas por los grupos e individuos más vulnerables en cada momento.

The AI Risk Repository

"The AI Risk Repository: A comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence" propone una aproximación basada en dominios y subdominios que los divide en:

- Discriminación y toxicidad
    - Discriminación injusta e infra/sobrerrepresentación
    - Exposición a contenido tóxico
    - Funcionamiento desigual en función del grupo
- Privacidad y seguridad
    - Comprometimiento de la privacidad
    - Vulnerabilidades del sistema de IA y ataques al mismo
- Desinformación

    - Información falsa o confusa que se sube de forma inadvertida
    - Contaminación del ecosistema informativo y pérdida de la realidad consensuada
- Actores maliciosos y usos fraudulentos
    - Desinformación, vigilancia e influencia sistemáticas
    - Ciberataques
    - Fraudes y manipulaciones
- Interacción humano-máquina
    - Exceso de confianza en la IA, incluida la generada por la antropoformización
    - Pérdida de la agencia y autonomía humana
- Daños socioeconómicos y ambientales
    - Centralización del poder y los beneficios de la IA
    - Incrementos de la desigualdad y pérdida de calidad del empleo
    - Devaluación económica y cultural del esfuerzo humano
    - Efectos derivados de la carrera empresarial por el desarrollo de la IA
    - Fallos en la gobernanza
    - Daños ambientales
- Seguridad, fallos y limitaciones de la IA
    - Priorización de los objetivos de las máquinas frente a los de los humanos
    - Posesión de capacidades peligrosas por parte de la IA
    - Falta de capacidad o robustez de la IA que provoque fallos en la fiabilidad de los sistemas
    - Falta de transparencia e interpretabilidad
    - Supuestos derechos de los robots

Sociotechnical Harms of Algorithmic Systems

Mientras tanto, en "Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction"[1] se presentan cinco categorías principales, divididas a su vez en un número variable de casos.

[1] https://doi.org/10.1145/3600211.3604673

En cualquier caso, esta taxonomía se centra en daños.

Figure 1: Sociotechnical harms taxonomy overview.

Mitigación de las externalidades

Mientras que las externalidades tecnológicas requieren de soluciones tecnológicas para su mitigación, las técnicas deberán abordarse desde los ámbitos de la concienciación y la educación. Es decir, la mitigación de las vulnerabilidades que surgen del uso de la tecnología y de la explotación de los factores psicológicos o sociales no puede basarse completamente en la implementación de soluciones de carácter puramente tecnológico. Para conseguirlo se requerirán técnicas legales (legislación), didácticas (concienciación y formación) y políticas.

Aunque es relativamente sencillo proceder a un desglose mayor de las vulnerabilidades y las externalidades, no cabe explorar soluciones diferenciadas para distintos grupos sociales afectados por las mismas. No obstante, estas soluciones genéricas sí deberán ser adaptadas -y presentadas de forma diferencial- a cada uno de los diferentes grupos sociales afectados.

Implementación de medidas para la mitigación de las externalidades

Para abordar la mitigación de las externalidades asociadas con la inteligencia artificial cabe desarrollar un enfoque integrador que considere tanto la dimensión técnica como la social y psicológica de estas externalidades. Este enfoque debe estructurarse en torno a principios éticos claros, estrategias normativas y acciones prácticas, teniendo en cuenta las particularidades de los grupos o individuos en situación de vulnerabilidad.

Priorización de principios éticos en el desarrollo de la IA

- **Equidad**: garantizar que los algoritmos no reproduzcan ni amplifiquen sesgos existentes, y que promuevan un trato equitativo hacia todos los usuarios, independientemente de su contexto socioeconómico, cultural o geográfico.
- **Transparencia**: fomentar sistemas de IA comprensibles y auditables, que permitan a los usuarios y reguladores identificar y mitigar potenciales externalidades negativas.
- **Responsabilidad**: establecer mecanismos claros para que los desarrolladores, implementadores y usuarios finales sean responsables de los efectos que generan los sistemas de IA.

Estrategias normativas para la regulación de la IA

- **Regulación proactiva de los sesgos algorítmicos**: obligar a las entidades desarrolladoras a realizar auditorías independientes y periódicas que identifiquen y mitiguen sesgos.
- **Protección de los datos personales**: asegurar que la recolección, almacenamiento y uso de datos por parte de sistemas de IA cumplan con estrictos estándares de privacidad y ética, con especial atención a la protección de los datos de grupos vulnerables.
- **Requisitos de inclusión**: fomentar que las bases de datos utilizadas para entrenar modelos de IA representen adecuadamente la diversidad de la población y eviten la infrarrepresentación de minorías.

Educación y concienciación social

- **Fomento de la alfabetización digital**: implementar programas educativos que reduzcan la brecha digital, especialmente entre poblaciones vulnerables. Esto incluye desde formación básica hasta talleres avanzados que permitan un uso crítico y autónomo de las tecnologías.
- **Sensibilización sobre las externalidades**: desarrollar campañas informativas que expliquen cómo las externalidades pueden afectar a distintos grupos y qué medidas pueden tomar los usuarios para protegerse.
- **Formación continua para desarrolladores**: incluir módulos éticos y de impacto social en los programas de formación técnica, asegurando que los desarrolladores sean conscientes de las implicaciones de sus decisiones algorítmicas.

Políticas específicas para grupos vulnerables

- **Acceso garantizado a la tecnología**: establecer programas gubernamentales que faciliten el acceso a dispositivos y redes digitales en regiones remotas o desconectadas.
- **Subsidios para la formación tecnológica**: ofrecer incentivos económicos o becas para que poblaciones económicamente desfavorecidas puedan acceder a formación en IA.
- **Protección de los niños en el entorno digital**: diseñar políticas específicas para la protección de menores, enfocadas en garantizar un uso seguro y adecuado de tecnologías basadas en IA.

Fortalecimiento de la gobernanza internacional

- **Establecimiento de estándares internacionales**: desarrollar normativas globales para la gestión de externalidades, que sirvan como referencia para los marcos regulatorios nacionales.
- **Promoción de colaboraciones transnacionales**: fomentar iniciativas conjuntas entre gobiernos, instituciones académicas y empresas tecnológicas para investigar y mitigar las externalidades.
- **Transferencia tecnológica**: garantizar que los avances en IA no profundicen las desigualdades entre naciones desarrolladas y en desarrollo, mediante programas de cooperación y transferencia de conocimientos.

Evaluación de impacto y mejora continua

Para garantizar la eficacia de las medidas propuestas, es fundamental implementar sistemas de evaluación continua que permitan monitorear el impacto de las externalidades y la efectividad de las políticas de mitigación:

- **Indicadores de vulnerabilidad**: establecer métricas específicas que permitan identificar y medir el nivel de vulnerabilidad de diferentes grupos frente a las externalidades de la IA.
- **Auditorías regulares**: realizar evaluaciones periódicas de los sistemas de IA y de las políticas implementadas, para identificar áreas de mejora.
- **Incorporación de retroalimentación**: garantizar que las comunidades afectadas tengan voz en el diseño y ajuste de las políticas, asegurando así que estas sean pertinentes y efectivas.

Concreción del impacto de las externalidades de la IA en los grupos vulnerables

Este enfoque permite identificar patrones recurrentes, analizar los factores que agravan las externalidades y proponer soluciones personalizadas pero

extrapolables. A continuación, se presentan diversos estudios que ilustran la diversidad y profundidad de estas problemáticas.

1. Desinformación dirigida en poblaciones con acceso limitado a la alfabetización digital

La propagación de contenido desinformativo es una externalidad crítica en el uso de sistemas de IA diseñados para generar y personalizar información. Grupos con baja alfabetización digital, como comunidades rurales o personas mayores con escaso acceso a educación tecnológica, son particularmente vulnerables:

- **Manifestación del problema**: Algoritmos diseñados para maximizar el *engagement* priorizan contenido sensacionalista o desinformativo, exponiendo de forma desproporcionada a estas poblaciones a narrativas falsas o manipulativas.
- **Consecuencias**: En contextos electorales, por ejemplo, estas externalidades han resultado en la manipulación de opiniones políticas; mientras que, en entornos de salud, la circulación de información falsa ha dificultado campañas de vacunación y el acceso a tratamientos efectivos.
- **Medidas propuestas**: Programas de alfabetización digital adaptados a las necesidades de cada grupo y regulación más estricta sobre la transparencia y los objetivos de los sistemas de recomendación algorítmica.

2. Exclusión de lenguas minoritarias en sistemas de procesamiento de lenguaje natural

Los sistemas de IA de procesamiento de lenguaje natural han demostrado un progreso significativo en la comprensión y generación de texto, pero su diseño suele priorizar lenguas mayoritarias, dejando rezagadas a las lenguas minoritarias:

- **Manifestación del problema**: Idiomas hablados por pequeñas comunidades o en regiones específicas, como lenguas indígenas, suelen quedar excluidos de los conjuntos de datos de entrenamiento, perpetuando su invisibilidad en las herramientas digitales.
- **Consecuencias**: Esta exclusión amplifica la marginación cultural y lingüística, dificultando la integración de estas comunidades en el mundo digital y limitando su acceso a servicios críticos que dependen del PLN.
- **Medidas propuestas**: Incentivar la inclusión de lenguas minoritarias en los conjuntos de datos de entrenamiento mediante financiamiento público, desarrollo de estándares abiertos y colaboración con comunidades locales para la recopilación ética de datos lingüísticos.

3. Sesgos algorítmicos en sistemas de selección laboral

Las plataformas de IA utilizadas en procesos de selección de personal han sido objeto de críticas por reforzar sesgos discriminatorios en perjuicio de ciertos grupos:

- **Manifestación del problema**: Algoritmos entrenados con datos históricos de contratación tienden a replicar patrones de exclusión, como la discriminación por género, edad o raza, invisibilizando a candidatos pertenecientes a minorías o con trayectorias atípicas.
- **Consecuencias**: Estos sesgos no solo reducen las oportunidades laborales de los grupos afectados, sino que también perpetúan estructuras de desigualdad económica y social en los mercados laborales.
- **Medidas propuestas**: Implementación de auditorías algorítmicas obligatorias para detectar y corregir sesgos en los sistemas de selección y promoción de una mayor diversidad en los datos de entrenamiento, con supervisión externa independiente.

Propuestas específicas para la mitigación de externalidades en casos concretos

Además de identificar las manifestaciones específicas de las externalidades, resulta prioritario proponer medidas adaptadas que reduzcan su impacto en los grupos vulnerables:

1. **Herramientas inclusivas y adaptativas**: Crear sistemas que prioricen la accesibilidad y la personalización para contextos diversos, con énfasis en poblaciones con capacidades limitadas.
2. **Colaboración con comunidades afectadas**: Involucrar a las comunidades directamente impactadas en el diseño y evaluación de sistemas de IA, garantizando que sus necesidades y perspectivas sean consideradas.
3. **Pruebas piloto en contextos críticos**: Antes de la implementación masiva de sistemas de IA, realizar estudios piloto en comunidades vulnerables para evaluar impactos potenciales y ajustar los sistemas en función de los resultados.