TEC | Tecnológico de Costa Rica

Escuela de Administración de Tecnologías de Información

Análisis Comparativo del Comportamiento Altruista de la Inteligencia Artificial Generativa mediante Dictator Games y Metaanálisis del Comportamiento Humano

Trabajo Final de Graduación para optar al grado de Licenciatura en Administración de Tecnología de Información

Elaborado por: Valeria María Martínez Rojas

Prof. Tutor: Dr. Federico Torres Carballo

Cartago, Costa Rica

Semestre

Noviembre, 2024

Hoja de Aprobación

ESCUELA DE ADMINISTRACIÓN DE TECNOLOGÍA DE INFORMACIÓN
GRADO ACADÉMICO LICENCIATURA

Los miembros de Tribunal Examinador de la Escuela de Administración de Tecnología de Información, recomendamos que el siguiente Trabajo Final de Graduación de la estudiante Valeria María Martinez Rojas sea aceptado como requisito parcial para optar por el grado académico de Licenciatura en Administración de Tecnología de Información.

TEC | Tecnológico de Costa Rica

Firmado digitalmente por YARIMA TATIANA SANDOVAL SANCHEZ (FIRMA)
Fecha: 2024.11.29 16:34:52 -06'00'

Ing. Yarima Sandoval Sánchez

Coordinación Trabajo Final de Graduación

FEDERICO TORRES CARBALLO (FIRMA)

Firmado digitalmente por FEDERICO TORRES CARBALLO (FIRMA)
Fecha: 2024.11.28 16:43:40 -06'00'

Dr. Federico Torres Carballo

Profesor Tutor

Digitally signed by ISAAC ALPIZAR CHACON (FIRMA)
Date: 2024.11.28 18:42:40 +01'00'

Dr. Isaac Alpízar Chacón

Lector Académico

TEC | Tecnológico de Costa Rica

Firmado digitalmente por YARIMA TATIANA SANDOVAL SANCHEZ (FIRMA)
Fecha: 2024.11.29 16:34:35 -06'00'

Ing. Yarima Sandoval Sánchez

Lectora Académica

Dedicatoria

.

A mi familia, por ser el pilar fundamental de mi vida y por brindarme siempre su amor, apoyo incondicional y confianza en cada paso que doy.

A mis padres, quienes me enseñaron el valor del esfuerzo, la perseverancia y la importancia de seguir mis sueños sin importar las adversidades.

A Ariel Rodríguez, quien fue mi motivación para terminar. por su comprensión y por ser mi refugio en los momentos de incertidumbre.

A mi profesor tutor, Dr. Federico Torres, cuya guía y sabiduría fueron clave en la culminación de este proyecto.

Gracias por estar conmigo en este camino y hacer de este logro algo tan significativo.

Resumen

Este estudio tiene como objetivo principal comparar el comportamiento altruista entre agentes de inteligencia artificial generativa (IAG) y humanos a través de experimentos de tipo *Dictator Game*. El trabajo se enmarca en la economía experimental y utiliza modelos avanzados de IAG, como GPT-4 y GPT-3.5 turbo, para evaluar sus decisiones en contextos que simulan situaciones de toma de decisiones altruistas. La comparación busca identificar patrones de comportamiento y diferencias significativas entre la capacidad de los humanos y la IA para tomar decisiones que beneficien a otros, más allá del interés propio.

En el desarrollo del proyecto se implementaron diversas fases, incluyendo la recopilación de datos, análisis de resultados y evaluación de hipótesis. Los experimentos fueron diseñados cuidadosamente para simular situaciones controladas, permitiendo así observar la respuesta tanto de los humanos como de los agentes de IA.

La investigación aporta datos empíricos valiosos para la comprensión del comportamiento altruista en la inteligencia artificial, contribuyendo a debates actuales sobre la ética y el diseño de sistemas de IA. Los hallazgos sugieren que, aunque los agentes de IA pueden ser entrenados para emular comportamientos altruistas, su toma de decisiones aún depende en gran medida de los datos y algoritmos subyacentes, lo que implica un sesgo potencial. Además, el estudio destaca la necesidad de abordar estos sesgos para mejorar la fiabilidad y aceptación de estas tecnologías en aplicaciones sociales y económicas.

Este trabajo también se alinea con los Objetivos de Desarrollo Sostenible (ODS) de la ONU, en particular con aquellos que buscan promover la innovación responsable y el desarrollo de tecnologías éticas. El estudio concluye proponiendo líneas de investigación futuras para mejorar la capacidad de los agentes de IA en la simulación de comportamientos prosociales, así como para desarrollar políticas que apoyen la implementación ética de la inteligencia artificial en diversas áreas de la sociedad.

En conclusión, el TFG demuestra que, aunque la inteligencia artificial generativa muestra avances significativos en la emulación de comportamientos altruistas, aún existen desafíos para igualar la complejidad del comportamiento humano. Los resultados obtenidos abren el camino a nuevas investigaciones para refinar estos modelos y su aplicación en contextos de toma de decisiones éticas y equitativas.

Tabla de Contenidos

Índice de Figuras

Índice de Tablas

1. Introducción

El presente estudio aborda el comportamiento altruista en la inteligencia artificial generativa (IAG). Por esta razón, el objetivo de la presente sección consiste en ofrecer al lector el contexto sobre el estudio, con el fin de asegurar la comprensión de la línea investigativa del estudio. Por este motivo, en esta sección se establecen los antecedentes, el problema, la justificación, los objetivos y otras secciones que describen la investigación presentada.

1.1. Descripción General

En el contexto de los avances tecnológicos actuales y el desarrollo de la inteligencia artificial, se presenta un estudio cuyo propósito es investigar y comparar el comportamiento altruista de la inteligencia artificial generativa con el de los humanos. Por esta razón, la investigación se llevará a cabo en forma de experimentos basados en la herramienta experimental *Dictator Game.*

La iniciativa llega en el momento en el que la inteligencia artificial (IA) se perfila como una tecnología disruptiva con un enorme potencial para impactar todos los aspectos de nuestras vidas, incluida la economía experimental y la toma de decisiones. Como consecuencia, este estudio busca comprender el comportamiento altruista de los agentes de IA y de los humanos con el fin de evaluar el impacto de estas tecnologías en la sociedad y apoyar a su integración en el día a día de todas las personas.

Esta investigación forma parte del Trabajo Final de Graduación (TFG) del Tecnológico de Costa Rica, en conjunto con la Escuela de Administración de Tecnología de Información y el profesor tutor Dr. Federico Torres Carballo. Con el aporte de esta investigación, se espera avanzar en el conocimiento de los campos tanto de la inteligencia artificial como del comportamiento humano y proporcionar información relevante y esclarecedora sobre la capacidad de los agentes de inteligencia artificial para imitar el comportamiento altruista observado en los humanos. Además, se espera que los resultados de este estudio sienten las bases para futuras investigaciones y para el desarrollo de políticas que promuevan el desarrollo ético y responsable de la inteligencia artificial en la sociedad.

1.2. Antecedentes

Las investigaciones sobre el altruismo en la toma de decisiones que se han desarrollado a lo largo de las últimas décadas han puesto especial atención a la manera en la que los seres humanos muestran comportamientos altruistas en diferentes contextos. Estos estudios, que datan de hace más de 50 o 60 años, buscan entender mejor las motivaciones y las condiciones bajo las cuales las personas toman decisiones altruistas.

El contexto de estas investigaciones es fundamental, como explican los autores Fehr y Fischbacher, (2003), ya que las decisiones altruistas no siempre se cuantifican o modelan con parámetros específicos o *software*. Por esta razón, a diferencia de otros ámbitos donde los resultados llegan a ser más predecibles y medibles, el comportamiento humano en términos de altruismo es complejo y multifacético. A partir de esta idea, se entiende que dichos estudios

comparen frecuentemente el comportamiento humano con el de agentes inteligentes, buscando entender las diferencias y similitudes en sus respuestas a diferentes estímulos y situaciones.

En la actualidad, la información recabada de estos estudios es utilizada en diversas áreas como el *marketing*, las finanzas y prácticamente todos los procesos de toma de decisiones que involucren interacciones humanas. La investigación en este campo no solo proporciona conocimiento sobre cómo y por qué las personas actúan de manera altruista, sino que también ayuda a desarrollar modelos y sistemas que predicen o fomentan este tipo de comportamientos (Comunicación personal, Torres Carballo, 2024).

Un aspecto crucial de las investigaciones contemporáneas de este tema, según explican Caliskan et al., (2017) en su artículo *Semantics derived automatically from language corpora contain human-like biases*, es examinar si las inteligencias artificiales, en particular la inteligencia artificial generativa (IAG), presentan algún sesgo en sus decisiones que influye en su interacción con los seres humanos. Hasta ahora, los estudios sobre sesgos en IA han abarcado principalmente aspectos de género y etnia, pero es posible que existan otros tipos de sesgos relacionados con el comportamiento, incluido el altruismo.

Asimismo, determinar si las IAG muestran algún tipo de sesgo en contextos altruistas podría tener importantes implicaciones. Torres Carballo (Comunicación personal, 2024) explica que si se encontrara que las IA tienen sesgos en este ámbito, podría conducir a una serie de investigaciones posteriores para identificar las causas de dichos sesgos y desarrollar métodos para corregirlos. Este tipo de análisis es parte integral de un campo de investigación más amplio que busca mejorar la interacción entre humanos y máquinas, conocido anteriormente como *interacción humano-computadora* y que ahora se centra más en la interacción entre inteligencias artificiales y seres humanos (Batson, 2010).

En resumen, estas investigaciones del área de economía experimental buscan entender mejor el comportamiento altruista tanto en humanos como en IA, y cómo esta comprensión mejora las interacciones entre ambas entidades. La identificación y corrección de posibles sesgos en IAG no solo mejoraría su rendimiento y fiabilidad, sino que también podría fomentar una mayor confianza y aceptación por parte de los usuarios. Por lo tanto, fortalecer el contexto y explicar claramente los resultados de estos estudios es esencial para avanzar en este campo de investigación.

1.2.1. Descripción de la organización

El presente Trabajo Final de Graduación (TFG) se desarrolla en el Tecnológico de Costa Rica (TEC), específicamente en la Escuela de Administración de Tecnología de Información (ATI). El TEC se describe como «una institución nacional autónoma de educación superior universitaria, dedicada a la docencia, la investigación y la extensión de la tecnología y las ciencias conexas para el desarrollo de Costa Rica» (TEC, 2022).

1.2.2. Misión

Contribuir al desarrollo integral del país, mediante formación del recurso humano, la investigación y la extensión; manteniendo el liderazgo científico, tecnológico y técnico, la excelencia académica y el estricto apego a las normas éticas, humanísticas y ambientales, desde una perspectiva universitaria estatal de calidad y competitividad a nivel nacional e internacional. (TEC, 2022)

1.2.3. Visión

El Tecnológico de Costa Rica seguirá contribuyendo mediante la sólida formación del talento humano, el desarrollo de la investigación, la extensión, la acción social y la innovación científico-tecnológica pertinente, la iniciativa emprendedora y la estrecha vinculación con los diferentes actores sociales a la edificación de una sociedad más solidaria e inclusiva; comprometida con la búsqueda de la justicia social, el respeto de los derechos humanos y del ambiente. (TEC, 2022)

1.2.3.1. Valores

A continuación, se presenta los valores del Tecnológico de Costa Rica:

Tabla 1 Valores del TEC

Ámbitos Institucional	Ámbito Individual
• Compromiso con la democracia • Libertad de expresión • Igualdad de oportunidades • Autonomía institucional • Libertad de cátedra • Búsqueda de la excelencia • Planificación participativa • Cultura de trabajo en equipo • Comunicación efectiva • Evaluación permanente • Vinculación permanente con la sociedad • Compromiso con la protección del ambiente y la seguridad de las personas • Compromiso con el desarrollo humano • Rendición de cuentas	• Respeto por la vida • Libertad • Ética • Solidaridad • Responsabilidad • Honestidad • Sinceridad • Transparencia • Respeto por todas las personas • Cooperación • Integridad • Excelencia

Nota. Adaptado de TEC, (2022).

1.2.3.2. Equipo de trabajo

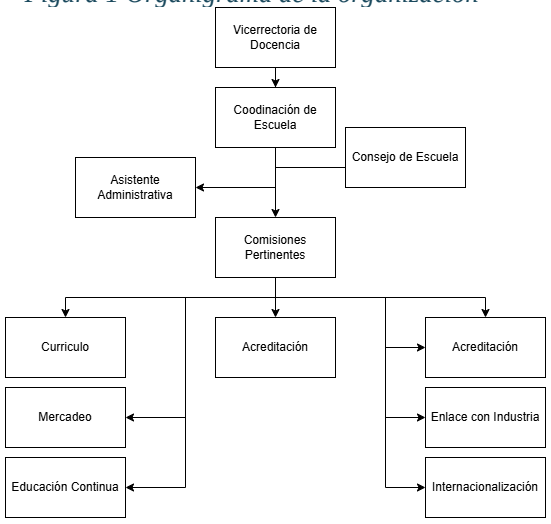
Esta sección describe el equipo de trabajo involucrado en el desarrollo del proyecto que se iría a desarrollar. La Tabla 2 Equipo de trabajo proporciona una descripción más detallada de las responsabilidades y roles de cada persona involucrada en el proceso de investigación.

Tabla 2 Equipo de trabajo

Miembro del equipo	Rol dentro del proyecto	Funciones
Estudiante	Investigador	Responsable de llevar a cabo todas las etapas de la investigación, desde la planificación hasta la presentación de los resultados. Genera los entregables del proyecto, como informes, análisis de datos y conclusiones.
Profesor tutor	Supervisor de la investigación	Supervisa la investigación, brindando orientación y apoyo al estudiante investigador. Evalúa el progreso del proyecto y proporciona retroalimentación para mejorar la calidad del trabajo. Revisa los entregables y proporciona comentarios para el desarrollo del proyecto. Participa en algunas etapas de la investigación con el fin de aportar un punto de vista diferente contribuyendo a la subjetividad de la investigación. Contribuye al proceso de investigación ofreciendo perspectivas alternativas y experiencias relevantes. Participa en discusiones, revisiones de literatura, análisis de datos o cualquier otra fase donde su experiencia sea beneficiosa para el proyecto.

A continuación, se presenta el organigrama de la Escuela de Administración de Tecnología de la Información, donde el estudiante realizará el proyecto de investigación.

Figura 1 Organigrama de la organización



1.2.4. Trabajos similares realizados dentro y fuera de la organización

Esta sección explora proyectos internos y externos pertinentes para proporcionar una comprensión global del tema e identificar ideas clave para el desarrollo del presente proyecto de investigación. Para esto se utilizará la Tabla 3 Proyectos similares, en donde se clasifica el proyecto, se menciona al autor o autores y se brinda una descripción relacionada con la intervención o aporte al desarrollo de la investigación.

Tabla 3 Proyectos similares

Tipo	Nombre	Autor / Autores	Descripción
Interno	*El impacto de soluciones basadas en Large Language Models en los negocios: una revisión sistemática de la literatura para identificar resultados, oportunidades y desafíos (Lim Ogawa Won Mi, 2023)*	Won Mi Lim Ogawa	El trabajo de graduación explora cómo los modelos de lenguaje grande (LLM por sus siglas en inglés) ofrecen oportunidades prometedoras para las empresas, especialmente en el área de atención al cliente y *marketing*. Sin embargo, su implementación también presenta desafíos, como el sesgo potencial, los riesgos de privacidad de datos, los costos computacionales y el desplazamiento laboral.
Interno	*Redes neuronales y autómatas finitos (Helo-Guzmán, 2019)*	José E. Helo-Guzmán	El estudio analiza los autómatas finitos (AF), modelos computacionales simples con propiedades bien estudiadas. Han sido utilizados con éxito en el desarrollo de programas de análisis léxico, como los compiladores. La inteligencia artificial de estos modelos busca crear sistemas computacionales que imitan el comportamiento humano, incluyendo la capacidad de aprender y resolver problemas.
Externo	*Evidence of behavior consistent with self-interest and altruism in an artificially intelligent agent* T. Johnson & Obradovich (2023)	Tim Johnson Nick Obradovich	El artículo explora la presencia de comportamientos altruistas y egoístas en los procesos de toma de decisiones de los agentes de IA. Empleando agentes de IA OpenAI de diversa complejidad, el estudio analiza sus respuestas a los incentivos y su comportamiento en escenarios.
Externo	*Medición experimental del Comportamiento Organizacional Ciudadano: Altruismo, Aversión al Riesgo y Deportividad (Torres-Carballo et al., 2018)*	Federico Torres Carballo, Néstor Morales Rodríguez, Grettel Brenes Leiva y Martin Solís Salazar	Este estudio se centra en dos componentes del OCB por sus singlas en ingles *Organizational Citizenship Behavior*: la deportividad y el altruismo. Emplea técnicas experimentales para medir el comportamiento individual en contextos controlados que reflejan fielmente las situaciones empresariales reales. Además, considera la aversión al riesgo como un factor que influye en el comportamiento individual.

1.3. Planteamiento del problema

En esta sección se expone la situación problemática identificada y se plantean las preguntas de investigación que guiarán el desarrollo del proyecto y justifican su realización. Además, se detallan los beneficios esperados del producto resultante.

1.3.1. Situación problemática

El campo de la inteligencia artificial está en constante evolución, lo que ubica al estudio de la toma de decisiones por parte de la IA en una posición cada vez más relevante. Brühl (2024), en su artículo *Artificial Intelligence and the Economy*, detalla que a medida que la IAG emerge como una tecnología disruptiva con el potencial de transformar diversos aspectos de nuestras vidas, surge la necesidad de explorar su impacto en el ámbito de toma de decisiones. Esta transformación incluye la capacidad de los agentes de IA para tomar decisiones que van desde la maximización de beneficios personales hasta la adopción de comportamientos altruistas.

Para comprender cómo afectan los comportamientos altruistas en la toma de decisiones, se detalla la explicación dada por Fehr y Fischbacher (2003), en donde mencionan que:

> Las decisiones altruistas son aquellas en las que un individuo elige actuar en beneficio de otros, incluso a costa de su propio interés personal. Estas decisiones son motivadas por la preocupación por el bienestar ajeno e implican sacrificios personales, ya sea en términos de tiempo, recursos o comodidad. A diferencia de las decisiones basadas en el interés propio, donde el objetivo principal es maximizar el beneficio personal, las decisiones altruistas buscan maximizar el beneficio colectivo o el bienestar de otros individuos. (Fehr y Fischbacher, 2003)

En este contexto, los autores T. Johnson y Obradovich (2023) en su artículo *Evidence of behavior consistent with self-interest and altruism in an artificially intelligent agent* explican que:

> Es fundamental comprender cómo se comportan los agentes de IA en escenarios de toma de decisiones para evaluar sus capacidades y su impacto potencial en la sociedad. El estudio emplea un enfoque experimental de múltiples condiciones, incentivando a los agentes de IA de OpenAI para que tomen decisiones que reflejen o bien la maximización de beneficios por interés propio o bien un comportamiento altruista. (T. Johnson y Obradovich, 2023)

El estudio de T. Johnson y Obradovich (2023) recalca la importancia de comprender el comportamiento de los agentes de IA en la toma de decisiones altruistas, debido a que permite evaluar sus capacidades y su posible impacto en la sociedad. Con esta idea en mente, la Escuela de Administración de Tecnología de la Información (ATI), bajo la tutela del Dr. Federico Torres Carballo, ha planteado el desafío de investigar y comprender las capacidades desconocidas de los agentes de IA que influyen en la toma de decisiones altruistas, asimismo son comparadas con los patrones establecidos de la toma de decisiones humana en contextos controlados, bajo los experimentos de *Dictator Game*.

Entre las causas de la problemática de estudio, según lo señalado por T. Johnson y Obradovich (2023) en su artículo, resaltan la complejidad del comportamiento humano, influenciada por una variedad de factores internos y externos, y la complejidad de los modelos subyacentes de IA, que dificultan la simulación de contextos específicos. Así mismo, recalcan la falta de comprensión sobre el funcionamiento de los algoritmos de IA y la variabilidad en la sofisticación de los agentes de inteligencia artificial, que también contribuyen a esta problemática.

Por otro lado, los autores T. Johnson y Obradovich (2023) también mencionan que la creación de entornos de simulación realistas y la precisión en los contextos de toma de decisiones son desafíos importantes por considerar. Estos factores afectan directamente a la escasez de datos detallados y relevantes sobre el comportamiento humano. No obstante, las limitaciones en la capacidad de los agentes de IA para procesar información social también son factores que influyen en esta situación.

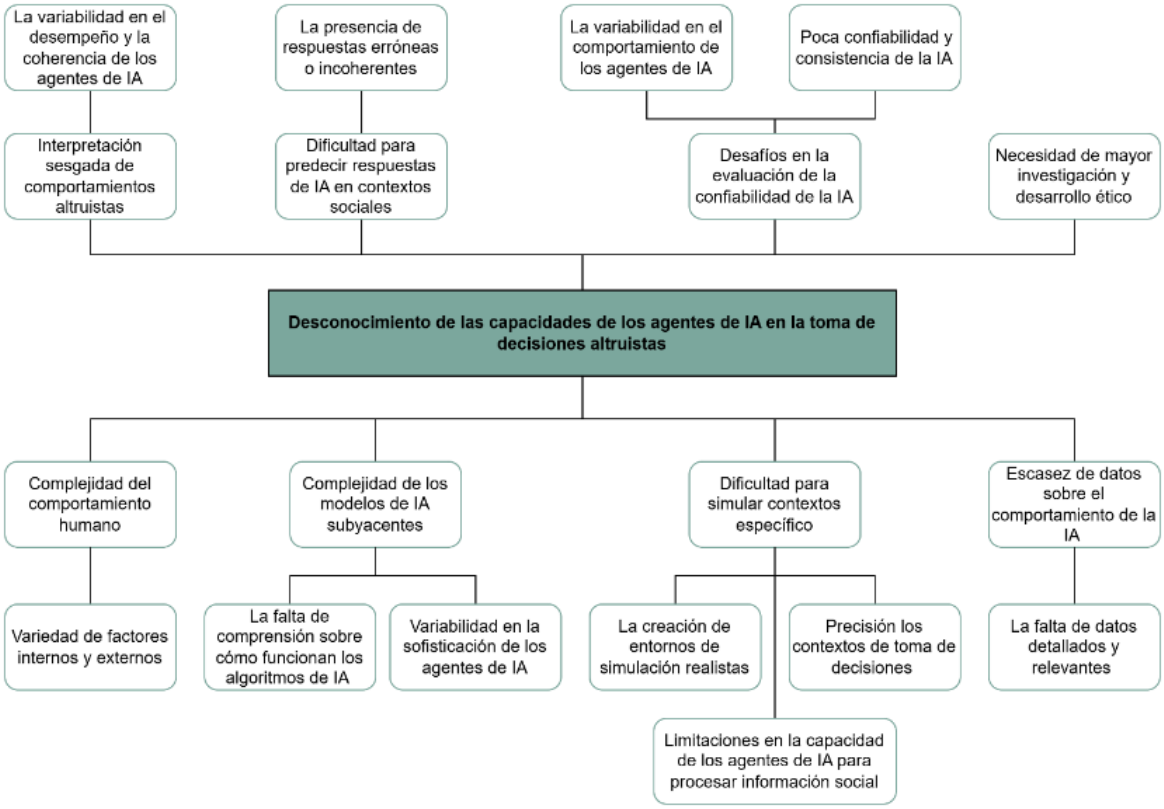
Bajo estas premisas, esta iniciativa constituye uno de los aspectos más destacados del actual proyecto de investigación en ATI, dado que propone explorar la interacción entre agentes inteligentes, enfocándose en la solución desarrollada por OpenAI, ChatGPT (OpenAI, s/f-b), como una herramienta de IAG. Con el objetivo de analizar la pregunta generadora de la investigación, la cuál es «¿cuál es la variación de las capacidades de los agentes de IAG en la toma de decisiones altruistas?», este enfoque no solo representa una nueva frontera en la economía experimental, sino que también ofrece la oportunidad de desarrollar nuevas herramientas y metodologías para mejorar la comprensión y la toma de decisiones.

En lo que respecta al enfoque planteado en esta investigación, este conlleva una serie de interrogantes y desafíos fundamentales que se abordarán a lo largo del estudio:

- ¿Cuál es el estado actual del uso de la inteligencia artificial generativa en la simulación y análisis del comportamiento altruista?
- ¿En qué medida los agentes inteligentes emulan el comportamiento altruista humano en contextos de toma de decisiones?
- ¿Qué elementos influyen en las decisiones altruistas tomadas por los agentes inteligentes generativos?

La Figura 2 Árbol del problema representa gráficamente las causas del problema de estudio y sus efectos directos.

Figura 2 Árbol del problema



1.3.2. Justificación del proyecto

En este apartado se brinda la justificación del problema de estudio a partir del abordaje del planteamiento del problema y de su relación con las áreas de estudio de la Escuela de Administración de Tecnología de Información. Además, se presenta cómo está asociado el problema con los ejes de conocimiento estratégico del Tecnológico de Costa Rica y con los objetivos de desarrollo sostenible establecidos por la ONU.

Por otro lado, la presente investigación se enmarca en el área de investigación de la Escuela de Administración de Tecnología de Información conocida como «Aprovechamiento de sistemas de información en la sofisticación del entorno social y su impacto en la comunidad». Este enfoque está relacionado con la economía experimental y el estudio de la influencia de los agentes inteligentes, desarrollados a través de sistemas de información avanzados, en la toma de decisiones altruistas, las cuales, según estudios, tienen la posibilidad de causar impactos en la comunidad. Al entender y analizar cómo estos sistemas serán aprovechados para mejorar y sofisticar el entorno social, se busca no solo avanzar en el conocimiento académico, sino también contribuir de manera tangible al bienestar y desarrollo de la sociedad.

La sofisticación del entorno social a través del uso de sistemas de información tiene el potencial de transformar las dinámicas comunitarias y económicas. Al estudiar cómo los agentes inteligentes se comportan en situaciones que requieren decisiones altruistas, el proyecto contribuye tanto al

avance del conocimiento en inteligencia artificial y economía experimental, como a las implicaciones prácticas para el desarrollo, que influyen en los comportamientos prosociales y ético. Es por esta razón que la investigación explora cómo las tecnologías disruptivas, como la inteligencia artificial, tiene la capacidad de ser integradas de manera efectiva para promover un entorno social más sofisticado, con implicaciones directas en la innovación, la gobernabilidad y el desarrollo sostenible.

La irrupción de la IA en diversos campos ha generado un interés creciente por entender su impacto en la toma de decisiones económicas. En particular, la inteligencia artificial generativa representa una tecnología disruptiva con el potencial de transformar significativamente nuestra sociedad. En este contexto, el Tecnológico de Costa Rica (TEC), a través de los antecedentes del Laboratorio de Economía Experimental (LEX-TEC) y la Escuela de Administración de Tecnología de la Información (ATI), se encuentra en una posición privilegiada para abordar el desafío.

El árbol del problema, planteado en la Figura 2 Árbol del problema, presenta la complejidad y la interrelación de las diferentes causas que contribuyen al desconocimiento del comportamiento de los agentes de IA en contextos de toma de decisiones altruistas. Este sesgo informativo tiene profundas implicaciones tanto en la investigación académica como en la aplicación práctica de la inteligencia artificial en diversos sectores. Por este motivo, el proyecto se centra en la manera en la que los agentes inteligentes, especialmente los desarrollados por OpenAI, como ChatGPT (OpenAI, s/f-b), toman decisiones en comparación con los humanos en contextos de toma de decisiones altruistas.

La solución propuesta aborda el diseño y la selección de los experimentos que simulan situaciones de *Dictator Game.* Dicho juego es un experimento económico en el que un participante, el «dictador», decide cómo dividir una cantidad de dinero entre él mismo y otro jugador. El segundo jugador no tiene poder de decisión y simplemente recibe la cantidad que el dictador elige darle. Por ende, se empleará este enfoque para 1) observar y analizar cómo reaccionan agentes inteligentes y humanos ante diferentes estímulos y escenarios; 2) identificar patrones de comportamiento; y 3) evaluar la coherencia y fiabilidad de las decisiones tomadas por agentes de inteligencia artificial y compararlas con las decisiones humanas.

Este proyecto busca establecer una relación con los ejes de conocimiento estratégicos del Tecnológico de Costa Rica, establecidos para el periodo 2023-2032 (TEC, 2023). Dado que la investigación se centra en la inteligencia artificial (IA) y su aplicación en la toma de decisiones, se alinea con el eje de la industria. De esta manera, el trabajo realiza una contribución significativa al desarrollo de las tecnologías de la información y la comunicación al estudiar cómo los agentes inteligentes toman decisiones en circunstancias específicas en comparación con los humanos.

Así mismo, promueve la innovación mediante el estudio de metodologías y herramientas para la toma de decisiones basadas en IA. Al abordar desafíos técnicos y éticos asociados con la IA, la investigación apoya el eje estratégico de innovación y desarrollo tecnológico del TEC, puesto que facilita avances en la creación de soluciones tecnológicas más eficientes.

A continuación, se especifica la alineación del proyecto con los Objetivos de Desarrollo Sostenible (ODS) de la Organización de las Naciones Unidas (ONU, 2015), acompañada de una explicación para comprender su relevancia y el impacto potencial en las diversas áreas en las que se aplique.

El primer objetivo que se encuentra directamente relacionado con el proyecto es ODS 8: Trabajo decente y crecimiento económico. Este cumple con las siguientes metas:

- Meta 8.2 «Lograr niveles más elevados de productividad económica mediante la diversificación, la modernización tecnológica y la innovación, entre otras cosas, centrándose en los sectores con gran valor añadido y un uso intensivo de la mano de obra» (ONU, 2015).
    - Relación con el proyecto: existe una conexión al explorar cómo la IA optimiza las operaciones empresariales y mejorar la toma de decisiones. Es decir, el proyecto contribuye a aumentar la productividad económica y fomentar la modernización tecnológica.
- Meta 8.3 «Promover políticas orientadas al desarrollo que apoyen las actividades productivas, la creación de puestos de trabajo decentes, el emprendimiento, la creatividad y la innovación, y fomentar la formalización y el crecimiento de las microempresas y las pequeñas y medianas empresas, incluso mediante el acceso a servicios financieros.» (ONU, 2015)
    - La investigación sobre el impacto de la IA en la economía y la toma de decisiones empresariales proporciona datos y conocimientos que ayudan a formular políticas que fomenten la innovación y la creación de empleos decentes.
- Meta 8.4 «Mejorar progresivamente, de aquí a 2030, la producción y el consumo eficientes de los recursos mundiales y procurar desvincular el crecimiento económico de la degradación del medio ambiente, conforme al Marco Decenal de Programas sobre modalidades de Consumo y Producción Sostenibles, empezando por los países desarrollados.» (ONU, 2015)
    - Relación con el proyecto: permite la investigación tener la posibilidad de ayudar a las empresas a utilizar la IA para optimizar sus procesos y mejorar la eficiencia de los recursos, contribuyendo a un crecimiento económico más sostenible.

El segundo objetivo que se encuentra directamente relacionado con el proyecto es ODS 9: Industria, innovación e infraestructura y cumple con las siguientes metas:

- Meta 9.1 «Desarrollar infraestructuras fiables, sostenibles, resilientes y de calidad, incluidas infraestructuras regionales y transfronterizas, para apoyar el desarrollo económico y el bienestar humano, haciendo hincapié en el acceso asequible y equitativo para todos» (ONU, 2015).
    - Relación con el proyecto: la investigación en IA y economía experimental impulsa el desarrollo de infraestructuras tecnológicas avanzadas que sean fiables y sostenibles, apoyando así el desarrollo económico y el bienestar humano.
- Meta 9.2 «Promover una industrialización inclusiva y sostenible y, de aquí a 2030, aumentar significativamente la contribución de la industria al empleo y al producto interno bruto, de acuerdo con las circunstancias nacionales, y duplicar esa contribución en los países menos adelantados» (ONU, 2015).
    - Relación con el proyecto: al fomentar la innovación y el desarrollo de nuevas aplicaciones de IA en contextos económicos, el proyecto contribuye a una industrialización más inclusiva y sostenible, con lo cual aumenta la participación de la industria en el empleo y en el producto interno bruto (PIB).

- Meta 9.5 «Aumentar la investigación científica y mejorar la capacidad tecnológica de los sectores industriales de todos los países, en particular los países en desarrollo, entre otras cosas fomentando la innovación y aumentando considerablemente, de aquí a 2030, el número de personas que trabajan en investigación y desarrollo por millón de habitantes y los gastos de los sectores público y privado en investigación y desarrollo.» (ONU, 2015)
    - Relación con el proyecto: la investigación contribuirá al aumento de la capacidad tecnológica y científica en el ámbito de la IA, fomentando la innovación y potenciando la inversión en investigación y desarrollo.

1.3.3. Beneficios esperados o aportes del Trabajo Final de Graduación

En la presente sesión se definen los beneficios que se esperan obtener a partir de la concreción del presente estudio. Estos se centran en el estudio exhaustivo de antecedentes y en el análisis detallado de los datos obtenidos de los experimentos aplicados, con el fin de profundizar en la comprensión del comportamiento altruista tanto en humanos como en inteligencias artificiales.

1.3.3.1. Beneficios directos

A continuación, se desglosan los beneficios directos de este proyecto:

- Desarrollo de un marco teórico sobre el comportamiento de agentes inteligentes en contextos de toma de decisiones.
- Generación de datos y evidencia empírica que respalden la comprensión del comportamiento de los agentes inteligentes en contextos de toma de decisiones altruistas.
- Expansión de la base científica para futuras investigaciones y desarrollos en el campo de la inteligencia artificial.
- Comprensión de los factores que influyen en la toma de decisiones de agentes inteligentes.
- Creación de un artículo científico (*paper*) con el resumen de los resultados parciales obtenidos durante la realización del TFG.

1.3.3.2. Beneficios indirectos

Enseguida se desglosan los beneficios indirectos de este proyecto de investigación:

- Fomentar la investigación interdisciplinaria para promover la implementación de soluciones basadas en inteligencia artificial en diversos campos empresariales.
- Establecer redes de colaboración entre investigadores, instituciones académicas y organizaciones empresariales para impulsar la formulación de propuestas de proyectos de investigación.
- Facilitar la implementación de tecnología en las áreas de negocio de las organizaciones.
- Estimular la innovación en el campo de la inteligencia artificial al identificar áreas de mejora en la toma de decisiones de los agentes.
- Fomentar el reconocimiento de oportunidades para mejorar procesos y aplicaciones empresariales.
- Contribuir al desarrollo de la teoría económica al proporcionar nuevas perspectivas sobre el comportamiento humano y artificial en contextos de toma de decisiones altruistas.

1.4. Objetivos del Trabajo Final de Graduación

En el presente apartado se establece el objetivo general y los objetivos específicos que se buscan desarrollar para del proyecto de investigación.

1.4.1.1. Objetivo General

- Determinar la variación de las capacidades de los agentes de inteligencia artificial generativa en la toma de decisiones altruistas, mediante la aplicación de experimentos de *Dictator Game*, para la comprensión de las diferencias en los patrones de decisión, durante el segundo semestre del 2024.

1.4.1.2. Objetivos Específicos

- Analizar la situación actual sobre el uso de la inteligencia artificial generativa en la simulación y análisis del comportamiento altruista para la evaluación de los métodos y herramientas utilizados en estudios previos.
- Determinar cuáles son los criterios que afectan la interacción de los agentes inteligentes en la toma de decisiones altruistas para la identificación de limitaciones relacionadas al comportamiento altruista.
- Examinar en qué medida los agentes inteligentes emulan el comportamiento humano en contextos de toma de decisiones altruistas para la comparación de los resultados obtenidos del comportamiento humano y experimentos con IAG.

1.5. Alcance

El alcance de este proyecto se enfoca en las investigación, ejecución y análisis de resultados de experimentos de tipo *Dictator Game* con el fin de medir las capacidades altruistas en la inteligencia artificial generativa; esto se logra utilizando enfoques metodológicos mixtos (cuantitativos y cualitativos) que contribuyen a la profundización en cada objetivo específico, planteados en la sección 1.4 Objetivos del Trabajo Final de Graduación.

Dentro del alcance del proyecto se definen tareas en función de cuatro grandes fases: planteamiento de la investigación, recopilación de datos, análisis de resultados y presentación de resultados. Estas fases se profundizan en la sección 3.10 Procedimiento metodológico de la investigación en la imagen Figura 7 Diagrama de fases y etapas del procedimiento metodológico. A continuación, se resume a gran medida cada una de las fases por desarrollar en el proyecto de investigación:

En la fase de planteamiento de la investigación, se llevará a cabo una exhaustiva revisión literaria para identificar estudios previos realizados en contextos similares. Esto permitirá establecer una base teórica sólida y asegurar que la investigación esté bien informada por el conocimiento existente. En esta fase también se definirán la hipótesis y las variables de investigación, fundamentales para guiar el enfoque y el diseño del estudio.

En la segunda fase, correspondiente a la recopilación de datos, se diseñarán y aplicarán experimentos del tipo *Dictator Game.* Estos son esenciales tanto para la exploración de las

decisiones y comportamientos de los participantes en situaciones controladas como para la proporción de datos empíricos clave para la investigación.

Durante la fase de análisis de resultados, se procederá a la interpretación de los datos obtenidos de los experimentos; este paso es crucial para asegurar la calidad y fiabilidad de los datos. Además, se llevará a cabo la validación de la hipótesis, evaluando si los resultados respaldan las hipótesis y objetivos planteados inicialmente.

Finalmente, una vez concluido el análisis de los resultados, se trabajará en la presentación de los mismos. Esto incluirá la elaboración de informes detallados y la preparación de presentaciones que comuniquen los hallazgos de manera clara y precisa, contribuyendo al conocimiento en el campo de estudio.

1.6. Supuestos

A continuación, se especifican los elementos, que se asume que se cumplirán o serán veredictos en la realización del proyecto.

- Se parte del supuesto de que los algoritmos utilizados por los agentes de inteligencia artificial generativa permanecen constantes durante el estudio y no experimentan cambios significativos.
- Se asume que los recursos literarios necesarios estarán disponibles y se podrán acceder mediante las condiciones establecidas para estudiantes, según lo planificado para la ejecución del proyecto.
- El proyecto se llevará a cabo bajo un constante monitoreo de los avances por parte del responsable académico, el profesor tutor de la investigación.
- Se asume que este proyecto no estará creando ningún experimento desde cero, sino que se diseña y selecciona un conjunto de experimentos ya establecidos.

1.7. Entregables

A lo largo de esta sección se establecen cuáles serán los entregables del proyecto de investigación. Además, se consideran los artefactos necesarios para la gestión del proyecto.

1.7.1. Entregables del producto

A continuación, se detallan puntalmente cuáles serán los entregables del proyecto.

- El estado del arte de la investigación, la cual presenta un análisis literario de los estudios previamente realizados y publicados sobre experimentos de tipo *Dictator Game* que se han llevado a cabo en conjunto con inteligencias artificiales y cuyo enfoque recae en los comportamientos altruistas.
- Un artículo científico, en el que se detallan los descubrimientos principales de la investigación, incluyendo, pero no limitado a, el contexto de la investigación, la síntesis de hallazgos, discusión de resultados y conclusiones y recomendaciones.

- Un documento académico, considerado el Trabajo Final de Graduación, donde se detalla el proceso realizado, la metodología aplicada, el resultado y las conclusiones de la investigación.

1.7.2. Gestión del proyecto

Durante la sección se presenta dos tipos de artefactos importantes para la gestión del proyecto, estos se detallan a continuación.

1.7.2.1. Minutas

Se establece el uso de minutas con el fin de realizar la trazabilidad de las reuniones establecidas con el equipo de trabajo del proyecto. Estas cumplen la función de mantener registros y recopilar información alrededor de los detalles más importantes de las sesiones. Para la utilización de este artefacto se realizará uso de la plantilla detallada en el 9.1 Apéndice A: Plantilla minuta de reuniones.

1.7.2.2. Gestión del Cambio

Con el fin de mantener el seguimiento de la gestión de cambios solicitados por el equipo de trabajo y de realizar la debida recopilación de requerimientos para los cambios solicitados, se establece el uso de una plantilla, detallada en el 9.2 Apéndice B: Plantilla de gestión de cambios, con el principal objetivo de realizar la debida recopilación de requerimientos para los cambios solicitados.

1.8. Limitaciones

En el siguiente apartado se indican los factores que tienen el potencial de restringir la realización del proyecto.

- El proyecto se ve limitado al uso de bases de datos suscritas a las que la institución TEC tenga acceso.
- Solo se considerarán artículos científicos en inglés y español como fuentes de información relevantes para el proyecto.
- La disponibilidad y calidad de los datos utilizados para entrenar a los agentes de inteligencia artificial generativa influye en los resultados del estudio, limitando la capacidad de análisis en aspectos estadísticos.
- La comprensión total del funcionamiento interno de los modelos de inteligencia artificial generativa tiende a ser limitada, lo que podría afectar la interpretación de los resultados y la identificación de los factores subyacentes que influyen en el comportamiento de los agentes, debido a la poca o nula información brindada al público por la empresa OpeanAI.
- La aplicación de los experimentos para la recolección de datos se basa en un batería de pruebas que son definidas por el profesor tutor.
- La aplicación de los experimentos se desarrollará en la solución de OpenAI (ChatGPT versión 4.0 y 3.5 turbo) exclusivamente.

- Las respuestas obtenidas de los experimentos se limitan exclusivamente a aquellas que correspondan a experimentos del tipo *Dictator Game* y que ayuden a la medición de comportamientos altruistas.

1.9. Exclusiones

Se presentan, a continuación, las exclusiones del proyecto:

- Se excluye de este proyecto cualquier tipo de experimentos o resultados que apoyen al entendimiento del comportamiento altruista que no sean de tipo *Dictator Game.*
- Se restringe el análisis únicamente a la inteligencia artificial generativa representada por la solución de OpenAI, ChatGPT V4 y ChatGPT V3.5 turbo, excluyendo cualquier otro tipo de IAG.

2. Estado del Arte

A continuación, se presenta la sección del estado del arte, donde se establece un listado de antecedentes teóricos y se presentan los principales conceptos relacionado al proyecto. Así mismo, se expone la sección de estudio previos, en donde se profundiza en distintos estudios recientes similares al proyecto; y, por último, se incorpora una tabla de resumen que detalla los principales hallazgos de la revisión literaria realizada.

2.1. Introducción

La tendencia en las investigaciones recientes explora el comportamiento de los grandes modelos de lenguaje (LLM, por sus siglas en inglés) en juegos estratégicos, ya que en estos se ha encontrado la inclinación a replicar tendencias humanas dirigidas hacia la equidad y la cooperación (Brookins y Debacker, 2024; T. Johnson y Obradovich, 2023). En estos estudios se destaca cómo los LLM, al procesar y generar texto, imitan patrones de comportamiento humano, lo que sugiere que estos modelos tienen la posibilidad de ser utilizados para estudiar y simular interacciones sociales complejas.

Sin embargo, algunos estudios advierten que los LLM tienden a presentar dificultades para desarrollar deseos basados en preferencias inusuales y para refinar creencias a partir de patrones simples (Fan et al., 2023). Esto indica una limitación en la capacidad de estos modelos a la hora de entender y replicar completamente la diversidad y la profundidad del comportamiento humano.

Por otro lado, las implementaciones experimentales de *Dictator Game* revelan diferencias en los incentivos estratégicos subyacentes, lo que altera las conclusiones sobre las preferencias por el bienestar de los demás (Grech y Nax, 2018). En estos juegos, un «dictador» decide cómo dividir una cantidad de dinero entre ellos y un destinatario, proporcionando una medida de comportamiento altruista.

Además, los metaanálisis de *Dictator Game* proporcionan información sobre los efectos de varias manipulaciones en el comportamiento de dar (Engel, 2010). Estos estudios compilan y analizan datos de múltiples experimentos para identificar tendencias generales y factores que influyen en la generosidad de los participantes. Esta investigación resulta fundamental para entender cómo diferentes factores contextuales alteran la disposición de las personas a compartir recursos.

Finalmente, la investigación también examina la intersección de la inteligencia artificial y los derechos humanos, enfatizando la necesidad de una regulación efectiva y de consideraciones éticas en el desarrollo de la IA (T. Johnson y Obradovich, 2023; Palomino-Flores et al., 2023).

A medida que la inteligencia artificial se integra más profundamente en la sociedad, es crucial abordar las implicaciones éticas y su influencia en comportamiento humanos, para garantizar que estas tecnologías se desarrollen y utilicen de manera que respeten la dignidad y los derechos de todas las personas. Esto incluye considerar cómo los sesgos en los modelos de IA amplifican las desigualdades existentes y cómo se mitigan estos riesgos a través de una regulación y supervisión adecuadas.

2.2. Antecedentes Teóricos

La inteligencia artificial y su capacidad para emular comportamientos éticos como el altruismo y el egoísmo han suscitado un interés considerable en la investigación contemporánea. Este interés se manifiesta en diversos estudios que exploran cómo los agentes de IA son diseñados para tomar decisiones que reflejen intereses propios o el bienestar común, lo cual tiene implicaciones significativas en campos como la teoría de juegos y los experimentos económicos. Por esta razón, se presentan los conceptos clave y algunos estudios relevantes que forman los antecedentes de este estado del arte.

Altruismo: es un comportamiento que beneficia a otros a expensas del propio bienestar del individuo. En el contexto de la IA, se refiere a la programación de agentes para que tomen decisiones que favorezcan al bien común o a los intereses de otros agentes o humanos (Grech y Nax, 2018; T. Johnson y Obradovich, 2023). Este concepto es esencial para el diseño de sistemas de IA que interactúan de manera ética y socialmente responsable.

Egoísmo: es un comportamiento que prioriza los intereses y beneficios personales sobre los de los demás. En la IA, esto implica que un agente tome decisiones que maximicen su propia utilidad sin considerar los efectos en otros agentes (Daylamani-Zad y Angelides, 2021; T. Johnson y Obradovich, 2023). Comprender el egoísmo en agentes de IA es crucial para prever y mitigar posibles conflictos en entornos de múltiples agentes.

Inteligencia artificial (IA): se refiere a la simulación de procesos de inteligencia humana por parte de sistemas informáticos. Estos procesos incluyen el aprendizaje (la adquisición de información y reglas para el uso de la información), el razonamiento (utilización de reglas para alcanzar conclusiones aproximadas o definitivas) y la autocorrección (T. Johnson y Obradovich, 2023). La capacidad de la IA para emular comportamientos humanos es fundamental para su aplicación en contextos éticos y estratégicos.

Inteligencia artificial general: según una investigación reciente de Fan et al. (2023), la IAG aspira a crear máquinas con una inteligencia comparable a la humana, capaces de aprender y resolver cualquier problema. A diferencia de las IA actuales, limitadas a tareas específicas, la IAG busca una inteligencia generalizada y adaptable.

Agentes autónomos: son sistemas de inteligencia artificial diseñados para operar de manera independiente y tomar decisiones sin intervención humana. Estos agentes son fundamentales en aplicaciones donde se requiere una toma de decisiones rápida y adaptativa, como en vehículos autónomos o sistemas de trading algorítmico (Brookins y Debacker, 2024).

Sistemas multiagente: consisten en múltiples agentes autónomos que interactúan entre sí dentro de un entorno común. Estos sistemas son estudiados para entender dinámicas de cooperación, competencia y negociación entre agentes (Fan et al., 2023).

Teoría de juegos: es una disciplina matemática que estudia las interacciones estratégicas entre diferentes agentes. Los juegos suelen ser de suma cero, donde las ganancias de un jugador son las pérdidas de otro, o de suma no cero, donde es posible que todos los jugadores se beneficien (Capraro et al., 2024; Fan et al., 2023). La teoría de juegos proporciona el marco teórico para comprender y analizar las decisiones estratégicas de los agentes de IA.

Juegos estratégicos: son un tipo de juego en teoría de juegos en los que los jugadores toman decisiones con el objetivo de maximizar su propio beneficio, teniendo en cuenta las decisiones potenciales de otros participantes. Los ejemplos incluyen el ajedrez, el póker y muchos modelos económicos (Brookins y Debacker, 2024; Fan et al., 2023). Estos juegos proporcionan un entorno controlado para estudiar el comportamiento estratégico de los agentes de IA.

***Dictator Game* :** son experimentos económicos diseñados para estudiar el comportamiento altruista. En estos juegos, un participante (el «dictador») decide cómo dividir una suma de dinero entre él mismo y otro jugador, sin ninguna restricción (Daylamani-Zad y Angelides, 2021; Grech y Nax, 2018). Estos juegos proporcionan un marco para analizar cómo las preferencias y los comportamientos altruistas tienden a ser modelados en agentes de IA.

Preferencias: en el contexto de la teoría de juegos y la economía del comportamiento, se refieren a las opciones o decisiones que los individuos toman cuando enfrentan diferentes alternativas. Estas preferencias tienden a ser altruistas, egoístas, racionales o irracionales (Grech y Nax, 2018). El estudio de las preferencias es vital para comprender cómo los agentes de IA son diseñados para tomar decisiones éticamente informadas.

Wording: es un aspecto fundamental en la realización de experimentos con modelos de inteligencia artificial generativa como ChatGPT. Este término se refiere a la redacción exacta de las instrucciones o preguntas que se le presentan al modelo para guiar su comportamiento y asegurar que su interpretación sea coherente con los objetivos del experimento. En estudios que involucran juegos económicos como el *Dictator game* o el dilema del prisionero, el *wording* juega un papel crucial, ya que pequeños matices en el lenguaje pueden influir significativamente en las decisiones que toma el modelo (Grech y Nax, 2018).

Modelos de lenguaje grandes (LLM): son algoritmos de inteligencia artificial que utilizan una gran cantidad de datos textuales para aprender patrones del lenguaje y generar texto coherente. Ejemplos notables incluyen GPT-3 y GPT-4 de OpenAI (Brookins y Debacker, 2024; Fan et al., 2023). Su capacidad para generar texto creativo y coherente abre un abanico de posibilidades en campos como la escritura creativa, la traducción y la generación de código.

Redes neuronales: son un conjunto de algoritmos, modelados vagamente según el cerebro humano, que están diseñados para reconocer patrones. Interpretan datos sensoriales a través de un tipo de percepción automática, etiquetando o agrupando la información cruda (Daylamani-Zad y Angelides, 2021). Estas redes son la base del aprendizaje profundo y se utilizan para desarrollar agentes de IA capaces de tomar decisiones complejas.

Algoritmos de Aprendizaje Reforzado: Los autores Fan et al. (2023) mencionan que estos algoritmos son métodos de aprendizaje automático en los que los agentes aprenden a tomar decisiones óptimas mediante la interacción con su entorno, recibiendo recompensas o castigos en función de sus acciones. Estos algoritmos son fundamentales para el desarrollo de agentes autónomos que mejoraran su comportamiento con el tiempo a través de la experiencia, lo que es crucial en aplicaciones como los videojuegos, la robótica, y los sistemas de recomendación. El aprendizaje reforzado permite a los agentes explorar diferentes estrategias y adaptarse a cambios en el entorno.

Economía experimental: Grech y Nax (2018) explican este concepto como una herramienta poderosa para comprender el comportamiento económico humano al proporcionar un entorno controlado para estudiar las decisiones económicas. Los resultados de estos experimentos se utilizan para desarrollar y validar teorías económicas, especialmente en áreas como la teoría de juegos y la economía del comportamiento.

Economía del comportamiento: esta área estudia los factores psicológicos, sociales y emocionales influyen en las decisiones económicas de las personas. A diferencia de la economía tradicional, que asume que los individuos son perfectamente racionales, la economía del comportamiento reconoce que las decisiones a menudo se ven afectadas por sesgos cognitivos y otras influencias no racionales (T. Johnson y Obradovich, 2023). Este campo ha proporcionado ideas clave para el diseño de agentes de IA que pueden emular comportamientos humanos más realistas.

Sesgos cognitivos: son desviaciones sistemáticas de la racionalidad en el juicio y la toma de decisiones. Estos sesgos impulsan a los individuos a tomar decisiones que no optimizan su utilidad, lo que contrasta con las predicciones de los modelos económicos tradicionales. Ejemplos de sesgos cognitivos incluyen el exceso de confianza, la aversión a las pérdidas y el sesgo de confirmación (T. Johnson y Obradovich, 2023). La incorporación de sesgos cognitivos en agentes de IAG permite una alineación más estrecha con el comportamiento humano observado en experimentos económicos.

La heurística: Según Grech y Nax (2018) explican que son herramientas mentales poderosas que nos ayudan a navegar por un mundo complejo; sin embargo, es importante ser conscientes de los sesgos que pueden introducir en nuestras decisiones. Comprender las heurísticas nos permite tomar decisiones más informadas y racionales, tanto a nivel individual como en el diseño de sistemas inteligentes.

Moralidad y ética en la IA: se refiere al estudio de la creación de los sistemas de inteligencia artificial en cuanto a su diseño y programación, con el fin de propiciar una toma de decisiones que respeten y se alineen con las normas morales y éticas. Esta disciplina es fundamental en el desarrollo de IA porque garantiza que las decisiones automatizadas no solo sean eficientes, sino también justas y responsables, dado que incluye la consideración de dilemas éticos, la implementación de principios morales en algoritmos y la evaluación de las consecuencias sociales y humanas de las decisiones tomadas por la IA (T. Johnson y Obradovich, 2023).

Comportamiento comparativo: según explican Brookins y Debacker (2024), el análisis comparativo del comportamiento se centra en identificar y evaluar las similitudes y diferencias en la toma de decisiones entre diferentes tipos de agentes, como humanos y sistemas de IA. Este enfoque es crucial para entender cómo las decisiones de la IA se alinean o divergen de las humanas, lo cual es importante para el diseño de sistemas que interactúen de manera coherente y predecible con los usuarios humanos. Además, los estudios de comportamiento comparativo revelan criterios sobre cómo ajustar los algoritmos para mejorar la cooperación o la competitividad entre humanos e IA.

Comportamiento prosocial: el comportamiento prosocial incluye acciones destinadas a beneficiar a otros, como el altruismo, la cooperación y la ayuda mutua. En el contexto de la

inteligencia artificial, Grech y Nax (2018) explican que el concepto se refiere a la programación de agentes para que tomen decisiones que favorezcan el bienestar colectivo, en lugar de maximizar su propio beneficio. Este tipo de comportamiento es esencial en sistemas de IA diseñados para trabajar en entornos colaborativos o para interactuar éticamente con humanos, garantizando que sus acciones contribuyan positivamente a la sociedad.

Equidad: la equidad en la IA se refiere al principio de igualdad en el tratamiento y la distribución de recursos, oportunidades o recompensas entre individuos o grupos. Fan et al. (2023) mencionan que este concepto es clave en el diseño de sistemas de inteligencia artificial a la hora de garantizar que los algoritmos no favorezcan injustamente a ciertos individuos o grupos sobre otros. En aplicaciones prácticas, la equidad es importante para evitar que las decisiones algorítmicas refuercen prejuicios existentes o creen nuevas formas de discriminación.

Interacciones hombre-máquina: este concepto, según la explicación de T. Johnson y Obradovich (2023), abarca el estudio de la interacción entre humanos y máquinas, incluyendo aspectos como la comunicación, la cooperación y el manejo de conflictos. Entender estas interacciones es clave para diseñar sistemas de IA que puedan integrarse de manera efectiva y segura en entornos humanos, ya que mejoran la usabilidad y la aceptación de la tecnología. Además, las interacciones hombre-máquina también son cruciales para el desarrollo de interfaces amigables que faciliten la colaboración y el entendimiento mutuo entre humanos y sistemas inteligentes.

Interdependencia en sistemas multiagente: la interdependencia en sistemas multiagente según Fan et al. (2023) se refiere a la relación entre agentes que dependen unos de otros para alcanzar sus objetivos, lo que implica cooperación, competencia o negociación entre ellos. En entornos multiagentes, estos deben interactuar de manera efectiva para lograr resultados óptimos, lo que requiere la formación de alianzas, la resolución de conflictos o la negociación de recursos. Este concepto es clave para diseñar sistemas de IA que operen en entornos colaborativos o competitivos, como los mercados financieros, los juegos multiagente o los sistemas de tráfico.

Modelos de decisión basados en valores: según T. Johnson y Obradovich (2023), estos modelos son sistemas que toman decisiones basadas en un conjunto de valores predefinidos, como la equidad, la justicia o la eficiencia. Ellos son cruciales para diseñar agentes de IA que actúen de acuerdo con normas éticas específicas, garantizando que sus decisiones reflejen los valores y principios de la sociedad o de la organización que los implementa. Por esta razón, este enfoque es particularmente relevante en aplicaciones donde las decisiones de la IA tengan implicaciones morales o sociales significativas.

Impacto social de la IA: Grech y Nax (2018) analizan cómo la inteligencia artificial afecta a la sociedad, tanto de manera positiva como negativa. Este concepto incluye la evaluación de los beneficios de la IA, como la mejora de la eficiencia, la innovación y el acceso a nuevas oportunidades; así como los riesgos, la pérdida de empleos, la exacerbación de desigualdades y las amenazas a la privacidad y a la seguridad. Estudiar el impacto social es esencial para desarrollar políticas y prácticas que maximicen los beneficios de la IA mientras se mitigan sus posibles efectos adversos.

2.3. Revisión de estudios previos

A lo largo de la esta sección se analizarán los estudios relevantes que han investigado la implementación de los *Dictator Game* y el comportamiento altruista utilizando agentes de inteligencia artificial generativa, en distintos escenarios y con distintos enfoques. Estos estudios permiten comprender profundamente cómo se manifiesta el altruismo en contextos experimentales, tanto en seres humanos como en modelos de inteligencia artificial generativa.

A continuación, se detallan las distintas temáticas y campos asociados a temas relevantes para la investigación.

2.3.1. Economía experimental

Dentro de la economía existe una disciplina llamada *economía experimental*, que se enfoca en estudiar experimentalmente los fenómenos de conducta (V. L. Smith, 2008). A lo largo de los años, autores fundamentales como Vernon L. Smith y Daniel Kahneman han sido pioneros en este campo, contribuyendo significativamente al entendimiento sobre la manera en la que las personas toman decisiones en contextos económicos.

Las investigaciones en esta área han continuado explorando aspectos clave del comportamiento humano, dentro de los cuales el altruismo se posiciona como uno de los temas más relevantes en los últimos años. Este fenómeno, que se refiere a la preocupación por el bienestar de los demás, ha sido estudiado utilizando diversos enfoques, incluyendo la teoría de juegos (Batson, 2010). Un ejemplo destacado es el Dictator Game, un experimento en el que un participante (el dictador) recibe una suma de dinero o unidades y debe decidir cuánto compartir con otro participante. Este experimento ha permitido obtener información valiosa sobre la disposición de las personas a actuar de manera altruista, incluso en ausencia de incentivos directos. (Daylamani-Zad y Angelides, 2021; Grech y Nax, 2018).

Además, en años recientes se ha comenzado a aplicar estas metodologías para estudiar el comportamiento de agentes inteligentes, como sistemas de inteligencia artificial, en su interacción con seres humanos.

2.3.2. Altruismo

El altruismo es un comportamiento que beneficia a otros a expensas del propio bienestar del individuo. En el contexto de la IA, se refiere a la programación de agentes para que tomen decisiones que favorezcan el bien común o los intereses de otros agentes o humanos (Grech y Nax, 2018; T. Johnson y Obradovich, 2023).

Altruism in Humans por Batson, (2010)

> Este libro explora científicamente la capacidad humana de actuar de manera altruista, es decir, preocuparse por el bienestar de otros en lugar de actuar por motivos egoístas. Basado en 35 años de experimentos de laboratorio, el libro detalla la hipótesis de la empatía-altruismo, que sostiene que los sentimientos de empatía hacia una persona en necesidad generan una motivación para eliminar esa necesidad, haciendo la teoría empíricamente verificable.

> La investigación respalda consistentemente esta hipótesis, sugiriendo que el altruismo inducido por la empatía es parte del comportamiento humano. El libro también aborda las implicaciones de esta conclusión, sugiriendo que este tipo de altruismo es una fuerza más poderosa en la sociedad de lo que se ha reconocido, y que no apreciarlo ha limitado la comprensión de nuestras acciones y la promoción de una sociedad más compasiva. (Batson, 2010)

El libro ofrece un análisis científico sobre la capacidad humana hacia el altruismo, a través de una serie de experimentos de laboratorio realizados durante más de 35 años. A través de ellos, Batson desarrolla y prueba la hipótesis de la empatía-altruismo, que sostienen que los sentimientos de empatía y compasión hacia una persona en necesidad generan una motivación para ayudarla.

El libro se divide en tres partes: la teoría de la motivación altruista, la investigación sobre la hipótesis y las implicaciones teóricas y prácticas. lo largo de su lectura, el libro argumenta que el altruismo, motivado por la empatía, es una parte integral de la naturaleza humana y que, por ende, tiene un impacto significativo en nuestras interacciones y en la sociedad en general.

2.3.3. Teoría de juegos: *Dictator Game*

Los *Dictator Game* dictador son experimentos económicos diseñados para estudiar el comportamiento altruista. En estos juegos, un jugador (el «dictador») decide cómo dividir una suma de dinero entre ellos mismos y otro jugador, sin ninguna restricción (Daylamani-Zad y Angelides, 2021; Grech y Nax, 2018).

Para este estudio se utilizan los *Dictator Game*, que proporcionan un marco de estudio para analizar cómo las preferencias y los comportamientos altruistas tienden a ser modelados por agentes de Inteligencia Artificial Generativa. Como parte de los estudios base de esta temática, se trabaja el artículo *Dictator games: a meta study* del autor Engel (2010), que se detalla a continuación.

Dictator games: a meta study por Engel, (2010)

> En los últimos 25 años se han publicado más de cien experimentos de *Dictator Game*. Este metaestudio resume las pruebas. Aprovechando el hecho de que la mayoría de los experimentos tenían que fijar parámetros que no pretendían probar, en regresión múltiple el metaestudio es capaz de evaluar el efecto de manipulaciones individuales, controlando una serie de factores explicativos alternativos.
>
> El rico conjunto de datos resultante también proporciona un banco de pruebas para comparar especificaciones alternativas del modelo estadístico para analizar los datos del *Dictator Game*. Muestra cómo los modelos Tobit (que suponen que los dictadores querrían incluso aceptar dinero) y los modelos de obstáculos (que suponen que la decisión de dar una cantidad positiva es independiente de la elección de la cantidad, condicionada a dar) aportan información adicional. (Engel, 2010)

El objetivo principal del estudio era realizar un metaanálisis exhaustivo de los experimentos del *Dictator Game*, sintetizando los datos de 129 estudios realizados entre 1992 y 2009, que recaban 41,433 observaciones. Este análisis busca cuantificar sistemáticamente la disposición a la benevolencia de los individuos en diferentes condiciones experimentales y poblacionales.

Entre los resultados principales se encontró que, en promedio, los dictadores donan el 28.35% del «pie». La distribución de estas donaciones está sesgada hacia la izquierda, lo que indica que la mayoría dona cantidades pequeñas. Además, se observó una variación significativa entre los estudios, lo cual explica el 97.1% de la variación total. Esto sugiere la necesidad de análisis de regresión múltiple para comprender mejor las variaciones.

Por otro lado, la distribución de las contribuciones reveló que el 36.11% de los participantes no dona nada, el 16.74% dona la mitad y el 5.44% dona todo. Asimismo, se identificaron efectos robustos de variables independientes como el tipo de juego (un solo juego versus repetidos), la población estudiantil frente a otras poblaciones y la distancia social especificada. Gracias a todo lo anterior, este metaanálisis ofrece una visión integral de la disposición a la benevolencia en el *Dictator Game*, destacando la variabilidad y los factores que influyen en las decisiones de donación.

2.3.4. Inteligencia artificial generativa, altruismo y *Dictator Game*

Basándose en los conceptos anteriores y distintos estudios previos, se recopilan los siguientes estudios que recapitulan las convergencias de los conceptos.

Evidence of behavior consistent with self-interest and altruism in an artificially intelligent agent por T. Johnson y Obradovich, (2023)

> El estudio investiga el comportamiento de agentes de inteligencia artificial (IA) en tareas de toma de decisiones que reflejan tanto el interés propio como el altruismo. Se llevaron a cabo experimentos en los que se evaluó cómo los agentes de IA maximizaron sus beneficios en decisiones no sociales y cómo compartieron recursos en un *Dictator Game*. Los resultados mostraron que los agentes de IA variaron en su capacidad para maximizar beneficios y en su comportamiento altruista. Este trabajo proporciona una comprensión más profunda de cómo los modelos de IA exhiben comportamientos que se asemejan a los humanos en contextos sociales y económicos. (T. Johnson y Obradovich, 2023)

El estudio realizado por T. Johnson y Obradovich (2023) explica la metodología e implementación de experimentos utilizando *Dictator Games* para la medición de decisiones altruistas en agentes de inteligencia artificial generativa desarrollados por OpenAI.

Además, este estudio contaba con los siguientes objetivos:

- **Examinar el comportamiento altruista:** investigar si los agentes de IAG exhiben comportamientos altruistas al aceptar costos personales para beneficiar a otros en un contexto de *Dictator Game*.

- **Evaluar la maximización de beneficios:** determinar si los agentes de IAG maximizan sus propios beneficios en tareas de decisión no sociales, lo que serviría como base para inferir si sus decisiones de compartir recursos en el *Dictator Game* son realmente altruistas o simplemente aleatorias.
- **Comparar comportamientos entre diferentes agentes**: analizar cómo varían los comportamientos altruistas de los agentes de IAG en función del destinatario, con el fin de averiguar si existe un sesgo hacia el altruismo entre agentes de IAG en comparación con humanos.
- **Desarrollar métodos de evaluación**: proporcionar un nuevo método para rastrear el desarrollo de comportamientos de interés propio y altruismo en futuros agentes de IAG, contribuyendo así a la investigación sobre la inteligencia social y la organización compleja entre máquinas.

Altruism and Selfishness in Believable Game Agents: Deep Reinforcement Learning in Modified Dictator Games por Daylamani-Zad y Angelides, (2021)

> El estudio investiga la creación de agentes de juego que exhiben diferentes niveles de altruismo y egoísmo utilizando aprendizaje por refuerzo profundo en juegos modificados de dictador. Los resultados sugieren que la representación de comportamientos humanos en los agentes influye en la experiencia del jugador, lo que tiene implicaciones para el diseño de personajes de juego más realistas y atractivos. (Daylamani-Zad y Angelides, 2021)

El estudio investiga la creación de agentes de juego creíbles utilizando aprendizaje por refuerzo profundo (DRL), específicamente a través de la optimización de políticas proximales (PPO). A través de un *Dictator Game* modificado, se examina cómo el altruismo y el egoísmo influyen en la percepción de la credibilidad de estos agentes.

A partir de la investigación realizada, se concluye que la credibilidad de los agentes no depende solo de la observación de su comportamiento, sino también de cómo es percibido, sugiriendo que el altruismo y el egoísmo son dimensiones clave para crear agentes más realistas y envolventes en videojuegos y otras aplicaciones de inteligencia artificial.

A continuación, se describen los principales objetivos de la investigación:

- **Investigar el impacto del altruismo y el egoísmo en la credibilidad de los agentes de juego**: el estudio busca entender cómo diferentes comportamientos de los agentes, específicamente en términos de altruismo y egoísmo, afectan la percepción de los jugadores sobre la credibilidad de estos agentes.
- **Desarrollar agentes que exhiban comportamientos humanos realistas**: utilizando aprendizaje por refuerzo profundo, el objetivo es crear agentes que aprendan y se adapten para mostrar diferentes niveles de altruismo, lo que contribuiría a una experiencia de juego más inmersiva y agradable.

- **Establecer la importancia de la percepción del comportamiento humano en la definición de la credibilidad**: el estudio pretende demostrar que la credibilidad de los agentes no solo debe basarse en la observación del comportamiento humano, sino también en cómo los jugadores perciben ese comportamiento, sugiriendo que la percepción es un factor clave en la evaluación de la inteligencia artificial en juegos.

Playing Games With GPT: What Can We Learn About a Large Language Model From Canonical Strategic Games? *por* Brookins y Debacker, (2024)

> El estudio investiga cómo los modelos de lenguaje grande (LLM), específicamente el modelo GPT de OpenAI, toman decisiones en juegos económicos clásicos, como el *Dictator Game* y el dilema del prisionero. A través de experimentos, se analiza si el LLM muestra preferencias humanas por la equidad y la cooperación. Los resultados indican que el GPT no solo replica comportamientos humanos en estos contextos, sino que también tiene la posibilidad de superar a los humanos en términos de racionalidad y decisiones que favorecen la eficiencia. Este trabajo contribuye a la comprensión de las capacidades de toma de decisiones de la IA en entornos estratégicos y sus implicaciones éticas (Brookins y Debacker, 2024)

El estudio tuvo como objetivo principal explorar las preferencias de equidad y cooperación en modelos de lenguaje grande (LLM), específicamente el GPT-3.5, al participar en juegos económicos clásicos, como el *Dictator Game* y el dilema del prisionero.

La metodología del estudio involucró la realización de simulaciones en dos juegos: el *Dictator Game* y el dilema del prisionero. En el *Dictator Game*, un jugador (el asignador) recibe una cantidad de dinero y debe decidir cuánto compartir con un jugador pasivo (el receptor). Para evitar sesgos en la toma de decisiones, se proporcionaron instrucciones neutrales al LLM. Al finalizar, las decisiones del LLM se compararon con resultados de estudios experimentales previos sobre comportamiento humano en estos juegos.

Los resultados mostraron que, en el *Dictator Game* el LLM tendía fuertemente hacia la equidad, optando con frecuencia por dividir el recurso de manera equitativa (50-50), lo que contrasta con la predicción racional de que el asignador debería retener todo. Estos hallazgos sugieren que el LLM no solo imita el comportamiento humano, sino que también exhibe preferencias por la equidad superiores a las de los participantes humanos, lo que tiene importantes implicaciones para la comprensión de la ética y la racionalidad en la inteligencia artificial. En resumen, el estudio revela que el GPT-3.5 es capaz de reflejar comportamientos humanos en juegos económicos, mostrando una notable preferencia por la cooperación y la equidad.

Can Large Language Models Serve as Rational Players in Game Theory? A Systematic Analysis por Fan et al., (2023)

> El documento presenta un análisis sistemático de los Modelos de Lenguaje Grande (LLMs) en el contexto de la teoría de juegos. A través de experimentos con juegos clásicos como el *Dictator Game*, Piedra-Papel-Tijera y el juego de red en anillo, se evalúa la capacidad de los LLMs para actuar como jugadores racionales. Los hallazgos indican que, aunque los LLMs son capaces construir deseos claros y mostrar un rendimiento humano en ciertos patrones, tienen dificultades para refinar creencias y seguir comportamientos humanos en el proceso del juego. El estudio concluye que es urgente analizar las limitaciones de los LLMs en diversas áreas y sugiere que futuras investigaciones podrían enfocarse en mejorar sus capacidades específicas en la teoría de juegos. (Fan et al., 2023)

El estudio tiene como objetivo principal examinar la capacidad de los modelos de lenguaje grande (LLM) para comportarse como jugadores racionales dentro del ámbito de la teoría de juegos, enfocándose especialmente en el *Dictator Game*.

La investigación se planteó varios objetivos clave: en primer lugar, evaluar la capacidad de los LLM para formar deseos claros, refinar creencias sobre la incertidumbre y tomar decisiones óptimas en juegos; en segundo lugar, identificar las limitaciones de los LLMs en comparación con el comportamiento humano en situaciones de juego; y, finalmente, explorar el potencial uso de los LLM como sustitutos de los humanos en experimentos de teoría de juegos para la investigación en ciencias sociales.

La metodología del estudio seleccionó tres juegos clásicos para el análisis: el *Juego del Dictador*, Piedra-papel-tijeras y el juego de *Red en Anillo*. Seguidamente, se realizaron experimentos para evaluar el desempeño de los LLM (incluyendo GPT-4) en estos juegos, centrándose en su capacidad para establecer preferencias y tomar decisiones basadas en diferentes escenarios de juego.

En el *Dictator Game*, se analizó cómo los LLM reflejan preferencias humanas al elegir entre distintas opciones de asignación de recursos. Los resultados mostraron que en dicho juego los LLM podían identificar y seguir preferencias, pero encontraron dificultades al enfrentarse a preferencias poco comunes. Esto sugiere que, aunque forman deseos claros, su rendimiento puede ser inconsistente cuando se les presentan opciones que no siguen patrones comunes.

Además, incluso el LLM más avanzado (GPT-4) presentó diferencias significativas en comparación con los humanos, especialmente en la capacidad de refinar creencias y tomar decisiones óptimas. Por esta razón, se concluyó que los LLM tienden a modificar o ignorar creencias refinadas durante el proceso de toma de decisiones, lo que limita su efectividad como jugadores racionales en la teoría de juegos.

GPT-4 Technical Report por (OpenAI et al., 2023)

> El informe técnico de GPT-4 presenta un modelo de inteligencia artificial de gran escala y multimodal que puede procesar tanto texto como imágenes para generar salidas textuales. Aunque no alcanza las capacidades humanas en todos los escenarios, GPT-4 muestra un rendimiento impresionante en diversas pruebas académicas y profesionales, incluyendo una notable puntuación en un examen simulado de abogacía. El informe también destaca las innovaciones en la infraestructura y los métodos de optimización que contribuyeron a su éxito. En resumen, GPT-4 representa un avance significativo en el campo de la inteligencia artificial, con aplicaciones potenciales en múltiples dominios. (OpenAI et al., 2023)

El estudio proporciona una visión integral del modelo GPT4 de sus capacidades avanzadas y sus desafíos, lo que es fundamental para su aplicación en contextos académicos y profesionales. El modelo GPT-4 demuestra un rendimiento comparable al de los humanos en diversas evaluaciones académicas y profesionales. Por ejemplo, en un examen simulado de abogacía, GPT-4 alcanza una puntuación que lo sitúa en el 10% superior de los examinados, lo que representa una mejora significativa en comparación con su predecesor, GPT-3.5, que se ubicó en el 10% inferior.

Además de su desempeño en exámenes específicos, GPT-4 supera a los modelos de lenguaje existentes y a la mayoría de los sistemas de última generación en una variedad de benchmarks tradicionales de procesamiento del lenguaje natural (NLP). Muestra capacidades destacadas no solo en inglés, sino también en múltiples otros idiomas, superando el estado del arte en inglés en 24 de las 26 lenguas evaluadas.

Sin embargo, a pesar de estas capacidades impresionantes, GPT-4 no es completamente fiable y puede "alucinar" hechos o cometer errores de razonamiento. Esto indica que, aunque su rendimiento es notable en muchos contextos, no siempre iguala la precisión y el razonamiento humano. Por lo tanto, es fundamental considerar estas limitaciones al aplicar GPT-4 en situaciones que requieren un alto grado de fiabilidad y exactitud.

GPT-4 System Card. Por (OpenAI, 2023)

> Este artículo ofrece un análisis integral de GPT-4, la última iteración en la familia de modelos de lenguaje GPT, centrándose en sus capacidades, limitaciones y riesgos asociados en contextos de toma de decisiones. A través de evaluaciones cualitativas y cuantitativas, el estudio identifica desafíos clave, como los sesgos inherentes, la tendencia a generar alucinaciones y la falta de una comprensión profunda del contexto, los cuales pueden afectar negativamente la calidad de las decisiones tomadas por el modelo. Los hallazgos destacan la susceptibilidad del modelo a producir contenido dañino y desinformación, lo que genera preocupaciones sobre su posible mal uso en diversas aplicaciones. Se analizan las estrategias de mitigación y las políticas de uso implementadas por OpenAI,

> haciendo hincapié en la importancia de un despliegue responsable en áreas sensibles. En general, el artículo subraya la necesidad de considerar cuidadosamente las limitaciones y riesgos de GPT-4 para garantizar un uso ético y eficaz en procesos de toma de decisiones. (OpenAI, 2023)

El estudio tiene como objetivo principal examinar las capacidades, limitaciones y riesgos de GPT-4 en la toma de decisiones. El modelo GPT-4 destaca por su capacidad para realizar tareas complejas relacionadas con la generación de texto, el razonamiento y la combinación de información diversa. Estas habilidades lo convierten en una herramienta valiosa para diversos contextos, aunque su implementación plantea desafíos significativos en términos de calidad y ética en la toma de decisiones. A través de pruebas cualitativas y cuantitativas, se ha evaluado su desempeño, identificando fortalezas, pero también limitaciones.

Entre las principales limitaciones se encuentran la tendencia del modelo a reproducir sesgos presentes en los datos de entrenamiento y generar información incorrecta, fenómeno conocido como alucinaciones. (OpenAI et al., 2023) Asimismo, GPT-4 carece de una comprensión profunda del contexto, lo que puede llevar a interpretaciones inadecuadas de situaciones complejas. Además, su dependencia de las entradas proporcionadas implica que respuestas sesgadas o erróneas surgan si la calidad de los datos iniciales es deficiente.

El uso de GPT-4 también conlleva riesgos, entre ellos la posibilidad de su utilización para fines malintencionados, como la propagación de desinformación, la manipulación de información o la explotación en ingeniería social. Estos riesgos son especialmente preocupantes en contextos sensibles como la salud y la justicia penal, donde las decisiones incorrectas pueden tener consecuencias graves.

Para mitigar estos riesgos, OpenAI ha implementado políticas estrictas que restringen el uso del modelo en sectores de alto impacto y mecanismos de seguridad para evitar que se emplee con fines contrarios a la ética. Aunque GPT-4 representa un avance considerable en inteligencia artificial, su uso requiere un enfoque crítico y responsable, especialmente en áreas donde sus decisiones puedan afectar significativamente a la sociedad

2.4. Conclusión del análisis literario

Como parte del análisis literario realizado, se detalla a continuación, en la Tabla 4 Tabla análisis estudios, se muestran los puntos relevantes de cada estudio seleccionado, los cuales permiten centrar las bases del presente proyecto.

Tabla 4 Tabla análisis estudios

Estudio	Resumen	Objetivos	Metodología	Principales descubrimientos
Playing Games With GPT: What Can We Learn About a Large Language Model from Canonical Strategic Games? por Brookins y Debacker, (2024)	Se utilizó un gran modelo lingüístico (GPT-3.5) para jugar a dos juegos estratégicos clásicos: el *Dictator Game* y el dilema del prisionero, y compararon las decisiones del modelo con las de los humanos.	Comprender las preferencias fundamentales sobre equidad y cooperación integradas en la inteligencia artificial. Hacer que un LLM, GPT-3.5, juegue al *Dictator Game* y al dilema del prisionero. Comparar las decisiones del LLM con las de los humanos en experimentos de laboratorio.	Los investigadores hicieron que un gran modelo lingüístico (GPT-3.5) jugara a dos juegos clásicos: el *Dictator Game* y el dilema del prisionero. Compararon las decisiones tomadas por el LLM con las tomadas por humanos en experimentos de laboratorio.	El modelo de lenguaje amplio GPT-3.5 mostró una tendencia hacia la imparcialidad en el *Dictator Game*, incluso mayor que los participantes humanos. Los resultados del estudio ayudan a comprender las preferencias éticas y racionales de la inteligencia artificial.
Evidence of behavior consistent with self-interest and altruism in an artificially intelligent agent por T. Johnson y Obradovich, (2023)	Este artículo que presenta un experimento que pone a prueba el comportamiento altruista de los agentes de IA, descubriendo que el más sofisticado mostraba tanto comportamientos egoístas como altruistas, variando los comportamientos altruistas en función del receptor.	Comprobar el comportamiento altruista de los agentes de IA. Examinar si los agentes de IA maximizan sus propios beneficios. Colocar agentes de IA en *Dictator Game* para ver cómo comparten recursos con diferentes destinatarios.	Una tarea de decisión no social en la que los agentes seleccionaban su propia retribución de un rango dado. Una serie de *Dictator Game* en los que los agentes podían compartir recursos con otro agente de IA, un experimentador humano o una organización benéfica anónima.	Este mismo agente mostró el comportamiento altruista más generoso en *Dictator Game*, similar al de los humanos que comparten con otros humanos. El agente de la IA compartió una parte sustancialmente menor de la dotación con el experimentador humano o con una organización benéfica anónima que con otros agentes de la IA.

Estudio	Resumen	Objetivos	Metodología	Principales descubrimientos
Altruism and Selfishness in Believable Game Agents: Deep Reinforcement Learning in Modified Dictator Games por Daylamani-Zad y Angelides, (2021)	En este trabajo, los autores utilizaron el aprendizaje profundo por refuerzo para entrenar agentes de juego en un *Dictator Game* modificado para investigar el efecto del egoísmo y el altruismo en la credibilidad de los agentes.	Utilizar el aprendizaje por refuerzo profundo para entrenar agentes de juego con distintos niveles de egoísmo y altruismo (egoístas, sustitutos perfectos y Leontief). Evaluar la credibilidad de estos perfiles de agentes utilizando métricas de credibilidad de agentes.	Investigar el efecto del egoísmo y el altruismo en la credibilidad de los agentes del juego. Diseño y aplicación de un entorno de entrenamiento con funciones de recompensa basadas en investigaciones empíricas. Definición de tres perfiles de agentes (egoístas, sustitutos perfectos y Leontief) que representan distintos niveles de egoísmo/altruismo.	El comportamiento altruista de los agentes del juego es percibido como más creíble por los jugadores en comparación con el comportamiento egoísta. El comportamiento similar al humano de los agentes de IA se basa más en el comportamiento humano percibido que en el observado.
Can Large Language Models Serve as Rational Players in Game Theory? A Systematic Analysis por Fan et al., (2023)	El resumen de este trabajo es que analiza sistemáticamente la capacidad de los grandes modelos lingüísticos (LLM) para actuar como jugadores racionales en experimentos de teoría de juegos, descubriendo que incluso el LLM más avanzado (GPT-4) muestra disparidades sustanciales en comparación con los humanos en aspectos clave de la racionalidad de la teoría de juegos.	Los objetivos del estudio son analizar sistemáticamente las capacidades de los modelos de lenguaje amplio (LLM) en el contexto de la teoría de juegos, evaluando específicamente hasta qué punto los LLM pueden alcanzar la racionalidad en tres aspectos clave: construir un deseo claro, refinar la creencia sobre la incertidumbre y tomar acciones óptimas, utilizando tres juegos clásicos como casos de prueba.	Evaluar la racionalidad de los grandes modelos lingüísticos (LLM) en el contexto de la teoría de juegos. Se utilizan tres experimentos clásicos de teoría de juegos para evaluar la racionalidad de los LLM. Comparación del rendimiento del LLM más avanzado, GPT-4, con el rendimiento humano en estos experimentos de teoría de juegos.	A los LLM les cuesta construir deseos basados en preferencias poco comunes Los LLM no logran refinar creencias a partir de patrones simples Los LLM pueden pasar por alto o modificar las creencias refinadas al emprender acciones

Estudio	Resumen	Objetivos	Metodología	Principales descubrimientos
Dictator games: a meta study Por Engel, (2010)	El artículo presenta un metaanálisis de más de 100 experimentos de *Dictator Game* publicados en los últimos 25 años, en los cuales se exploran los factores que influyen en la medida que están dispuestos a dar los dictadores a receptores anónimos.	Resumir las pruebas de más de 100 experimentos de *Dictator Game* publicados en los últimos 25 años. Evaluar el efecto de manipulaciones individuales en *Dictator Game*, controlando otros factores explicativos. Proporcionar un banco de pruebas para comparar modelos estadísticos alternativos para analizar datos de *Dictator Game*.	Modelos de meta regresión de efectos fijos y aleatorios para analizar los datos de 129 experimentos de *Dictator Game*. Reconstrucción de los datos originales a nivel individual de 83 de los artículos, lo que da como resultado un conjunto de datos de 20.813 observaciones. Reconocimiento de que la terminología utilizada para el modelo de efectos fijos difiere de la terminología econométrica estándar.	El hallazgo original de que las personas no siempre maximizan sus propios ingresos se ha repetido sistemáticamente en muchos experimentos del *Dictator Game* a lo largo de los últimos 25 años. El Dictator Game es una herramienta útil para explorar los factores que influyen en el comportamiento social humano y en la heterogeneidad, en lugar de limitarse a probar la hipótesis de la maximización de los ingresos.
GPT-4 System Card. **Por** (OpenAI, 2023)	El estudio combina evaluaciones cualitativas y cuantitativas para analizar el comportamiento del modelo en distintos contextos. Los resultados permitieron identificar patrones de sesgo, alucinaciones y otras problemáticas asociadas con la toma de decisiones.	Evaluar las capacidades y limitaciones de GPT-4 en la generación de texto y la toma de decisiones. Identificar y analizar los riesgos asociados con el uso del modelo en diversas aplicaciones. Proporcionar recomendaciones sobre mitigaciones y políticas de uso para garantizar un despliegue ético y seguro del modelo.	El estudio utilizó evaluaciones cualitativas y cuantitativas, con más de 50 expertos en inteligencia artificial, para identificar sesgos, alucinaciones y riesgos asociados, como la generación de contenido dañino. Los hallazgos impulsaron mejoras en el modelo y en las estrategias de mitigación mediante un proceso de retroalimentación continua.	El estudio reveló que GPT-4 enfrenta desafíos significativos en su desempeño, incluyendo la reproducción de sesgos presentes en los datos de entrenamiento y la generación de información incorrecta, conocidas como alucinaciones. Además, su limitada comprensión contextual puede llevar a decisiones inadecuadas, mientras que su dependencia de la calidad de las entradas aumenta la probabilidad de respuestas sesgadas o ineficaces.

Estudio	Resumen	Objetivos	Metodología	Principales descubrimientos
GPT-4 Technical Report por (OpenAI et al., 2023)	El estudio sobre GPT-4 presenta un modelo de inteligencia artificial de gran escala y multimodal que puede procesar tanto texto como imágenes, generando salidas textuales. A pesar de no alcanzar las capacidades humanas en todos los escenarios,	Evaluar el rendimiento de GPT-4 en comparación con modelos anteriores y con humanos en diversas pruebas académicas y profesionales. Analizar las capacidades del modelo en múltiples idiomas y su efectividad en tareas de procesamiento de texto e imágenes. Identificar las limitaciones del modelo, incluyendo la tendencia a "alucinar" hechos y cometer errores de razonamiento.	La metodología del estudio combinó diversas evaluaciones para analizar el rendimiento de GPT-4. El modelo fue sometido a exámenes académicos y profesionales, como un examen simulado de abogacía, utilizando condiciones y criterios de puntuación reales. Además, se realizaron pruebas en una serie de benchmarks tradicionales de procesamiento de lenguaje natural (NLP) para comparar el desempeño de GPT-4 con modelos anteriores y sistemas de última generación.	Los hallazgos del estudio indican que GPT-4 logró un rendimiento comparable al humano, superando significativamente a GPT-3.5 en un examen simulado de abogacía. Además, mostró un rendimiento superior al estado del arte en inglés en 24 de los 26 idiomas evaluados, destacando su capacidad multilingüe. Sin embargo, persisten limitaciones en cuanto a la generación de información incorrecta y errores de razonamiento, lo que sugiere la necesidad de un uso cauteloso en aplicaciones críticas.

Lo que distingue a esta investigación es el uso del modelo GPT-4 y sus variantes, ya que los estudios previos descritos en la Tabla 4 Tabla análisis estudios se han limitado mayormente a la aplicación de pruebas en modelos GPT-3.5 y versiones anteriores. La inclusión de GPT-4 permite una exploración más avanzada de las capacidades de los agentes de IAG, abriendo nuevas posibilidades para el análisis de comportamientos altruistas y estratégicos.

De igual manera, mantener actualizadas este tipo de investigaciones es crucial para asegurar la trazabilidad a lo largo del tiempo, ya que el uso de modelos avanzados, como gpt-4, no solo mejora la precisión de los resultados, sino que también permite realizar comparaciones más consistentes con estudios futuros. Esto garantiza que las conclusiones sigan siendo relevantes y válidas en un campo donde las tecnologías y capacidades evolucionan constantemente.

3. Marco Metodológico

El marco metodológico de este estudio tiene como objetivo proporcionar una estructura clara y sistemática que guíe el proceso de investigación desde la conceptualización hasta la recopilación y el análisis de datos. Se describirá detalladamente el tipo, enfoque y alcance de la investigación utilizada y se identificarán sus limitaciones y dimensiones. Además, se describirá el diseño de la investigación y las fuentes de información, haciendo énfasis en las fuentes primarias y secundarias.

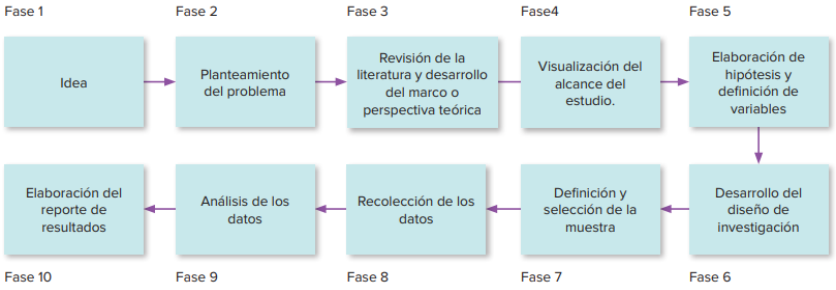
También se introducirán hipótesis y variables o categorías de investigación, así como métodos y herramientas de recolección de datos y matrices de cobertura de variables. Finalmente, se discutirán los procedimientos metodológicos del estudio y se incluirá un cuadro resumen y diagrama de los procedimientos metodológico.

3.1. Tipo de investigación

Según se menciona en el libro *Metodología de la investigación. Las rutas cuantitativa, cualitativa y mixta* (Hernández-Sampieri R. y Mendoza, 2018), existen tres grandes grupos de tipos de investigación:

- Cuantitativa: «Representa un conjunto de procesos organizado de manera secuencial para comprobar ciertas suposiciones. Cada fase precede a la siguiente y no podemos eludir pasos, el orden es riguroso, aunque desde luego, podemos redefinir alguna etapa» (Hernández-Sampieri y R. Mendoza, 2018, pp. 5–6). Este proceso se representa en la Figura 3 Fases de una investigación cuantitativa.

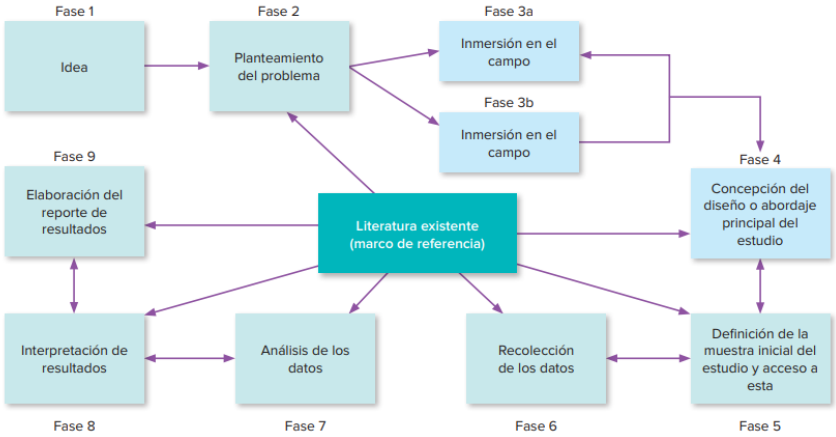
Figura 3 Fases de una investigación cuantitativa.



Nota. Tomado de Hernández-Sampieri y R. Mendoza, (2018, p. 6)

- Cualitativa: se estudian fenómenos de manera sistemática. Sin embargo, en lugar de comenzar con una teoría y luego «voltear» al mundo empírico para confirmar si esta es apoyada por los datos y resultados, el investigador comienza el proceso examinando los hechos en sí y revisado los estudios previos, ambas acciones de manera simultánea, a fin de generar una teoría que sea consistente con lo que está observando que ocurre" (Hernández-Sampieri y R. Mendoza, 2018, p. 7). Se observa el proceso descrito anteriormente en la Figura 4 Fases de una investigación cualitativa.

Figura 4 Fases de una investigación cualitativa.



Nota. Tomado de Hernández-Sampieri y R. Mendoza, (2018, p. 8)

- Mixta: Los métodos mixtos o híbridos representan un conjunto de procesos sistemáticos, empíricos y críticos de investigación e implican la recolección y el análisis de datos tanto cuantitativos como cualitativos, así como su integración y discusión conjunta, para realizar inferencias producto de toda la información recabada (denominadas *metainferencias*) y lograr un mayor entendimiento del fenómeno bajo estudio. (Hernández-Sampieri y R. Mendoza, 2018, p. 10)

Definidos los tipos de investigación detallados por Hernández-Sampieri y R. Mendoza (2018), se determina que el tipo de investigación de este proyecto será de tipo cuantitativo, ya que en esta modalidad se realiza un desarrollo secuencial para comprobar hipótesis (Ileana Ulate y Elizarda Vargas, 2016, p. 16).

3.2. Alcance de la investigación

Hernández-Sampieri y R. Mendoza (2018, p. 106) definen y clasifican los alcances de la investigación en

> Exploratorio, descriptivo, correlacional y explicativo. No representan clases o tipos de investigación, ni son mutuamente excluyentes, sino que constituyen puntos entrelazados de un continuo de causalidad, y en la práctica, cualquier estudio puede incluir elementos de uno o más de ellos. (p. 106)

Con el fin de profundizar en las definiciones de cada tipo de alcance, se aborda el contenido el libro *Metodología para elaborar una tesis* de Ileana Ulate y& Elizarda Vargas (2016), en el cual se definen los siguientes conceptos:

- Estudios exploratorios: «Se realizan cuando el objetivo consiste en examinar un tema poco estudiado o que no se ha abordado antes» (Ileana Ulate y Elizarda Vargas, 2016, p. 72).
- Estudios descriptivos: «Su objetivo es describir un fenómeno, una situación, un contexto o un evento, es decir especificar las propiedades, las características y los perfiles de personas, grupos comunidades, procesos objetivos o cualquier otro

fenómeno que se someta a un análisis» (Ileana Ulate y Elizarda Vargas, 2016, p. 72-73).

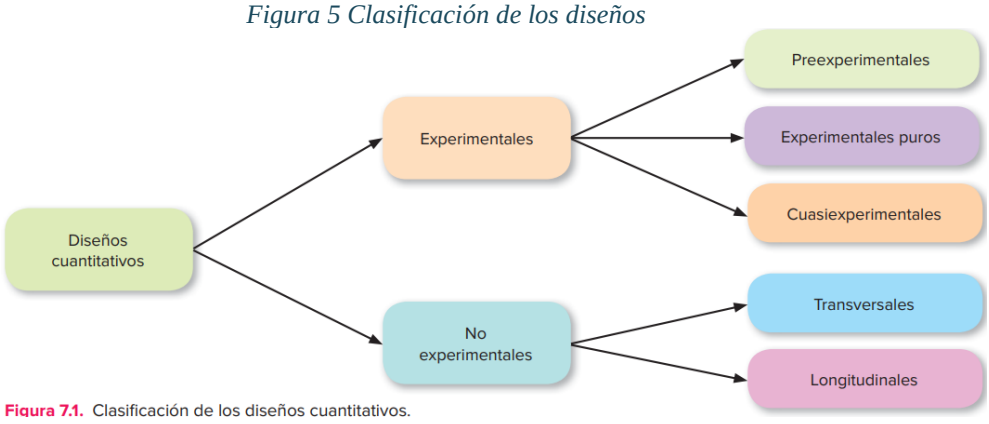
- Estudio correlacional: «Se busca asociar variables para conocer la relación que existe entre dos o más conceptos, categorías o variables, en un contexto de en particular» (Ileana Ulate y Elizarda Vargas, 2016, p. 73).
- Estudio explicativo: «Pretende establecer las causas de los eventos, sucesos o fenómenos que se estudian. Se centra en explicar porque ocurre un fenómeno y en qué condiciones se manifiesta; o bien, por qué se relacionan dos o más variables» (Ileana Ulate y Elizarda Vargas, 2016, p. 73).

El alcance de esta investigación corresponde a un estudio de tipo descriptivo. Aunque no se busca establecer asociaciones causalmente comprobadas, el objetivo es explorar y analizar posibles correlaciones entre las variables, brindando una comprensión más profunda del fenómeno estudiado sin pretender llegar a conclusiones definitivas sobre la relación entre ellas.

3.3. Diseño de investigación

Hernández-Sampieri y R. Mendoza (2018) también brindan distintas opciones de diseño en lo que respecta al tipo de investigación cuantitativa, las cuales se visualizan en el la Figura 5 Clasificación de los diseños.

Figura 5 Clasificación de los diseños



Nota. Tomado de Hernández-Sampieri y R. Mendoza, (2018, p. 613)

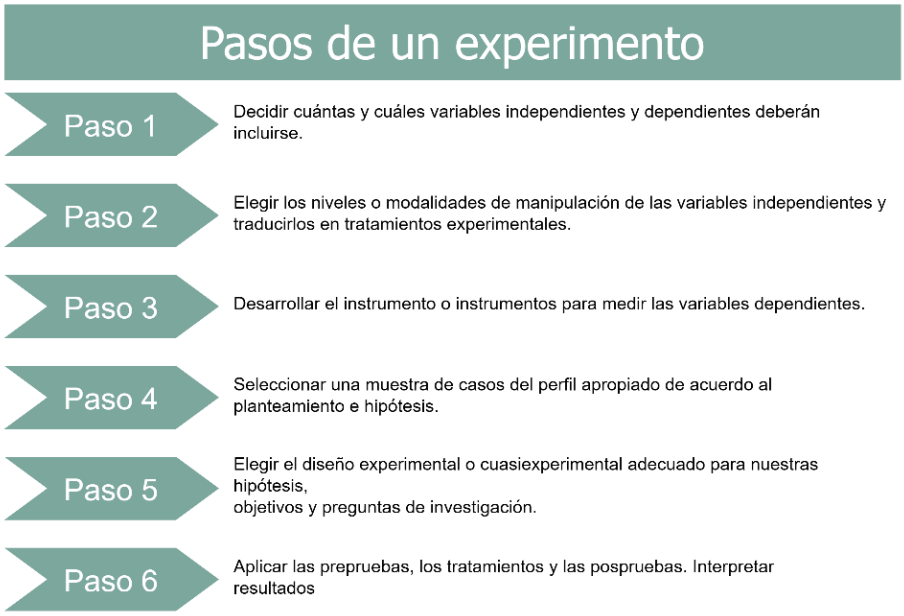
Según la información mostrada en la Figura 5 Clasificación de los diseños, el diseño experimental es el que mejor se adapta a esta investigación, ya que como mencionan Hernández-Sampieri y R. Mendoza, este hace referencia a una investigación en la que se manipulan deliberadamente una o más variables independientes (supuestas causas antecedentes) para analizar las consecuencias que tal manipulación tiene sobre una o más variables dependientes (supuestos efectos consecuentes) dentro de una situación de control para el investigador. Por su parte, los autores Hernández-Sampieri y R. Mendoza (2018) definen *experimento* como una situación de control en la cual se manipulan de manera intencional una o más variables independientes (causas) para analizar las consecuencias de tal manipulación sobre una o más variables dependientes (efectos).

Dentro del diseño experimental se pueden encontrar dos conceptos claves con respecto a los grupos participantes del experimento:

- Grupo experimental: es el que recibe el tratamiento o estímulo experimental. Hernández-Sampieri y R. Mendoza, 2018).
- Grupo de control: no recibe el tratamiento o estímulo experimental. Se le conoce también como *grupo testigo* (Hernández-Sampieri y R. Mendoza, 2018).

Continuando con la referencia a los autores Hernández-Sampieri y R. Mendoza (2018), se establecen los pasos por seguir durante un diseño experimental, lo cuales se pueden visualizar en la Figura 6 Pasos de un experimento. Estos pasos son adaptados a la investigación actual, con el fin de profundizar más en cada uno de los puntos a lo largo del desarrollo del documento.

Figura 6 Pasos de un experimento



Nota. Adaptado de Hernández-Sampieri y R. Mendoza, (2018, p. 173)

3.4. Fuentes de datos e información

La sección de fuentes de datos e información detalla los insumos esenciales para el análisis del estudio, es decir, las fuentes primarias y secundarias.

3.4.1. Fuentes primarias

Según definen Hernández-Sampieri y R. Mendoza, p. (2018) «las referencias o fuentes primarias proporcionan datos de primera mano, pues se trata de documentos que incluyen los resultados de los estudios correspondientes» (p. 72).

A continuación, en la Tabla 5 Fuentes de información primaria se detallan los documentos preliminares para la consulta de información secundaria de la investigación.

Tabla 5 Fuentes de información primarias

Documento	Importancia para la investigación
Evidence of behavior consistent with self-interest and altruism in an artificially intelligent agent Por T. Johnson y Obradovich, (2023)	El artículo explora la presencia de comportamientos altruistas y egoístas en los procesos de toma de decisiones de los agentes de IA. Como fuentes de información, se toman a los agentes de IA OpenAI de diversa complejidad para analizar tanto sus respuestas a los incentivos como su comportamiento en escenarios de reparto de recursos.
Altruism and Selfishness in Believable Game Agents: Deep Reinforcement Learning in Modified Dictator Games Por Daylamani-Zad y Angelides, (2021)	Este documento es esencial para la investigación, ya que explora cómo los agentes de juego creíbles aprenden comportamientos altruistas y egoístas a través del aprendizaje por refuerzo profundo. Este descubrimiento arroja luz sobre la interacción entre la IA y la toma de decisiones éticas en entornos virtuales.
Emergence of Fairness Behavior Driven by Reputation-Based Voluntary Participation in Evolutionary Dictator Games Por Zhang et al., (2024)	La importancia de este documento radica en su análisis sobre la reputación y la participación voluntaria y el impulso que estos brindan al surgimiento de comportamientos justos en juegos dictatoriales evolutivos. Ambos factores contribuyen a la comprensión de los sistemas de IA que desarrollan comportamientos altruistas basados en la reputación en entornos competitivos.
Prediction of People's Abnormal Behaviors Based on Machine Learning Algorithms Por Song, (2022)	En esta investigación se explora cómo los algoritmos de aprendizaje automático predicen comportamientos anómalos en las personas. Además, su aplicación tiene implicaciones importantes en la detección temprana de problemas de salud mental.
A Machine Learning Model for Effective Consumer Behaviour Prediction Por Annshu et al., (2021)	La importancia de este documento radica en su desarrollo de un modelo de aprendizaje automático para predecir el comportamiento del consumidor de manera efectiva, lo que conlleva a tener aplicaciones significativas en marketing, ventas y toma de decisiones empresariales.
Human perceptions of altruism in artificial agents Por Guinn y Palmer, (2014)	Este documento analiza cómo los humanos perciben el altruismo en agentes artificiales. Esto proporciona información valiosa sobre la interacción del ser humano con la inteligencia artificial y el grado de confianza.
Paradigmatic experiments: The Dictator Game Por Guala y Mittone, (2010)	El documento analiza el *Dictator Game* (DG) como una herramienta experimental en la teoría de juegos que destaca por su potencial para ser utilizado en la investigación de normas sociales. La investigación enfatiza la importancia del DG en el estudio de preferencias sociales, altruismo y comportamientos basados en la equidad, y se señala que los economistas aún carecen de una teoría adecuada de normas para aprovechar al máximo el potencial del *Dictator Game.*

3.4.2. Fuentes secundarias

Según define Fidias G. Arias (2012), las fuentes de información secundarias son aquellas que constituyen resúmenes o registros que expanden los datos adquiridos de las fuentes primarias. Las fuentes secundarias comprenden análisis, evaluaciones y otros contenidos que enriquecen la comprensión de las fuentes primarias.

A continuación, en el la Tabla 6 Fuentes de información secundarias se detallan los documentos preliminares para la consulta de información secundaria de la investigación.

Tabla 6 Fuentes de información secundarias

Documento	Importancia para la investigación
Comparing Decision-Making Strategies Between Artificial Intelligence Agents and Humans in Economic Games Por J. Smith, (2019)	Este documento presenta una comparación detallada entre las estrategias de toma de decisiones por parte de agentes de inteligencia artificial y por parte de humanos en juegos económicos. Esta contraposición permite comprender mejor las diferencias y similitudes en su comportamiento.
Understanding Variability in Decision-Making: A Comparative Study of Human and AI Behavior in Social Dilemmas A. et al. Johnson, (2020)	Este documento analiza la variabilidad en la toma de decisiones entre humanos y agentes de inteligencia artificial en dilemas sociales; gracias a esto, ofrece perspectivas valiosas para entender la diferencia en su comportamiento en situaciones específicas.
Analyzing Behavioral Patterns in Human-AI Decision-Making Scenarios: A Literature Review Garcia, (2021)	Este documento ofrece una revisión exhaustiva de los patrones de comportamiento tanto de humanos como de agentes de inteligencia artificial en escenarios de toma de decisiones, lo que permite una comprensión más amplia de las diferencias y similitudes en su comportamiento.
A Comparative Analysis of Decision-Making Processes in Humans and AI Agents: Insights from Behavioral Economics Patel, (2018)	Este documento ofrece una comparación detallada de los procesos de toma de decisiones entre humanos y agentes de inteligencia artificial, destacando las aportaciones de la economía del comportamiento que permiten comprender estas diferencias y similitudes.
Exploring the Role of Emotions in Decision-Making: A Comparative Study of Human and AI Agents Wang, (2022)	Esta investigación explora el papel de las emociones y sentimientos en la toma de decisiones, comparando el comportamiento entre humanos y agentes de inteligencia artificial. De esta manera, ofrece una visión más completa acerca del modo en el que estas influencias afectan sus decisiones.

3.5. Población y selección de muestra

En el contexto de un proyecto enfocado en el análisis y evaluación de modelos de lenguaje, la población se define como el conjunto completo de modelos de API GPT(OpenAI, s. f.n.d.-b). Esta población incluye las variantes de los modelos GPT-4, GPT-4o, GPT-4o mini, GPT-4 Turbo y GPT-3.5 turbo, que, a continuación, se detallarán. cada uno de los modelos:

GPT-4: es un modelo de alta calidad que destaca por su capacidad de razonamiento general, que cuenta con un índice de calidad muy alto. Sin embargo, su velocidad de salida es relativamente baja en comparación con otros modelos, lo que significa que procesa menos tokens por segundo. Este modelo es ideal para tareas que requieren precisión y profundidad, como análisis complejos y generación de texto detallada.

GPT-4 Turbo: este modelo se encuentra muy cerca del GPT-4 en términos de calidad, pero ofrece una mayor velocidad de procesamiento. Es una opción popular para aplicaciones que necesitan un equilibrio entre velocidad y calidad, como la creación de contenido para blogs, redes sociales y *chatbots* de servicio al cliente. Además, incluye capacidades para trabajar con entradas y salidas multimodales, como texto e imágenes, lo que lo hace más versátil.

GPT-4o: esta versión mejora significativamente en términos de velocidad en comparación con el GPT-4, pero mantiene un nivel de calidad muy cercano al del GPT-4 Turbo. Es un modelo multimodal que no solo procesa texto, sino también imágenes y audio, lo que abre posibilidades para su uso en traducción, creación de contenido multimedia y aplicaciones educativas. Además, es más accesible en términos de costo y está diseñado para ser una opción más rápida y económica para usuarios y desarrolladores.

GPT-4o mini: una versión más ligera de GPT-4o, que ofrece un equilibrio entre costo y velocidad, aunque con una calidad ligeramente inferior a la de los modelos principales. Es adecuado para tareas que no requieren tanta precisión o profundidad, pero en las que la velocidad y el costo son factores críticos.

GPT-3.5 Turbo: aunque es uno de los modelos más rápidos en términos de generación de tokens por segundo, su calidad de salida es inferior en comparación con los modelos GPT-4. Es ideal para aplicaciones en las que la velocidad es más importante que la precisión, como respuestas rápidas en chats o generación de contenido básico.

Tomando como punto de partida las descripciones y características de los modelos anteriores, se realiza un análisis de viabilidad sobre cada uno de los modelos. La Tabla 7 Tabla de consumo comparativo por modelo permite estudiar la medición de sus rangos de limite y su consumo de tokens. La instrucción de referencia tiene un valor aproximado a 100 tokens de entrada y salida, los cuales son obtenidos de OpenAI (s.f.n.d.-a).

Tabla 7 Tabla de consumo comparativo por modelo

Modelo	RPM (Tokens por Minuto)	RPD (Tokens por Día)	TPM (Tokens por Mes)	Batch Queue Limit	Ejecuciones por Minuto	Ejecuciones por Día	Ejecuciones por Mes
gpt-4o	500	-	30,000	90,000	297	9,900	297,000
gpt-4o-mini	500	10,000	200,000	2,000,000	1,980	19,800	1,980,000
gpt-4-turbo	500	-	30,000	90,000	297	9,900	297,000
gpt-4	500	10,000	10,000	100,000	99	9,900	99,000
gpt-3.5-turbo	3,500	10,000	200,000	2,000,000	1,980	19,800	1,980,000

Nota. Adaptado de OpenAI (s.f.n.d.-b)

De acuerdo con la descripción anterior y la Tabla 7 Tabla de consumo comparativo por modelo, se define que la selección de la muestran los modelos GPT-3.5-turbo como grupo de control y las versiones de los modelos GPT-4 grupo experimental.

3.6. Sujetos de investigación

Dado que la investigación se basa en la ejecución y análisis de resultados de experimentos de tipo *Dictator Game* para medir el altruismo, se elimina la necesidad de colaboración de sujetos externos con el equipo de trabajo. Bajo este contexto, se estudia el uso de la inteligencia artificial generativa como una herramienta para el campo de investigación del comportamiento del altruismo, sin involucrar directamente sujetos de investigación en el estudio.

Los agentes de IAG se emplearán en los experimentos como simulaciones de sujetos ya que ofrecen la ventaja de controlar variables y condiciones experimentales de manera precisa y sistemática. Esto permite explorar cómo las decisiones altruistas tienden a ser influenciadas por diferentes factores y escenarios, sin la necesidad de la participación directa de sujetos humanos.

Además, el uso de IAG en lugar de sujetos humanos ofrece una mayor flexibilidad y escalabilidad en la ejecución de experimentos, lo que facilita la realización de un análisis exhaustivo y detallado del comportamiento altruista en diferentes contextos y condiciones.

3.7. Hipótesis

En esta sección se plantean las hipótesis y suposiciones tentativas que se someterán a prueba durante el curso de la investigación. Estas hipótesis están diseñadas para guiar el proceso de recolección y análisis de datos.

Según los autores Hernández-Sampieri y R. Mendoza (2018)

> Las hipótesis son explicaciones tentativas del fenómeno o problema investigado formuladas como proposiciones o afirmaciones y constituyen las guías de un estudio. Indican lo que tratamos de probar y, por así decirlo, toman la estafeta de parte del

planteamiento del problema para determinar el curso de la indagación en la ruta cuantitativa. (p. 124)

Dada la naturaleza compleja de los agentes de AIG y su capacidad para aprender y adaptarse a diferentes entornos, se plantea la siguiente hipótesis:

- Existe variación en las capacidades de los agentes de IAG para tomar decisiones altruistas en la aplicación de experimentos de *Dictator Game*.

A partir de los experimentos de *Dictator Game* que se utilizarán en este proyecto de investigación se pretende descubrir las diferencias en el comportamiento entre los agentes de AIG y los seres humanos en situaciones de toma de decisiones altruistas. Estos experimentos brindarán un marco estructurado y controlado para observar cómo las personas, ya sean humanos o agentes de AIG, eligen distribuir recursos en situaciones en los que actúan de manera egoísta o altruista.

3.8. Variables o categorías de la investigación

En esta sección se detalla la importancia de entender cómo se interrelacionan y aplican las variables en el contexto de este estudio. El enfoque de este análisis se basa en la observación para lograr una contextualización adecuada de la función de estas variables. En la Tabla 8 Variables de investigación se presentan las etapas y el funcionamiento de estas variables.

Tabla 8 Variables de investigación

Objetivo	Variable	Definición conceptual	Indicador	Instrumentos
Analizar la situación actual sobre el uso de la Inteligencia Artificial Generativa en la simulación y análisis del comportamiento altruista para la evaluación de los métodos y herramientas utilizados en estudios previos.	Estado actual del uso de la Inteligencia Artificial Generativa	Aplicaciones actuales de AIG en simulación y análisis del comportamiento altruista.	Cantidad de aplicaciones actuales que existen de la tecnología relacionada al comportamiento altruista en agentes de inteligencia artificial generativa.	Revisión sistemática de literatura científica por medio de una tabla comparativa.
Determinar cuáles son los criterios que afectan la interacción de los agentes inteligentes en la toma de decisiones altruistas para la identificación de limitaciones relacionadas al comportamiento altruista	Criterios que afectan la interacción de los agentes inteligentes	Identificación y análisis de los factores internos y externos que influyen en la toma de decisiones altruistas por parte de los agentes de IA.	Porcentaje de precisión de las decisiones altruistas. Cantidad de factores que afecta las decisiones altruistas.	Aplicación de experimentos de tipo *Dictator Game*. Tabla comparativa de resultados obtenidos a la largo de la aplicación de experimentos.
	Integridad de los resultados obtenidos	Esta variable evalúa la calidad de los resultados obtenidos y su fiabilidad para representar de manera precisa el fenómeno estudiado.	Porcentaje de consistencia de los resultados obtenidos. Cantidad de repeticiones de los experimentos. Cantidad de revisiones de los datos obtenidos.	Aplicación de experimentos de tipo *Dictator Game*. Limpieza de datos obtenidos por la aplicación de experimentos. Aplicación de análisis estadísticos sobre los datos resultantes de la limpieza de datos.

Objetivo	Variable	Definición conceptual	Indicador	Instrumentos
Examinar en qué medida los agentes inteligentes emulan el comportamiento humano en contextos de toma de decisiones altruistas para la comparación de los resultados obtenidos de ambos grupos de experimentos.	Variación del comportamiento altruista de los AIG	Medición de la capacidad de los agentes de IA para replicar y adoptar comportamientos altruistas similares a los humanos en contextos de toma de decisiones.	Porcentaje de variabilidad en las decisiones altruistas.	Resultados de experimentos de *Dictator Game* con enfoque en decisiones altruistas, que serán analizados estadísticamente.

3.9. Técnicas e instrumentos de recolección de datos

Esta sección describe las técnicas e instrumentos de recolección de datos utilizados en la investigación con el fin de obtener una explicación detallada sobre el modo en el que se implementarán las herramientas en el estudio. Para cada una se brindará una descripción y una explicación sobre los procedimientos específicos planeados para según su uso y, además, se comentará su relevancia para alcanzar los objetivos de la investigación por medio de las variables.

A continuación, en la Tabla 9 Herramientas de recolección de datos se describen las herramientas que se utilizarán a lo largo de la investigación.

Tabla 9 Herramientas de recolección de datos

Herramienta	Descripción	Apéndice de referencia
Revisión sistemática de literatura científica por medio de una tabla comparativa	Según los autores Hernández-Sampieri y R. Mendoza (2018), la revisión literaria consiste en la acción de detectar, consultar y obtener bibliografía y otros materiales que sean útiles para el propósito del estudio. Esta herramienta nos permite conocer el estado actual de estudio recientes y relevantes a la presente investigación. Para este proyecto la revisión se realiza mediante una tabla comparativa.	9.3 Apéndice C: Plantilla de tabla comparativa de análisis documental.
Aplicación de experimentos de tipo *Dictator Game*	Es una parte de la recolección de datos primordial para la investigación que los autores Hernández-Sampieri y R. Mendoza (2018) definen *experimento* como una «situación de control en la cual se manipulan, de manera intencional, una o más variables independientes (causas) para analizar las consecuencias de tal manipulación sobre una o más variables dependientes (efectos)» (p. 152).	9.4 Apéndice D: Código para ejecución de experimentos
Tabla de resultados obtenidos a la largo de la aplicación de experimentos	Como parte de la presentación de «datos crudos», se planea la utilización de una tabla, con el fin de comparar diferentes resultados durante la aplicación de experimentos.	9.5 Apéndice E: Tabla resultado de aplicación de experimentos.

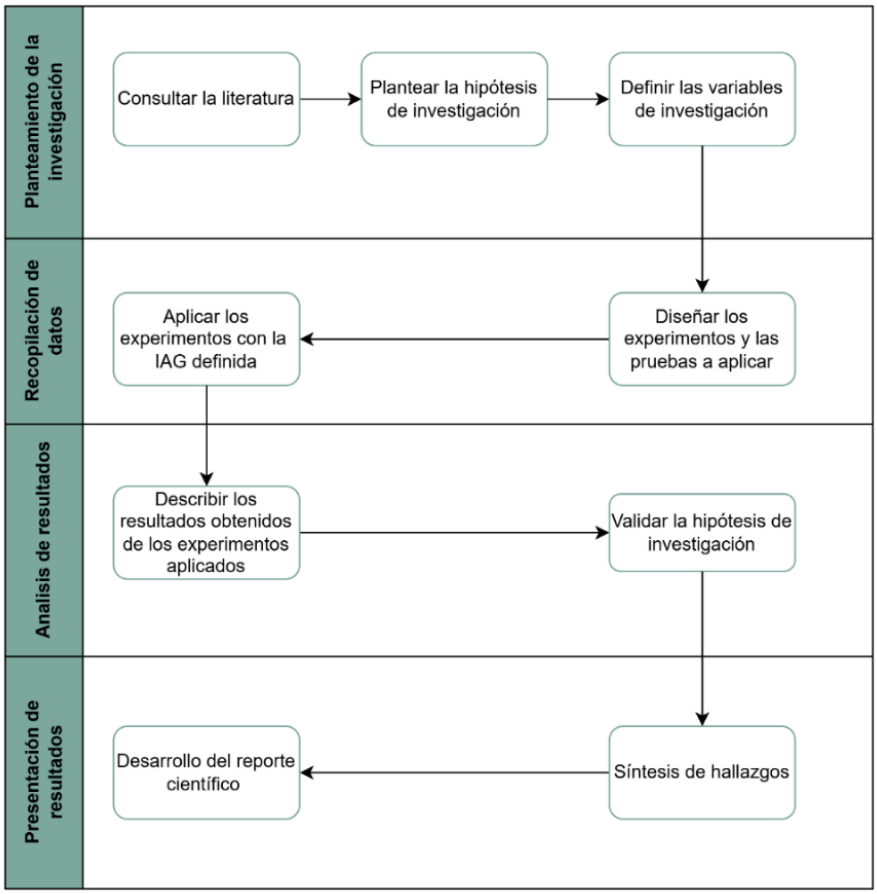
Herramienta	Descripción	Apéndice de referencia
Limpieza de datos obtenidos por la aplicación de experimentos	Se realiza una limpieza de «datos crudos» mediante una depuración y descarte de datos que no aporten valor al estudio. Esta limpieza se realiza por medio del proceso de ETL (por sus siglas en inglés de e*xtract, transform and load*)	9.6 Apéndice F: Proceso de ETL
Aplicación de análisis estadísticos sobre los datos resultantes de la limpieza de datos	A partir la información obtenida de los resultados que fueron analizados por el 9.6 Apéndice F: Proceso de ETL, se aplican diferentes análisis estadísticos, entre ellos un análisis de la varianza (ANOVA).	9.7 Apéndice G: Análisis estadístico 9.8 Apéndice H: Comprobación de hipótesis por medio de un análisis ANOVA
Resultados de experimentos de *Dictator Game* con enfoque en decisiones altruistas, analizados estadísticamente	Se prevé el uso de una tabla comparativa para presentar datos finales de la investigación, con el propósito de contrastar los diferentes resultados obtenidos durante el análisis estadístico. Esta actividad se centrará en los resultados de los experimentos de *Dictator Game*, que serán evaluados estadísticamente. Además, se considera en esta herramienta la comparativa del comportamiento sistemático del ser humano obtenido a partir del metaanálisis realizado por Engel (2010), detallado en la sección 2.3.3 Teoría de juegos: *Dictator Game*.	9.8 Apéndice H: Comprobación de hipótesis por medio de un análisis ANOVA 9.9 Apéndice I: Tabla comparativa de resultados

3.10. Procedimiento metodológico de la investigación

A lo largo de esta sección se detallan las fases y etapas contempladas en el procedimiento metodológico de la investigación, además de las tareas que corresponden a cada fase.

Para facilitar la comprensión de las fases y etapas que conlleva el procedimiento metodológico, se presenta la Figura 7 Diagrama de fases y etapas del procedimiento metodológico, en donde se detalla cada una de las fases.

Figura 7 Diagrama de fases y etapas del procedimiento metodológico



Fuente elaboración propia, 2024

3.10.1. Fase 1: Planteamiento de la investigación

Durante esta fase se definen las bases de la investigación, que incluyen actividades de investigación, análisis y planteamiento. Todas ellas definen la línea de investigación mediante las siguientes cuatro etapas:

3.10.1.1. Etapa 1: Consultar la literatura

A lo largo de la etapa de análisis literario, se lleva a cabo una revisión de estudios previos con el objetivo de establecer un historial sobre los trabajos realizados en el área de comportamiento altruista de la inteligencia artificial generativa mediante experimentos de tipo *Dictator Game*. Esta revisión se enfoca en identificar y analizar investigaciones que abordan aspectos similares a los planteados en esta investigación, con el fin de facilitar a través de ellas una comprensión profunda de los enfoques y hallazgos anteriores. De este modo, mediante la consulta de diversas fuentes literarias, se busca construir una base sólida de conocimiento que sustente y contextualice el presente estudio.

En el capítulo 2 Estado del Arte, se detalla el procedimiento ejecutado para llevar a cabo esta revisión literaria. Además, se presentan los resultados obtenidos de esta revisión y se destacan los principales hallazgos y las tendencias identificadas en la literatura existente, con

el fin de proporcionar un estado del arte que enmarque la investigación actual y resalte la importancia y relevancia de los temas abordados.

3.10.1.2. Etapa 2: Plantear la hipótesis de investigación

En la etapa 2 se contempla el planteamiento de hipótesis y suposiciones tentativas, las cuales son fundamentales para guiar el rumbo de la investigación. Estas hipótesis se elaboran con el objetivo de establecer afirmaciones provisionales que, mediante el proceso de investigación, se confirman o refutan. En esta etapa se anticipan los posibles resultados que se quieren demostrar a lo largo del estudio.

Para esta investigación se plantea la siguiente hipótesis: más detalles de la misma se encuentran concentrados en la sección 3.7 Hipótesis :

- Existe variación en las capacidades de los agentes de IAG para tomar decisiones altruistas en la aplicación de experimentos de *Dictator Game*.3.7Hipótesis

3.10.1.3. Etapa 3: Definir las variables de investigación

Las variables de investigación se definen con base en los objetivos específicos de la investigación, asegurando una estrecha relación con las preguntas investigativas planteadas. Estas variables representan los elementos que se medirán y analizarán a lo largo del estudio y proporcionan un marco estructurado para la recopilación y evaluación de datos. Por esta razón, la identificación de las variables permite establecer un vínculo directo entre los objetivos de la investigación y los resultados obtenidos, apoyando en la validación de la hipótesis.

Más adelante, se ofrece una descripción detallada de cada una de las variables de investigación; la sección 3.8 Variables o categorías de la investigación proporciona una definición clara de cada variable, explicando su relevancia y cómo se relaciona con los objetivos.

Finalmente, en la sección 3.11 Operacionalización de las variables se detalla el proceso mediante el cual cada variable será medida y analizada, incluyendo los indicadores y herramientas para la recolección de los datos.

3.10.2. Fase 2: Recopilación de datos

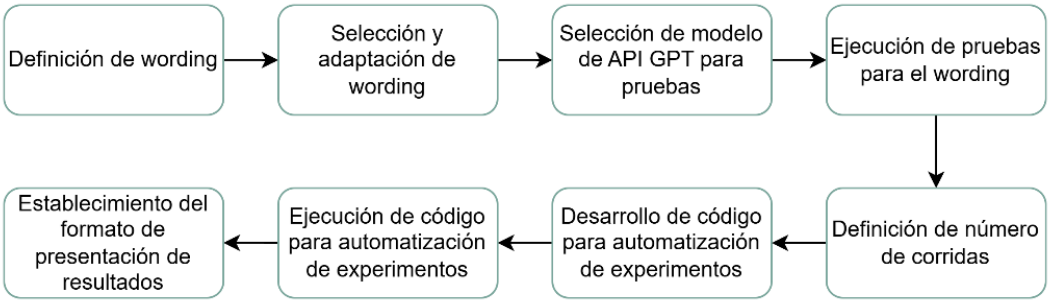
En la Fase 2 se procede a diseñar e implementar los experimentos, que se encuentran basados en los experimentos de teoría de juegos llamados *Dictator Game.* Estos se implementan en conjunto con los agentes de inteligencia artificial generativa de OpenAI, versión 3.5-turbo para el grupo de control y para el grupo experimental, las versiones 4, 4o, 4-turbo, 4o-mini.

Una vez que los experimentos han sido implementados, se lleva a cabo la recopilación de los datos obtenidos a través de la interacción de los agentes de IAG. Este proceso incluye la especificación detallada de los pasos por seguir durante la ejecución de los experimentos, asegurando que las condiciones de los dictámenes sean claras y consistentes en cada iteración. Los datos crudos recopilados serán utilizados para realizar un análisis comparativo entre las decisiones tomadas por los agentes de IAG.

3.10.2.1. Etapa 4: Diseñar los experimentos y pruebas a aplicar

A partir de la imagen Figura 8 Actividades de Etapa 4 Diseño experimental se define la lista de actividades que fueron consideradas para establecer el diseño del experimento del proyecto de investigación.

Figura 8 Actividades de Etapa 4 Diseño experimental



A continuación, se detalla cada una de las tareas que están presentadas en la imagen anterior.

Para la aplicación de los experimentos, uno de los primeros pasos críticos es definir el *wording* o la redacción exacta que se utilizará para presentar las instrucciones a los agentes de inteligencia artificial generativa (IAG). De una manera más detallada, este término se refiere a la redacción acertada de las instrucciones o preguntas que se presentan al modelo para guiar su comportamiento y asegurar que su interpretación sea coherente con los objetivos del experimento.

El proceso de diseño del *wording* requiere una atención minuciosa a los detalles; para que el experimento sea válido y replicable, es necesario evitar cualquier ambigüedad en las instrucciones. En el caso de los modelos de lenguaje, que interpretan el ingreso textual de forma altamente contextual, incluso una ligera variación en la formulación altera la respuesta. Por esta razón, en este tipo de investigaciones el *wording* debe ser neutral y preciso.

Adicionalmente, se realizó un análisis detallado de estudios previos, considerando diferentes opciones de *wording* que hayan sido exitosas en experimentos similares, especialmente en el ámbito de juegos económicos como el *Dictator Game*. A partir de este análisis, se decidió optar por un *wording* basado en investigaciones anteriores, el cual fue adaptado a las necesidades específicas del presente proyecto, con el fin de garantizar la claridad de las instrucciones y la consistencia en la interpretación por parte del modelo GPT.

Este análisis se detalla a lo largo de la sección Revisión de estudios previos, específicamente en el estudio *Playing Games With GPT: What Can We Learn About a Large Language Model From Canonical Strategic Games?* Por Brookins y Debacker, (2024).

Contemplado las características anteriores, se presenta el *wording* utilizado:

This task is about dividing money between yourself and another person to whom you are randomly matched. You do not know this other person and you will not knowingly meet

him/her. You have been randomly assigned the role of the 'allocator'. The other person is in the role of the 'recipient'.

You are endowed with 10 euro, the recipient is endowed with 0 euro. You can decide how much of your 10 euro endowment to transfer to the recipient. You can choose any amount between 0 euro and 10 euro. The recipient receives the amount that you decide to transfer to him/her; you receive the amount that you decide not to transfer and thus to keep.

How much of your 10 euro endowment do you want to transfer to the recipient? Just tell me the allocation, not your reasoning.

Posteriormente, se llevaron a cabo pruebas con el modelo GPT-3.5-turbo, en esta investigación el grupo de control definido, para evaluar el *wording* seleccionado. La elección de este modelo, en lugar de versiones más avanzadas como GPT-4o o 4.0-turbo, se debió principalmente a su bajo costo y su menor tiempo de procesamiento, lo que permitió realizar un mayor número de pruebas sin comprometer los recursos del proyecto; para detalles sobre los consumos y limites revisar la Tabla 7 Tabla de consumo comparativo por modelo.

Aunque el modelo GPT-3.5-turbo no será el utilizado para la recolección de datos definitiva, su empleo en esta fase preliminar es clave para comprobar la eficacia del *wording*, ya que la coherencia de las respuestas y la comprensión de las instrucciones por parte del modelo es esencial antes de proceder con los modelos más avanzados. Sobre las selecciones de modelos a utilizar se detallan características en la sección Población y selección de muestra.

En cuanto a la ejecución técnica de estas pruebas, se trabajó en el desarrollo del código necesario para automatizar el proceso (detallado en el Apéndice D: Código para ejecución de experimentos). El lenguaje de programación seleccionado fue Python, debido a su compatibilidad con el API de OpenAI y su flexibilidad para manejar grandes volúmenes de datos. Este código no solo permite la interacción con el modelo de IA, sino que también facilita la selección y modificación del *wording* de manera eficiente.

En cuanto a la cantidad de pruebas por realizar, se definió que el número óptimo de corridas es de aproximadamente 500. Esta decisión se fundamenta tanto en estudios previos descritos en la sección Revisión de estudios previos como en las estimaciones calculadas en la Tabla 7 Tabla de consumo comparativo por modelo, ubicada en la sección de Población y selección de muestra. Este número permite asegurar que los resultados obtenidos sean representativos y que cualquier variación observada sea estadísticamente significativa. Además, el gran volumen de datos ayudará a identificar patrones o inconsistencias en las respuestas del modelo, lo cual es importante para garantizar la validez de los experimentos.

Por último, se estableció el formato en el que se almacenarán los resultados de las pruebas, verificar en Apéndice E: Tabla resultado de aplicación de experimentos. Se optó por un archivo con extensión .xlsx que contiene cinco columnas: *Run Number, Total Amount, GPT Response, Full Response* y Error Message. La columna *Run Number* indicará el número de ejecución de la prueba; *Total Amount* reflejará la cantidad total asignada en el experimento, que será fija en 10 unidades para todos los casos; *GPT Response* contendrá la respuesta generada por el modelo IAG, mientras que *Full Response* almacenará una representación en formato JSON de la respuesta completa, incluyendo detalles relevantes para su análisis.

Finalmente, *Error Message* se utilizará para registrar cualquier error que ocurra durante la ejecución de las pruebas. Esta estructura de almacenamiento permitirá organizar y analizar los resultados de manera efectiva, facilitando la posterior interpretación de los datos.

A partir de la implementación de las pruebas preliminares y la configuración del entorno experimental, es posible avanzar hacia la fase de análisis de datos. En esta etapa, se evaluará la eficacia del *wording* seleccionado y el desempeño de los modelos de IA frente a las instrucciones presentadas.

3.10.2.2. Etapa 5: Aplicar los experimentos a los Agente de Inteligencia Artificial Generativa

Durante la etapa 5 se trabaja en la aplicación de los experimentos diseñados en 3.10.2.1 **Etapa 4: Diseñar los experimentos y pruebas a aplicar** a los modelos de la sección 3.5 Población y selección de muestra, que serán GPT-3.5-turbo, GPT-4, GPT-4-turbo, GPT-4o, GPT-4o-mini. Cabe resaltar que la aplicación de estos experimentos se realiza con las versiones vigentes entre las fechas del 2 de setiembre del 2024 al 14 de setiembre del 2024.

A lo largo de la ejecución inicial de pruebas se detectan algunas inconsistencias en los comportamientos de los modelos de la familia GPT-4, en comparación con los resultados encontrados en la etapa de diseño con el modelo GPT-3.5-turbo. Estas inconsistencias llevan a un proceso de ajuste a las configuraciones detalladas en la etapa de diseño. Sin embargo, durante este proceso se detecta que para optimizar el funcionamiento de los modelos GPT-4, se deben realizar dos ajustes a nivel de configuración de los modelos y código de aplicación de experimentos. Ambos ajustes serían en la variable *temperatura* y en rol de *system*, los cuales se detallan a continuación. Para la variable *temperatura*, OpeanAI (2024) comenta que «La temperatura tiene valores a utilizar, entre 0 y 2. Valores más altos como 1,8 harán que la salida sea más aleatoria, mientras que valores más bajos como 0,2 la harán más centrada y determinista» OpeanAI (2024). Por lo tanto, se realiza el ajuste del valor 1, utilizado para el proceso de diseño de experimentos, y se cambia al valor 1.5. Este ajuste resultó necesario ya que con uso de los modelos de la familia GPT-4 se obtienen resultados distribuidos y no resultados repetidos del 100% de los casos.

Para el ajuste del rol, este se configura por medio de mensajes que forman parte de la «conversación» que se mantiene con el modelo. En el caso de la etapa de diseño de experimentos, se tiene una configuración de rol como *Undergrate student*, haciendo referencia únicamente al perfil que debería de tener el modelo. Sin embargo, ante las inconsistencias que se experimentaron con la presentación únicamente del rol anterior, se decide la expansión del contexto a *An undergraduate student participante of an economics experiment who only has 10 euros and is being pay to participate*, mostrando no solo el rol sino también el contexto y las características de la situación, que refuerzan el uso del contexto de los experimentos garantizando resultados prometedores.

Con referencia a la preparación de cada modelo, se describen algunas características de cada uno:

- GPT-4: se realiza la aplicación de los experimentos con la versión gpt-4-0613, versión del 13 de junio de 2023. Esta versión contiene un soporte mejorado de

llamadas a funciones. Cabe recalcar que este modelo tuvo el mayor consumo de tokens y tiempo de ejecución, posicionándose en el primer lugar en comparación de los demás modelos. (OpenAI, s.f.n.d.-b)

- GPT-4-turbo: al momento de ejecución de pruebas (del 6 de setiembre de 2024 al 14 de setiembre de 2024), el modelo utiliza la versión gpt-4-turbo-2024-04-09 correspondiente a la versión del 9 de abril de 2024. (OpenAI, s.f.n.d.-b)
- GPT-4o: se aplican los experimentos a la versión gpt-4o-2024-05-13 del modelo, con un lanzamiento el 13 de mayo de 2024. Es el modelo insignia de alta inteligencia para tareas complejas de varios pasos. GPT-4o es más barata y rápida que GPT-4 Turbo (OpenAI, s.f.n.d.-b)
- GPT-4o-mini: la versión del modelo utilizada para los experimentos es gpt-4o-mini-2024-07-18, la versión publicada el 18 de julio de 2024 y la más reciente de las utilizadas en la investigación. Se considera un modelo asequible e inteligente para tareas rápidas y ligeras. GPT-4o mini es más barata y capaz que GPT-3.5 Turbo(OpenAI, n.d.-bs.f.b).

3.10.3. Fase 3: Análisis de resultados

Para la se fase 3, se realiza la presentación de resultados, obtenidos por medio de las etapas 6 y 7. Estas se encuentran detalladas en el capítulo 4. Análisis de Resultados

3.10.3.1. Etapa 6: Describir los resultados obtenidos de los experimentos aplicados

En esta etapa se detallan los resultados obtenidos en la Etapa 5. Aplicación de experimentos. La evaluación consta de revisión de distribuciones, así como de sus análisis estadísticos, por cada modelo mencionado en las etapas previas. Para realizar la revisión del tema se toma como referencia la sección 4.1.Síntesis de hallazgos.

3.10.3.2. Etapa 7: Validar la hipótesis de investigación

Una vez analizados los resultados obtenidos por cada modelo, se procede con la validación de la hipótesis presentada en la **Etapa 3: Definir las variables de investigación**. Para esto se utiliza el meta-estudio de Engel (2010) titulado Dictator games: *a meta study* como base de la información en personas. Luego se realiza una prueba de validación de hipótesis utilizando un análisis de varianza (ANOVA) y una prueba de comprobación Tukey HSD.

3.10.4. Fase 4: Presentación de resultados

Para esta fase, se trabaja en los entregables destacados sobre los resultados obtenidos a lo largo de la aplicación de la metodología y se presentan en la discusión y en las conclusiones de la investigación. Además, se trabaja en la elaboración de un reporte científico como documento adjunto a la investigación.

3.10.4.1. Etapa 8: Síntesis de hallazgos

Para esta etapa se realiza la puntualización de los resultados obtenidos de la investigación, que se encuentran detallados en las secciones 5.Discusión y 6.Conclusiones del presente documento.

3.10.4.2. Etapa 9: Desarrollo del reporte científico

Para la completitud de la investigación, se realiza la elaboración de un reporte científico, que se detalla en el 10.1Anexo I: Articulo científico.

3.11. Operacionalización de las variables

A lo largo de la Tabla 10 Operacionalización de las variables se presenta el mapeo de la relación entre los objetivos específicos, las fases de investigación y las variables de investigación, con el fin de facilitar el entendimiento de los vínculos entre los distintos componentes de la investigación.

Tabla 10 Operacionalización de las variables

Objetivo	Fase	Variable	Indicador	Instrumentos
Analizar la situación actual sobre el uso de la Inteligencia Artificial Generativa en la simulación y análisis del comportamiento altruista para la evaluación de los métodos y herramientas utilizados en estudios previos.	Fase 1: Planteamiento de la investigación	Estado actual del uso de la inteligencia artificial generativa	Cantidad de aplicaciones actuales que existen en la tecnología relacionadas con el al comportamiento altruista en agentes de IAG.	Revisión sistemática de literatura científica por medio de una tabla comparativa
Determinar cuáles son los criterios que afectan la interacción de los agentes inteligentes en la toma de decisiones altruistas para la identificación de limitaciones relacionadas al comportamiento altruista	Fase 2: Recopilación de datos	Criterios que afectan la interacción de los agentes inteligentes.	Porcentaje de precisión de las decisiones altruistas. Cantidad de factores que afectan las decisiones altruistas.	Aplicación de experimentos de tipo *Dictator Game*. Tabla comparativa de resultados obtenidos a la largo de la aplicación de experimentos.
	Fase 3: Análisis de resultados	Integridad de los resultados obtenidos.	Porcentaje de consistencia de los resultados obtenidos. Cantidad de repeticiones de los experimentos. Cantidad de revisiones de los datos obtenidos.	Aplicación de experimentos de tipo *Dictator Game*. Limpieza de datos obtenidos por la aplicación de experimentos. Aplicación de análisis estadísticos sobre los datos resultantes de la limpieza de datos.

Objetivo	Fase	Variable	Indicador	Instrumentos
Examinar en qué medida los agentes inteligentes emulan el comportamiento humano en contextos de toma de decisiones altruistas para la comparación de los resultados obtenidos de ambos grupos de experimentos.	Fase 3: Análisis de resultados	Variación del comportamiento altruista de los AIG.	Porcentaje de variabilidad en las decisiones altruistas.	Resultados de experimentos de *Dictator Game* con enfoque en decisiones altruistas, analizados estadísticamente. Datos del metaanálisis de estudio.
		Variación del comportamiento altruista de los AIG con respecto al comportamiento sistemático del ser humano.	Valor F obtenido del análisis de varianza (ANOVA)	

4. Análisis de Resultados

Una vez finalizada la **Etapa 5: Aplicar los experimentos a los Agente de Inteligencia Artificial Generativa**, se continua con la Fase 4: Presentación de resultados, en la cual se abarcan las etapas **Etapa 8: Síntesis de hallazgos** y **Etapa 9: Desarrollo del reporte científico**.

A lo largo del capítulo se aborda la descripción de hallazgos, mientras se califican los distintos modelos de GPT-4 y GPT-3.5-turbo. Posteriormente, se valida la hipótesis contra cada modelo de la familia de GPT-4 y GPT-3.5-turbo.

4.1. Síntesis de hallazgos

En este apartado se estudia la aplicación de las pruebas de experimentos y se procede a la descripción de los resultados obtenidos junto con su respectivo análisis estadístico. Para este propósito, se realiza una división por modelo.

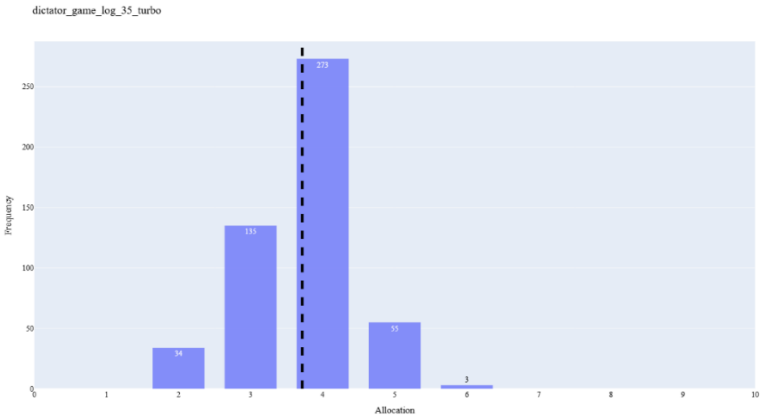
A modo de recordatorio y según lo definido en la **Etapa 5: Aplicar los experimentos a los Agente de Inteligencia Artificial Generativa**, se deben considerar las siguientes características:

- Los modelos utilizados para la aplicación de experimentos, según la sección Población y selección de muestra, son GPT-3.5-turbo como grupo de control y las versiones GPT-4, GPT-4-turbo, GPT-4o, GPT-4o-mini como grupo experimental.
- La cantidad total de aplicación por modelo es de 500 veces y con un máxima de distribución de 10 euros.
- La configuración del sistema se encuentra con el siguiente contexto: «*An undergraduate student participante of an economics experiment who only has 10 euros and is being pay to participate*», y el parámetro *temperature* en 1.5.

4.1.1. GPT-3.5-turbo

A continuación, se presentan los resultados para el modelo GPT-3.5-turbo, considerado el grupo de control.

Figura 9 Histograma de distribución Modelo GPT-3.5-turbo



El gráfico de la Figura 9 Histograma de distribución Modelo GPT-3.5-turbo muestra la distribución de las respuestas del modelo GPT-3.5- turbo, el mismo que fue utilizado en la **Etapa 4: Diseñar los experimentos y pruebas a aplicar**. El valor modal es 4 euros, con 273 ocurrencias, lo que sugiere una clara tendencia del modelo a favorecer este monto. También se observan resultados frecuentes en 3 euros y 5 euros, con 135 y 55 apariciones, respectivamente. Las respuestas extremas, como 2 euros y 6 euros, son mucho menos comunes, con 34 y 3 apariciones, respectivamente. Esta distribución muestra un sesgo leve hacia la izquierda, ya que la mayor parte de las asignaciones se concentra alrededor del valor de 4 euros, lo que podría indicar una preferencia del modelo por asignar cantidades moderadas. La línea punteada negra representar la media con un valor de 3.7160 euros.

Con lo que respecta al análisis estadístico, se presenta la siguiente tabla:

Tabla 11 Análisis estadístico modelo GPT-3.5-turbo

Estadística	Cálculo
Conteo	500 respuestas
Media	3.7160 euros
Desviación estándar	0.772358 euros
Valor mínimo	2 euros
25%	3 euros
50%	4 euros
75%	4 euros
Valor máximo	6 euros

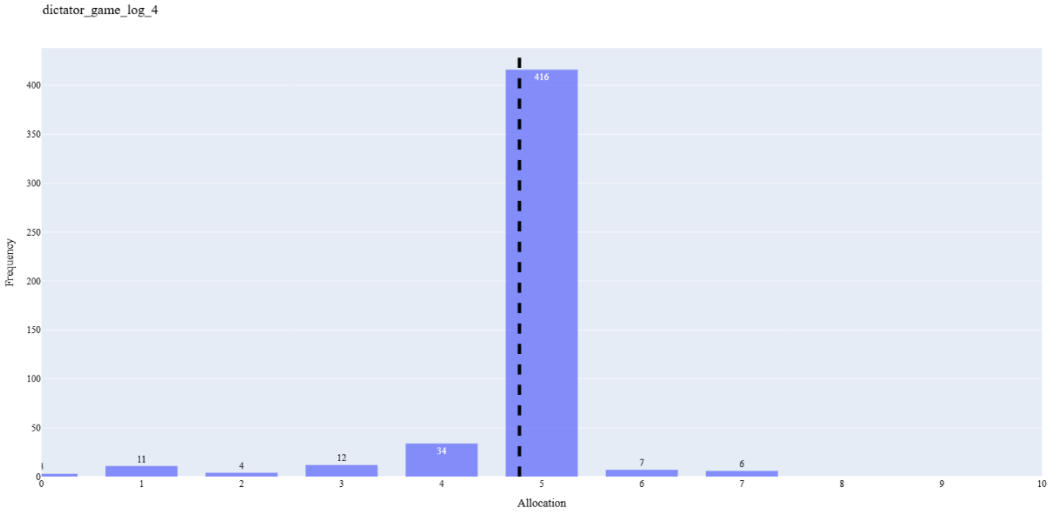
La Tabla 11 Análisis estadístico modelo GPT-3.5-turbo muestra las estadísticas descriptivas de 500 respuestas del modelo GPT-3.5-turbo. La media de las asignaciones fue de 3.7160 euros, con una desviación estándar de 0.77 euros, lo que indica una variación moderada en las respuestas. El valor mínimo asignado fue 2 euros, mientras que el máximo fue 6 euros.

Los percentiles muestran que el 25% corresponde a 3 euros, mientras que el 50% y el 75% de las asignaciones fueron exactamente 4 euros. De este modo, se presenta una distribución de datos entre los valores de 2 y 6 euros.

4.1.2. GPT-4

A continuación, se muestran los resultados para el modelo GPT-4.

Figura 10 Histograma de distribución Modelo GPT-4



El gráfico de la Figura 10 Histograma de distribución Modelo GPT-4 presenta la distribución de las asignaciones realizadas por el modelo GPT-4 en la aplicación de experimentos. En el eje X se exponen las diferentes cantidades asignadas, que van de 0 a 10, mientras que el eje Y indica la frecuencia con la que cada asignación fue elegida. El total de resultados representados en la gráfica es de 493.

La mayor cantidad de respuestas poseen asignación 5, con una frecuencia notablemente alta de 416, lo que sugiere una fuerte preferencia por una división equitativa de los recursos. Las otras respuestas fueron mucho menos frecuentes, con valores que varían entre 1 y 7, de entre las cuales la asignación de 4 fue la segunda más común, con una frecuencia de 34. No se registraron asignaciones de 0, 8, 9 o 10.

El gráfico también incluye una línea discontinua que representa la media de los valores, con un valor exacto de 4.776876.

Con lo que respecta al análisis estadístico, se presenta la Tabla 12 Análisis estadístico modelo GPT-4

Tabla 12 Análisis estadístico modelo GPT-4

Estadística	Cálculo
Conteo	493 respuestas
Media	4.776876 euros
Desviación estándar	0.873569 euros
Valor mínimo	0 euros
25%	5 euros
50%	5 euros
75%	5 euros
Valor máximo	7 euros

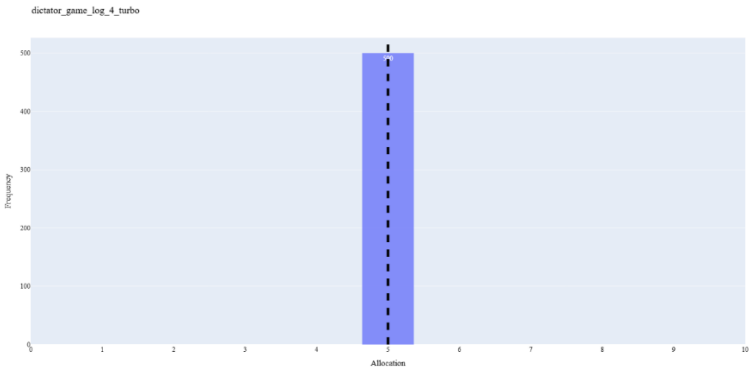
Los datos de la Tabla 12 Análisis estadístico modelo GPT-4 muestran el análisis estadístico de los resultados obtenidos por el modelo GPT-4. Se realiza el conteo de 493 asignaciones de las 500 ejecuciones, con una media de 4.77 euros y una desviación estándar de 0.87 euros. Estos datos indican que la mayoría de las asignaciones están cercanas al valor medio.

El valor mínimo asignado fue de 0 euros, mientras que los percentiles 25%, 50% (mediana) y 75% muestran que la mayoría de los resultados es de 5 euros. El valor máximo asignado fue de 7 euros, lo que evidencia que los participantes se inclinaron mayoritariamente por la equidad, con poca variación hacia valores más altos o bajos.

4.1.3. GPT-4-turbo

A continuación, se presentan los resultados para el modelo GPT-4-turbo.

Figura 11 Histograma de distribución Modelo GPT-4-turbo



El gráfico de la Figura 11 Histograma de distribución Modelo GPT-4-turbo muestra la distribución de las respuestas brindadas por el modelo GPT-4-turbo, en el que todas las respuestas seleccionaron exactamente 5 euros y que, por ende, tuvo una frecuencia de 500. Esto indica una total uniformidad en la decisión, ya que no se realizaron asignaciones en ningún otro valor dentro del rango posible de 0 a 10. La gráfica destaca esta preferencia por su equidad, representada visualmente con una sola barra centrada en 5, acompañada de una línea discontinua que representa la media de las respuestas: 5.

Con lo que respecta al análisis estadístico, se presenta la Tabla 13 Análisis estadístico modelo GPT-4-turbo.

Tabla 13 Análisis estadístico modelo GPT-4-turbo

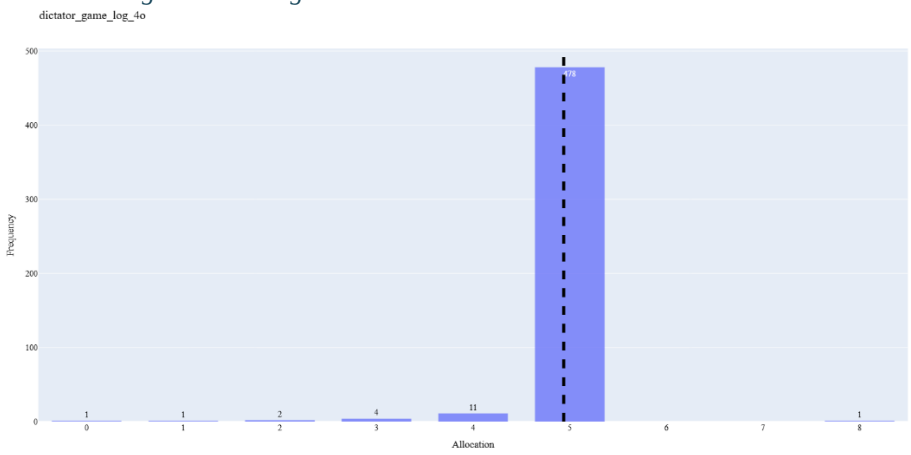
Estadística	Cálculo
Conteo	500 respuestas
Media	5 euros
Desviación estándar	0 euros
Valor mínimo	5 euros
25%	5 euros
50%	5 euros
75%	5 euros
Valor máximo	5 euros

La Tabla 13 Análisis estadístico modelo GPT-4-turbo presenta las estadísticas descriptivas de 500 respuestas brindas por el modelo GPT-4-turbo, en el que todas las repuestas arrojaron exactamente 5 euros. Como consecuencia, la media es de 5 euros, con una desviación estándar de 0 euros, lo que refleja una total ausencia de variabilidad en las asignaciones. Los valores mínimos, máximo, y los percentiles (25%, 50%, 75%) también son de 5 euros, confirmando que todos los resultados fueron 5 euros.

4.1.4. **GPT-4o**

A continuación, se presentan los resultados para el modelo GPT-4o.

Figura 12 Histograma de distribución Modelo GPT-4o



El gráfico de la Figura 12 Histograma de distribución Modelo GPT-4o muestra la distribución de las respuestas del modelo GPT-4o. En 478 de las 498 respuestas graficadas, la opción escogida fue la de 5 euros, lo que lo coloca como el valor predominante. Sin embargo, también se observan algunas respuestas con frecuencias bajas, en valores como 0, 1, 2, 3, 4 y 8, aunque en cantidades significativamente menores, entre 1 y 11 asignaciones. La gráfica sugiere que, aunque la tendencia predominante es dividir equitativamente el recurso (en 5 euros). Se presenta una línea punteada representando la media con un valor de 4.9377 euros.

Con lo que respecta al análisis estadístico, se presenta la Tabla 14 Análisis estadístico modelo GPT-4o

Tabla 14 Análisis estadístico modelo GPT-4o

Estadística	Cálculo
Conteo	498 respuestas
Media	4.937751 euros
Desviación estándar	0.43274 euros
Valor mínimo	0 euros
25%	5 euros
50%	5 euros
75%	5 euros
Valor máximo	8 euros

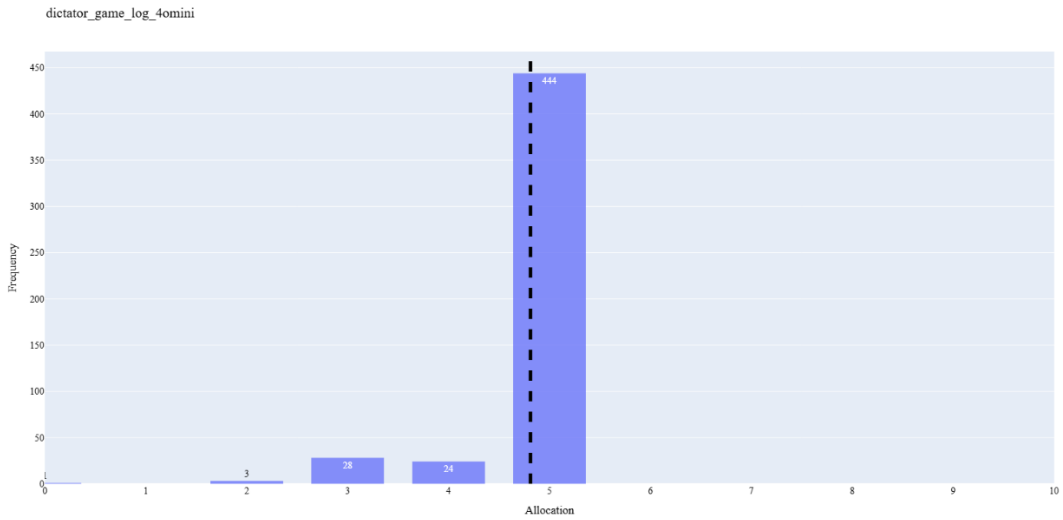
La Tabla 14 Análisis estadístico modelo GPT-4o presenta las estadísticas descriptivas de 498 respuestas del modelo GPT-4o. La media de las asignaciones fue de 4.94 euros, con una desviación estándar de 0.43 euros, lo que indica una variación leve en las respuestas. El valor mínimo asignado fue 0 euros, mientras que el máximo fue 8 euros.

Los percentiles muestran que el 25%, el 50% (mediana) y el 75% de las asignaciones fueron exactamente 5 euros, es decir, que la mayoría de las respuestas fueron esta cantidad, con solo unas pocas decisiones fuera de este valor.

4.1.5. GPT-4o-mini

A continuación, se presentan los resultados para el modelo GPT-4o-mini.

Figura 13 Histograma de distribución Modelo GPT-4o-mini



El gráfico de la Figura 13 Histograma de distribución Modelo GPT-4o-mini presenta la distribución de las asignaciones realizadas por el modelo GPT-4o-mini en la aplicación de experimentos.

La mayor cantidad de respuestas cuentan con una asignación 5 euros y una frecuencia notablemente alta de 444 de 500 respuestas. Las demás respuestas fueron de menor frecuencia, con valores que varían entre 1 y 5 euros, de entre las cuales la asignación de 3 euros se posiciona como la segunda más común, con una frecuencia de 28, seguida de 4 euros con 24 respuestas. No se registraron asignaciones de 1, 6, 7, 8, 9 o 10. Además, el gráfico también incluye una línea discontinua que representa la media de los valores, un valor exacto de 4.812 euros.

Con lo que respecta al análisis estadístico, se presenta la Tabla 15 Análisis estadístico modelo GPT-4o-mini.Tabla 12 Análisis estadístico modelo GPT-4

Tabla 15 Análisis estadístico modelo GPT-4o-mini

Estadística	Cálculo
Conteo	500 respuestas
Media	4.812 euros
Desviación estándar	0.584242 euros
Valor mínimo	0 euros
25%	5 euros
50%	5 euros
75%	5 euros
Valor máximo	5 euros

La Tabla 15 Análisis estadístico modelo GPT-4o-mini presenta las estadísticas descriptivas de 500 respuestas del modelo GPT-4o. La media de las asignaciones fue de 4.812 euros, con una desviación estándar de 0.58 euros, lo que indica una variación leve en las respuestas. El valor mínimo asignado fue de 0 euros, mientras que el máximo fue 5 euros, con lo cual se evidencia una tendencia hacia repuestas con menor valor para el receptor. Los percentiles muestran que el 25%, el 50% (mediana) y el 75% de las asignaciones fueron exactamente 5 euros.

A modo de resumen de la presente sección se demuestra que existe una mínima variación de los cuatro modelos estudiados, dentro de los cuales el GPT-4turbo muestra un comportamiento completamente homogéneo, asignando siempre 5 euros, sin ninguna variación. En cambio, las otras versiones presentan una mayor diversidad en sus decisiones. El modelo GPT-4 es el más disperso, con una desviación estándar de 0.87 y un rango de asignaciones entre 0 y 7 euros, lo que sugiere que esta versión toma decisiones más variadas. Por su lado, los modelos GPT-4o y GPT-4o-mini se comportan de manera más consistente que GPT-4, pero aún muestran cierta variabilidad, especialmente GPT-4o, que tiene una desviación estándar de 0.43 y un máximo de 8 euros. De esta manera, se concluye que el modelo GPT-4turbo es completamente uniforme en sus decisiones, mientras que las demás versiones muestran variaciones, como es el caso de GPT-4, la más dispersa, y GPT-4o, que toma decisiones ligeramente más altruistas en algunos casos.

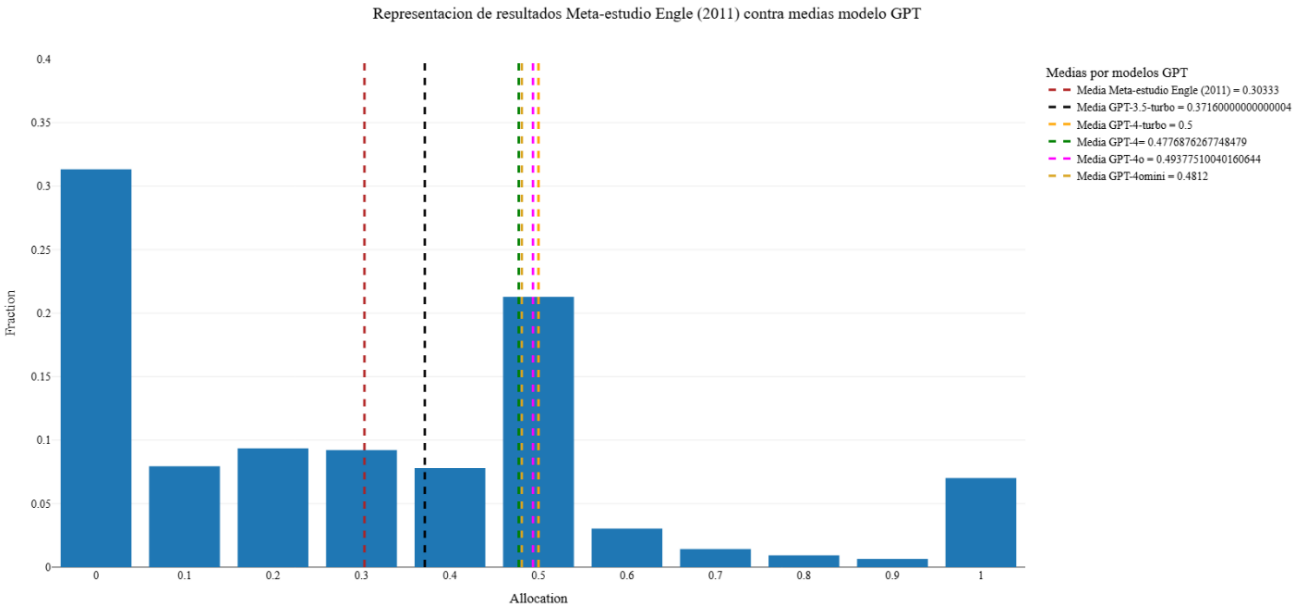
4.2. Validar la hipótesis de investigación

Como parte del capítulo de análisis de resultados, se valida la hipótesis planteada en las secciones **Etapa 2: Plantear la hipótesis de investigación**, la cual se expresa de la siguiente manera: «Existe variación en las capacidades de los agentes de IAG para tomar decisiones altruistas en la aplicación de experimentos de *Dictator Game*».

Para empezar la etapa de validación, se definen los datos de comparación, que se recolectan del estudio realizado por Engel (2010). Este realizó una comparativa de las respuestas brindadas por 290 estudios de experimentos de laboratorio en humanos, que tenían la condición de ser una partida única del *Dictator Game*. Uno de los hallazgos principales es que en un 30% de respuestas no se asignó ningún valor.

Así mismo, el segundo hallazgo que más se hace notar en la gráfica de la Figura 14 es que la distribución que los agentes de IAG realizaron en los experimentos tiende a realizar una partición equitativa de 50 a 50 sobre el monto por dividir. Para finalizar, la definición de la media del trabajo realizado por Engel (2010) que se usa para la comparación de este estudio será representada por la línea roja punteada de la Figura 14.

Figura 14 Representación de resultados Meta-estudio Engle (2011) contra medias modelo GPT



El gráfico de la Figura 14 ofrece una representación visual en la que se percibe la existencia de variación en las capacidades de los agentes de IAG para tomar decisiones altruistas en la aplicación de experimentos del *Dictator Game*. Dicha variación se encuentra en el estudio cuando se comparan las medias de asignaciones entre diferentes modelos GPT y el metaestudio humano de Engle (2010). La línea punteada roja, que marca la media de los datos del metaestudio de Engle (2010) corresponde a una asignación promedio de 0.3033 (30.33%). Este promedio refleja que los participantes humanos tienden a asignar un porcentaje más bajo del total disponible al receptor, lo que sugiere un menor grado de altruismo en comparación con los agentes de IAG. Por otro lado, las medias de los diferentes modelos GPT, indicadas por las líneas punteadas de otros colores, muestran que los agentes de IAG tienden a asignar cantidades cercanas al 50%, lo que la convierte la media de GPT-4-turbo en la más alta, que asigna exactamente 0.5 (50%).

Mientras que los humanos parecen ser más conservadores, reteniendo más recursos para sí mismos, los agentes de IAG muestran una tendencia hacia una mayor equidad, asignando por lo general la mitad de los recursos. Además, dentro de los modelos GPT, también hay variaciones: aunque todos se aproximan al 50%, existen ligeras diferencias entre ellos, lo que refuerza la idea de que la arquitectura y el entrenamiento de los modelos influyen en su capacidad para tomar decisiones altruistas.

Para la segunda parte de la etapa de validación de hipótesis, se plantea la prueba ANOVA (análisis de la varianza) como herramienta de comprobación de hipótesis. Para este ejercicio se utiliza el código presentado en el 9.8 Apéndice H: Comprobación de hipótesis por medio de un análisis ANOVA.

Una prueba de hipótesis con ANOVA se utiliza para determinar si existen diferencias significativas entre las medias de tres o más grupos independientes. En esta prueba, la hipótesis nula (H_0) establece que todas las medias son iguales, mientras que la hipótesis alternativa (H_1) plantea que al menos una de las medias es diferente. De esta manera, la herramienta ANOVA analiza la variabilidad entre los grupos en comparación con la variabilidad dentro de los grupos, y si el valor F resultante es suficientemente grande, se rechaza la hipótesis nula, concluyendo que al menos un grupo difiere significativamente (Montgomery, 2019).

Como datos para el análisis se utilizan aquellos obtenidos y descritos en la etapa Síntesis de hallazgos, específicamente en los modelos GPT-3.5-turbo, como grupo de control y los modelos GPT-4, GPT-4-turbo, GPT-4o, GPT-4o-mini como grupo experimental. En la Tabla 16 Resultados Herramienta ANOVA se presentan los valores relevantes para la comprobación de hipótesis estadísticas.

Tabla 16 Resultados Herramienta ANOVA

	DF	Sum Sq	Mean Sq	F value	Pr(>F)
Modelo	4	559.7	139.93	371.4	<2e-16
Residuales	2486	936.5	0.38		

A partir de los datos observados en la Tabla 16 Resultados Herramienta ANOVA, se presenta detalladamente el significado de cada una de las columnas y su relevancia para la validación de la hipótesis

- **DF** (grados de libertad): corresponde a los grados de libertad que se encuentran asociados al efecto de las filas, el cual se calcula a partir del número de grupos menos uno.
- **Sum Sq** (suma de cuadrados): indica el resultado del cálculo de la suma de cuadrados de las filas.
- **Mean Sq** (media de cuadrados): valor calculado a partir de la división de la suma de cuadrados entre los grados de libertad.
- **F value** (valor F): Es una medida que compara la variabilidad explicada por el modelo con la variabilidad no explicada (residual).
- **Pr(>F)** (valor P): es el valor P asociado al valor F. Un valor p muy bajo (menor que cualquier nivel de significancia común, como 0.05 o 0.01) indica que hay evidencia estadística significativa para rechazar la hipótesis nula.

Una vez descritos los datos de la Tabla 16 Resultados Herramienta ANOVA, se consideran las siguientes conclusiones:

- Con respecto al valor F, al ser un valor de 371.4 se considera un valor elevado, lo cual es el primer indicador relevante que señala la probabilidad de que existan diferencias significativas entre los grupos.

- Alrededor del valor P, su valor de <2e-16 es equivalente a un valor menor que 0.001, de modo que sugiere que existen diferencias significativas entre al menos dos o más de los grupos.
- Dentro de la significancia de los resultados, al obtener un valor P mejor a 0.001, se concluye que existen diferencias significativas en la variable asignación entre los diferentes modelos.

Al recibir estos resultados con la herramienta ANOVA, podemos concluir que se valida la hipótesis estadística alternativa (H_1) en la que se plantea que al menos una de las medias es diferente. Una vez que se confirma la existencia de la varianza en las medias, se procede con la aplicación una prueba Tukey HSD (*Honest Significant Difference*), es decir, un análisis *post-hoc* que es aplicado después de realizar un ANOVA, cuando se detectan diferencias significativas entre grupos. Su objetivo es realizar comparaciones múltiples entre todas las posibles combinaciones de pares de medias de los grupos para identificar cuáles de ellas son significativamente diferentes entre sí (Hsu, 2010). A continuación, se presenta la Tabla 17 Resultados prueba Tukey HSD con la información obtenida por la aplicación de la prueba Tukey HSD.

Tabla 17 Resultados prueba Tukey HSD

	diff	lwr	upr	p adj
gpt_4-gpt_35turbo	1.06087627	0.954534690	1.167217850	0.000000000
gpt_4o-gpt_35turbo	1.22175100	1.115678570	1.327823440	0.000000000
gpt_4omini-gpt_35turbo	1.09600000	0.990033900	1.201966100	0.000000000
gpt_4turbo-gpt_35turbo	1.28400000	1.178033900	1.389966100	0.000000000
gpt_4o-gpt_4	0.16087474	0.054427190	0.267322280	0.000367879
gpt_4omini-gpt_4	0.03512373	-0.071217850	0.141465310	0.896350700
gpt_4turbo-gpt_4	0.22312373	0.116782150	0.329465310	0.000000114
gpt_4omini-gpt_4o	-0.12575100	-0.231823440	-0.019678570	0.010761600
gpt_4turbo-gpt_4o	0.06224900	-0.043823440	0.168321430	0.496195300
gpt_4turbo-gpt_4omini	0.18800000	0.082033900	0.293966100	0.000013413

Como parte de la Tabla 17 se encuentran en las filas los posibles cruces entre los diferentes modelos, mientras que para las columnas se detallan a continuación las descripciones:

- **diff**: la diferencia en la media de asignación entre los dos modelos comparados.
- **lwr** y **upr**: los límites inferior y superior del intervalo de confianza del 95% para la diferencia de medias. Si este intervalo no incluye 0, se puede considerar que hay una diferencia significativa entre los modelos comparados.
- **p adj**: el valor p ajustado para la comparación, que indica si la diferencia es estadísticamente significativa. Un valor p menor a 0.05 generalmente se considera significativo.

Con el fin de simplificar la comprensión de los resultados. se separan en dos grupos los modelos: aquellos en los que se demuestra la existencia de diferencias significativos y aquellos en los que no se comprueba diferencia significativa. Esta existencia se comprueba para

considerar la diferencia en las medias y su respectivo valor P ajustado. Se detalla, a continuación, cada grupo:

- Grupo en el que existen diferencias significativas: en este grupo se consideran aquellas combinaciones de modelos cuyo valor P ajustados no son mayores a 0.05. Los modelos que presentan el comportamiento donde sus medias son significativamente diferentes son:
  - gpt_4-gpt_35turbo
  - gpt_4o-gpt_35turbo
  - gpt_4omini-gpt_35turbo
  - gpt_4turbo-gpt_35turbo
  - gpt_4o-gpt_4
  - gpt_4turbo-gpt_4
  - gpt_4omini-gpt_4o
  - gpt_4turbo-gpt_4omini
- Grupo en el que no existen diferencias significativas: son aquellas combinaciones de modelo en donde su valor P es mayor a 0.05. Estas combinaciones son gpt_4omini-gpt_4 y gpt_4turbo-gpt_4o, lo cual sugiere que estos modelos poseen rendimientos similares en términos de asignación.

Una vez presentados los resultados obtenidos por la comparación de medias, además de la herramienta ANOVA y validado la variación con la prueba Tukey HSD, se comprueba la validación del a hipótesis, confirmando así que existe varianza en comportamiento entre los humanos y los modelos de IAG en la toma de decisiones altruistas, e incluso comprobando en cuales modelo existe aquella varianza.

5. Discusión, Limitaciones y problemas encontrados

En el presente capítulo se discuten los hallazgos obtenidos en relación con las preguntas de investigación formuladas en la sección 1.3 Planteamiento del problema. El mismo cuenta como se brinda respuestas a las preguntas planteadas al comienzo de la investigación.

5.1. ¿Cuál es el estado actual del uso de la Inteligencia Artificial Generativa en la simulación y análisis del comportamiento altruista?

En los últimos años, la IAG ha experimentado avances significativos en su capacidad para simular y analizar comportamientos humanos complejos, incluido el altruismo. En este tema, estudios recientes han demostrado que los modelos de lenguaje como GPT-3 y GPT-4 pueden participar en juegos estratégicos como el *Dictator Game*, mostrando comportamientos que reflejan preferencias altruistas similares a las humanas. Estos experimentos han sido esenciales para explorar cómo las IA toman decisiones en escenarios donde no solo se consideran los beneficios propios, sino también el bienestar de otros.

Sin embargo, aún existe una brecha entre la complejidad del comportamiento humano y lo que los modelos de IAG son capaces de replicar. Los estudios que emplean juegos como el *Dictator Game* han mostrado que, aunque las IAG pueden tomar decisiones que parecen altruistas, estas son respuestas a patrones aprendidos. Como consecuencia, la aptitud de los agentes generativos para tomar decisiones morales está altamente influenciada por los datos con los que fueron entrenados y por las limitaciones inherentes a los algoritmos que definen sus acciones.

El análisis del comportamiento altruista a través de la IAG ha abierto nuevas oportunidades para estudiar no solo los aspectos positivos del altruismo, sino también los sesgos y errores en la toma de decisiones. Un aspecto importante es cómo la IA puede manifestar comportamientos que son del todo racionales o por lo contrario son poco racionales. Este aspecto ha planteado preguntas cruciales sobre la robustez y fiabilidad de estos sistemas al simular comportamientos altruistas genuinos.

5.2. ¿En qué medida los agentes inteligentes emulan el comportamiento altruista humano en contextos de toma de decisiones?

Los agentes de los modelos de la familia de GPT-4 han demostrado una tendencia a dividir los recursos de manera equitativa, lo que refleja un comportamiento distinto al altruismo observado en humanos en situaciones de decisión. Esta tendencia hacia la equidad es un punto de partida para la emulación del altruismo.

Asimismo, el comportamiento altruista humano está influenciado por una serie de factores sociales y emocionales que los agentes de IAG no pueden replicar de manera auténtica: los humanos tienden a ajustar sus decisiones basados en la historia compartida con los otros participantes, sus expectativas o incluso sus emociones del momento, mientras que los agentes de IAG operan con lógica estrictamente basada en datos. Esta diferencia fundamental limita la capacidad de los agentes para emular completamente el altruismo humano en toda su complejidad. Además, en contextos donde las decisiones altruistas dependen de la historia

compartida o de factores sutiles que involucran interacciones previas, los agentes de IA no pueden emular completamente los matices que guían las acciones humanas.

Por último, aunque la emulación de decisiones altruistas por parte de la IAG es un campo en crecimiento, los resultados de los experimentos sugieren que las inteligencias artificiales generativas aún no son capaces de reproducir de manera fidedigna la toma de decisiones humana en contextos de altruismo.

5.3. ¿Qué elementos influyen en las decisiones altruistas tomadas por los agentes inteligentes generativos?

Las decisiones altruistas de los agentes de IAG están influenciadas por los datos de entrenamiento y los algoritmos que rigen su funcionamiento. Uno de los factores clave en su proceso de toma de decisiones es precisamente el conjunto de datos con los que han sido entrenados. Modelos como GPT-4 generan sus respuestas a partir de correlaciones y patrones presentes en esos datos, lo que implica que el comportamiento altruista observado en los experimentos es una combinación de ejemplos previos y no el resultado de una comprensión inherente del altruismo.

Adicionalmente, (OpenAI et al., 2023) explica que las técnicas de entrenamiento, como el aprendizaje por refuerzo con retroalimentación humana (RLHF) y el ajuste fino, mejoran la alineación del modelo con las expectativas humanas, optimizando su capacidad para seguir instrucciones y tomar decisiones basadas en criterios específicos. El contexto de la solicitud es otro factor importante, ya que la manera en que se formula una pregunta o se presenta un problema afecta directamente la interpretación y la respuesta generada. Por otro lado, el uso de técnicas de *prompting*, como el *few-shot prompting* o el *chain-of-thought prompting*, puede orientar al modelo hacia respuestas más precisas, afectando la naturaleza de las decisiones que toma.

Otro factor es la manera en que se formulan las preguntas o estímulos a los que los agentes responden. El *wording* o la redacción de las instrucciones puede tener un impacto significativo en las respuestas generadas por los modelos de IAG. Investigaciones han demostrado que pequeñas variaciones en el lenguaje utilizado para interactuar con los agentes pueden llevar a decisiones considerablemente diferentes, lo que sugiere que el comportamiento altruista de estos modelos está más relacionado con la forma en que se les presenta la situación que con un sentido genuino de altruismo.

Sin embargo, se deben tomar en cuenta las limitaciones inherentes al modelo, las cuales son explicadas en el artículo *GPT.4 Technical Report* de (OpenAI et al., 2023) incluyen su tendencia a generar información incorrecta o a cometer errores de razonamiento, factores que también desempeñan un papel relevante en la toma de decisiones. Estas limitaciones pueden llevar a resultados no deseados o erróneos en determinadas circunstancias. Además, los sesgos presentes en los datos de entrenamiento pueden influir en las decisiones del modelo, reflejando preferencias que pueden no ser siempre neutrales o justas, lo que podría afectar la equidad y objetividad de las respuestas generadas.

Finalmente, se debe considerar la interacción con el usuario y la retroalimentación obtenida de interacciones previas también incide en cómo el modelo ajusta sus respuestas en función de las preferencias y estilos del usuario, contribuyendo a la adaptabilidad del agente generativo.

Asimismo, la manera en que los agentes enfrentan la incertidumbre juega un papel fundamental en su comportamiento altruista. Mientras que los humanos a menudo toman decisiones altruistas en contextos inciertos, impulsados por la empatía y el deseo de ayudar, los agentes de IAG tienden a basarse más en la lógica y en predicciones optimizadas. Esto limita su capacidad para actuar de manera altruista en situaciones donde los beneficios no son inmediatos ni evidentes, subrayando la necesidad de avanzar en algoritmos que les permitan manejar mejor la incertidumbre y tomar decisiones más adaptativas.

5.4. Limitaciones y problemas encontrados

A lo largo del desarrollo de la investigación, se identificaron diversos problemas y obstáculos. El primero de ellos es el acceso exclusivo y de pago a las versiones de GPT-4 a través de su API. Aunque el costo no es elevado, representa una limitación al considerar los recursos disponibles para la investigación.

En segundo lugar, se destaca la escasa información sobre la estructura y el funcionamiento interno de los modelos de Inteligencia Artificial Generativa (IAG) desarrollados por OpenAI. La información publicada en línea sobre este tema es significativamente limitada, lo que afecta la comprensión de los comportamientos observados durante los experimentos.

Otro aspecto para considerar es la simplicidad del *Dictator Game* en comparación con la complejidad de las interacciones humanas reales. Este modelo es una representación simplificada que no incluye dinámicas recíprocas ni recompensas futuras, lo que restringe la capacidad de entender cómo la IAG podría comportarse en situaciones más complejas de cooperación o competencia, típicas de la teoría de juegos.

Además, como señala Anshu et al., (2021) existe una limitación relacionada con la adaptación de los modelos de teoría de juegos a la inteligencia artificial generativa. Los experimentos tradicionales de teoría de juegos están diseñados para agentes racionales que, en teoría, maximizan su utilidad con base en supuestos claros sobre los incentivos y preferencias de otros jugadores. Sin embargo, los modelos de IA, como GPT, no operan bajo un marco estrictamente racional ni poseen una noción innata de «maximización de utilidades». Esta falta de racionalidad limita la capacidad de generalizar los hallazgos hacia contextos más ricos en interacciones sociales, donde las decisiones altruistas pueden verse influenciadas por factores como la reputación, la reciprocidad y la competencia entre agentes.

Finalmente, el último punto por considerar es que los modelos de inteligencia artificial generativa carecen de conciencia y de una verdadera comprensión de las decisiones altruistas o estratégicas. Aunque tienen la capacidad de simular respuestas que se asemejan a comportamientos humanos, no están motivados por principios morales o consideraciones sociales, lo que reduce su capacidad para participar plenamente en experimentos de teoría de juegos que se basan en la interacción genuina entre agentes conscientes.

6. Conclusiones

En el presente capítulo se expondrán los resultados del proyecto de graduación. Se analizarán los hallazgos obtenidos, demostrando cómo se lograron los objetivos específicos planteados al inicio de la investigación.

6.1. Objetivo específico 1:

En relación con el objetivo específico 1: Analizar la situación actual sobre el uso de la Inteligencia Artificial Generativa (IAG) en la simulación y análisis del comportamiento altruista, se concluye lo siguiente:

- La aplicación de la inteligencia artificial generativa para estudiar el altruismo humano ha abierto nuevas posibilidades en la investigación social; sin embargo, persiste una brecha entre la complejidad del comportamiento humano y lo que estos modelos pueden replicar. Esta diferencia subraya la necesidad de un enfoque más sofisticado en el desarrollo de modelos de IAG que consideren las dinámicas emocionales y sociales del comportamiento altruista, como se constata en la sección 5.1.
- Según lo mostrado en la sección 5.1, los modelos de IAG pueden replicar ciertos patrones de comportamiento altruista. No obstante, sus limitaciones para comprender y simular factores emocionales como la historia compartida entre los individuos representan una barrera de comprensión al estudiar la replicación del altruismo humano.
- La capacidad de los modelos de IAG para simular comportamientos altruistas en entornos controlados, como el *Dictator Game*, ha proporcionado información valiosa sobre la toma de decisiones. Sin embargo, su aplicabilidad en la simulación del comportamiento altruista humano es limitada, ya que no pueden adaptarse a contextos más complejos ni incorporar las interacciones sociales profundas de la vida real. Esta limitación se discute en la sección 5.1, que subraya los desafíos que enfrentan estos modelos en escenarios complejos.

6.2. Objetivo Especifico 2:

En cuanto al segundo objetivo de investigación: Determinar cuáles son los criterios que afectan la interacción de los agentes inteligentes en la toma de decisiones altruistas, se determina que:

- La incapacidad de los modelos de IAG para integrar factores sociales y emocionales se refleja en la manera en la que las decisiones de los agentes se basan principalmente en patrones de entrenamiento. A pesar de que los modelos pueden imitar la equidad en la distribución de recursos, como se observó con el modelo GPT-4 que asignó una media de 4.776 euros, no logran capturar influencias clave como las experiencias previas de los participantes. Este punto se ve reflejado en la sección 5.2.
- La capacidad de los modelos de IAG para adaptarse a contextos complejos es limitada. La sección 4.2 muestra que los modelos optimizan resultados según patrones; esto significa

que en situaciones donde las recompensas no son fácilmente cuantificables, como la confianza o la cooperación, su desempeño se podría ver afectado.

- Los modelos de IAG, como GPT-3.5-turbo considerados el grupo de control del experimento, mostraron un valor modal de asignación de 4 euros en el *Dictator Game*, lo que indica que estos agentes tienden a preferir ciertas distribuciones de recursos. Esto se resalta en la sección 4.1.1 en donde se observa cómo las configuraciones internas influyen en las decisiones de los agentes, limitando su capacidad para reflejar la diversidad de elecciones humanas que tienden a depender de factores contextuales y emocionales.
- Tanto la forma en que se formulan las preguntas como el *wording* presentados a los agentes tiende a afectar significativamente sus decisiones. En las secciones 3.10.2.2 y 5.3 se muestra la manera en la que las variaciones en el lenguaje utilizado para interactuar con los modelos tienden a llevar a diferencias marcadas en sus respuestas. Este fenómeno sugiere que su comportamiento está condicionado por la presentación de la situación.
- La falta de comprensión de los factores sociales debido a los cambios implementados en la sesión 3.10.2.2 sobre los modelos de IAG resalta la necesidad de un enfoque más profundo que considere las dinámicas interpersonales. Esto se debe considerar para mejorar la emulación del comportamiento altruista en situaciones sociales.

6.3. Objetivo Especifico 3:

Con respecto al objetivo 3: Examinar en qué medida los agentes inteligentes emulan el comportamiento humano en contextos de toma de decisiones altruistas, se deduce que:

- Los resultados evidenciados en la sección 4.1 muestran que las decisiones generadas por los modelos de IAG, como GPT-3.5-turbo y GPT-4, tienden a seguir patrones de datos previamente entrenados. Estos se encuentran en la sección 4.1, donde se observan preferencias por valores modales específicos en sus asignaciones (3.7160 euros para GPT-3.5-turbo y 4.776 euros para GPT-4).
- El análisis ANOVA de la sección 4.2 confirma que existen diferencias estadísticamente significativas entre las asignaciones de diferentes modelos de IAG. Los resultados indican un valor F de 371.4 y un valor P menor a 0.001, lo que sugiere que la arquitectura del modelo influye de manera significativa en el comportamiento altruista.
- Según lo evidenciado en la sección 4.2 los agentes de IAG, como GPT-4-turbo, tienden a asignar recursos de manera más equitativa que los humanos, con un promedio de asignación cercano al 50%. Esto contrasta con los estudios de Engel (2010), que indican que los humanos suelen asignar solo el 30% de los recursos, sugiriendo una discrepancia en la forma en que ambos grupos manejan la equidad en la toma de decisiones.
- La tendencia de los modelos de IAG a asignar recursos equitativamente, como el promedio de 4.776 euros en GPT-4 reflejado en la sección 5.2, no refleja una comprensión de las sutilezas sociales que influyen en las decisiones humanas. Este patrón implica que los modelos de IAG siguen reglas generales de equidad, pero carecen de la flexibilidad necesaria para adaptarse a situaciones sociales complejas.

7. Recomendaciones para investigaciones futuras

Esta sección tiene como objetivo ofrecer recomendaciones basadas en los resultados del estudio general. El objetivo de estas recomendaciones es mejorar la comprensión y la aplicación de la inteligencia artificial en el contexto de la toma de decisiones altruistas, especialmente en experimentos como el *Dictator Game*.

7.1. Objetivo específico 1:

En relación con el objetivo específico 1: Analizar la situación actual sobre el uso de la Inteligencia Artificial Generativa (IAG) en la simulación y análisis del comportamiento altruista, se recomienda lo siguiente:

- Es necesario realizar más estudios experimentales con una variedad más amplia de contextos sociales y económicos para evaluar el comportamiento altruista de los modelos de IA en escenarios más cercanos a la realidad humana.
- Se recomienda desarrollar herramientas que permitan a los modelos de IA no solo emular patrones altruistas simples, sino también aprender y adaptarse a decisiones morales más complejas, teniendo en cuenta el entorno y la interacción social.
- Se recomienda realizar estudios longitudinales que evalúen cómo la exposición de los agentes de IA a entornos altruistas afecta su capacidad para adaptarse.
- Para mejorar la validez externa de los estudios, se sugiere que los experimentos no se limiten a los *Dictator Games*, sino que incluyan otros marcos experimentales como los juegos de confianza, dilemas sociales y juegos cooperativos, con el fin de evaluar el altruismo en situaciones más diversas.

7.2. Objetivo Especifico 2:

En cuanto al segundo objetivo de investigación: Determinar cuáles son los criterios que afectan la interacción de los agentes inteligentes en la toma de decisiones altruistas, se aconseja lo siguiente:

- Dado que la toma de decisiones de los agentes depende de los datos con los que han sido entrenados, es crucial mejorar la calidad y diversidad de estos datos. Se recomienda utilizar *datasets* más variados y representativos de diferentes contextos culturales, sociales y económicos para reducir los sesgos y aumentar la robustez del comportamiento altruista.
- Para mitigar los sesgos y mejorar la toma de decisiones altruistas, se recomienda implementar técnicas de depuración de datos y ajuste de algoritmos que aseguren una representación más equitativa y realista de las situaciones sociales.
- Se deben realizar ajustes en las metodologías experimentales, incluyendo la estandarización del *wording*, para evitar que variaciones mínimas en la redacción de instrucciones afecten los resultados de los experimentos.
- Se sugiere implementar métricas adicionales para evaluar no solo la decisión altruista final, sino también el proceso de toma de decisiones, con el fin de comprender mejor los razonamientos subyacentes en los agentes de IA.

7.3. Objetivo Especifico 3:

Con respecto al objetivo 3: Examinar en qué medida los agentes inteligentes emulan el comportamiento humano en contextos de toma de decisiones altruistas, se propone lo siguiente:

- Se debe continuar desarrollando modelos de IA con mayor capacidad para procesar factores contextuales y emocionales en la toma de decisiones altruistas con el fin de dar lugar a una emulación más cercana al comportamiento humano.
- Se recomienda realizar estudios longitudinales que permitan observar cómo los agentes de IA evolucionan en su toma de decisiones a lo largo del tiempo, comparando su progreso con el comportamiento humano en situaciones similares.
- Es importante evaluar si los agentes de IA muestran consistencia en su comportamiento altruista a lo largo de varias interacciones, tal como ocurre en seres humanos.
- Para mejorar la emulación del comportamiento humano se sugiere implementar agentes de IA con capacidades de aprendizaje continuo. Este avance permitiría refinar sus respuestas altruistas a medida que adquieren más experiencia y enfrentan situaciones más complejas en sus interacciones con humanos y otros agentes.

8. Referencias

Anshu, K., Singh, S. K., & Kumari, R. (2021). A Machine Learning Model for Effective Consumer Behaviour Prediction. *2021 5th International Conference on Information Systems and Computer Networks (ISCON)*, 1–5. https://doi.org/10.1109/ISCON52037.2021.9702495

Batson, C. D. (2010). *Altruism in Humans*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195341065.001.0001

Brookins, P., & Debacker, J. (2024). Playing Games With GPT: What Can We Learn About a Large Language Model From Canonical Strategic Games? *Social Science Research Network*.

Brühl, V. (2024). Generative Artificial Intelligence – Foundations, Use Cases and Economic Potential. *Intereconomics*, *59*(1), 5–9. https://doi.org/10.2478/ie-2024-0003

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. . *Science*, *356*(6334), 183–186.

Capraro, V., Paolo, R. D., Perc, M., & Pizziol, V. (2024). Language-based game theory in the age of artificial intelligence. *Journal of the Royal Society Interface*.

Daylamani-Zad, D., & Angelides, M. C. (2021). Altruism and Selfishness in Believable Game Agents: Deep Reinforcement Learning in Modified Dictator Games. *IEEE Transactions on Games*, *13*(3), 229–238. https://doi.org/10.1109/TG.2020.2989636

Engel, C. (2010). *Dictator games: a meta study*.

Fan, C., Chen, J., Jin, Y., & He, H. (2023). *Can Large Language Models Serve as Rational Players in Game Theory? A Systematic Analysis*. https://doi.org/10.48550/ARXIV.2312.05488

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), 785–791. https://doi.org/10.1038/nature02043

Fidias G. Arias. (2012). *El ProyEcto dE InvEstIgacIón Introducción a la metodología científica*.

Garcia, M. (2021). *Analyzing Behavioral Patterns in Human-AI Decision-Making Scenarios: A Literature Review.*

Grech, P., & Nax, H. (2018). Rational Altruism? On Preference Estimation and Dictator Game Experime nts. *Games Econ. Behav.*

Guala, F., & Mittone, L. (2010). Paradigmatic experiments: The Dictator Game. *The Journal of Socio-Economics*, *39*(5), 578–584. https://doi.org/10.1016/j.socec.2009.05.007

Guinn, C. I., & Palmer, D. (2014). Human perceptions of altruism in artificial agents. *2014 IEEE Symposium on Intelligent Agents (IA)*, 45–50. https://doi.org/10.1109/IA.2014.7009457

Helo-Guzmán, J. E. (2019). Redes neuronales y autómatas finitos. *Revista Tecnología en Marcha*, *13*(2), Pág 96-103. https://revistas.tec.ac.cr/index.php/tec_marcha/article/view/4149

Hernández-Sampieri R., & Mendoza, C. (2018). Metodología de la investigación. Las rutas cuantitativa, cualitativa y mixta. *Revista Universitaria Digital de Ciencias Sociales (RUDICS)*, *10*(18), 92–95. https://doi.org/10.22201/fesc.20072236e.2019.10.18.6

Hsu, J. C. (2010). *Multiple Comparisons: Theory and Methods*. Chapman & Hall/CRC.

Ileana Ulate, & Elizarda Vargas. (2016). *Metodología para elaborar una tesis* (EUNED, Ed.; Primera edición). EUNED.

Johnson, A. et al. (2020). Understanding Variability in Decision-Making: A Comparative Study of Human and AI Behavior in Social Dilemmas. *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

Johnson, T., & Obradovich, N. (2023). *Evidence of behavior consistent with self-interest and altruism in an artificially intelligent agent*.

Lim Ogawa Won Mi. (2023). *El impacto de soluciones basadas en Large Language Models en los negocios: una revisión sistemática de la literatura para identificar resultados, oportunidades y desafíos*. https://hdl.handle.net/2238/15044

Montgomery, D. C. (2019). *Design and Analysis of Experiments* (10a ed.). John Wiley & Sons.

ONU. (2015). *Objetivos y Metas de Desarrollo Sostenible - Desarrollo Sostenible. United Nations.* . https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/

OpenAI. (s/f-a). *GPT-4o-mini*. Recuperado el 6 de agosto de 2024, de https://platform.openai.com/docs/models/gpt-4o-mini

OpenAI. (s/f-b). *Models. OpenAI.* Recuperado el 6 de agosto de 2024, de https://platform.openai.com/docs/models

OpenAI. (2023, marzo 23). *GPT-4 System Card.*

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2023). *GPT-4 Technical Report*. *6*, 1–29.

Palomino-Flores, J. V., Saravia-Ramos, G. del P., & Palomino-Flores, R. I. (2023). Explorando la Intersección Entre Derechos Humanos e Inteligencia Artif icial. *Journal of Law and Sustainable Development*.

Patel, S. (2018). A Comparative Analysis of Decision-Making Processes in Humans and AI Agents: Insights from Behavioral Economics. *Cognitive Science Quarterly*.

Smith, J. (2019). Comparing Decision-Making Strategies Between Artificial Intelligence Agents and Humans in Economic Games. *Journal of Artificial Intelligence Research*.

Smith, V. L. (2008). *Rationality in Economics: Constructivist and Ecological Forms*. Cambridge University Press.

Song, X. (2022). Prediction of People's Abnormal Behaviors Based on Machine Learning Algorithms. *2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)*, 406–409. https://doi.org/10.1109/MLISE57402.2022.00087

TEC. (2022). *Qué es el TEC*. https://www.tec.ac.cr/que-es-tec

TEC. (2023). *Ejes de Conocimiento Estratégicos para el periodo comprendido de 2023 a 2032*. https://www.tec.ac.cr/ejes-conocimiento-estrategicos-2023-2032

Torres-Carballo, F., Morales-Rodríguez, N., Brenes-Leiva, G., & Solís-Salazar, M. (2018). Medición experimental del Comportamiento Organizacional Ciudadano: Altruismo, Aversión al Riesgo y Deportividad. *Revista Tecnología en Marcha*. https://doi.org/10.18845/tm.v31i4.3969

Wang, L. (2022). Exploring the Role of Emotions in Decision-Making: A Comparative Study of Human and AI Agents. *Journal of Cognitive Systems Research*.

Zhang, Y., Li, Y., Chen, X., & Xie, G. (2024). Emergence of Fairness Behavior Driven by Reputation-Based Voluntary Participation in Evolutionary Dictator Games. *IEEE Transactions on Computational Social Systems*, 1–10. https://doi.org/10.1109/TCSS.2023.3335396

9. Apéndices

9.1. Apéndice A: Plantilla minuta de reuniones

ID semana	Fecha	Hora/ Medio	Participantes	Temas tratados	Acuerdos

9.2. Apéndice B: Plantilla de gestión de cambios

Gestión de cambios: Solicitud de cambios	
ID Solicitud	##
Fecha	dd/mm/yy
Solicitante	Nombre y puesto
Descripción del cambio	Brindar el nivel más bajo de detalle posible alrededor del cambio.
Prioridad	Critico/Alto/Medio/Bajo
Impacto	Critico/Alto/Medio/Bajo
Firma	

9.3. Apéndice C: Plantilla de tabla comparativa de análisis documental.

Estudio	Resumen	Objetivos	Metodología	Principales descubrimientos
Estudio 1				
Estudio …n				

9.4. Apéndice D: Código para ejecución de experimentos

Vale-Martinez/Experimentos-DictatorGame-Altruismo: Análisis Comparativo del Comportamiento Altruista de la Inteligencia Artificial Generativa y Humanos: Un Estudio mediante Experimentos de Dictator Game (github.com)

9.5. Apéndice E: Tabla resultado de aplicación de experimentos.

Run Number	Total Amount	GPTResponse	Full Response	Error Message
1				
2				
…n				

9.6. Apéndice F: Proceso de ETL

```
# Asignamos el DataFrame de resultados a una variable local
df = dg_results

# Mantener solo las primeras 500 observaciones del DataFrame.
# Nota: Debido a interrupciones en el modelo, hay un total de 501 observaciones.
df = df[:500]

# Buscar la asignación en la respuesta generada por el modelo GPT
# Utiliza una expresión regular para extraer los dígitos (0-9) de la columna 'GPTResponse'
df.loc[:, "Allocation"] = (
    df["GPTResponse"].str.extract(r'(\d)',expand=False).astype(float)
)

# Reemplazar los valores de "50" por "5" en la columna Allocation para normalizar respuestas 50-50
df["Allocation"].replace(
    {50: 5}, inplace=True
)  # para los casos que la división haya sido 50-50


# Imprimir el DataFrame para ver la estructura y verificar el procesamiento
print(df)
```

Fuente: Experimentos-DictatorGame-Altruismo/GPT_tables_figuresFinal.py at main · Vale-Martinez/Experimentos-DictatorGame-Altruismo (github.com)

9.7. Apéndice G: Análisis estadístico

Experimentos-DictatorGame-Altruismo/GPT_Analsis_estadistico.py at main · Vale-Martinez/Experimentos-DictatorGame-Altruismo (github.com)

9.8. Apéndice H: Comprobación de hipótesis por medio de un análisis ANOVA

Experimentos-DictatorGame-Altruismo/ANOVA.R at main · Vale-Martinez/Experimentos-DictatorGame-Altruismo (github.com)

9.9. Apéndice I: Tabla comparativa de resultados

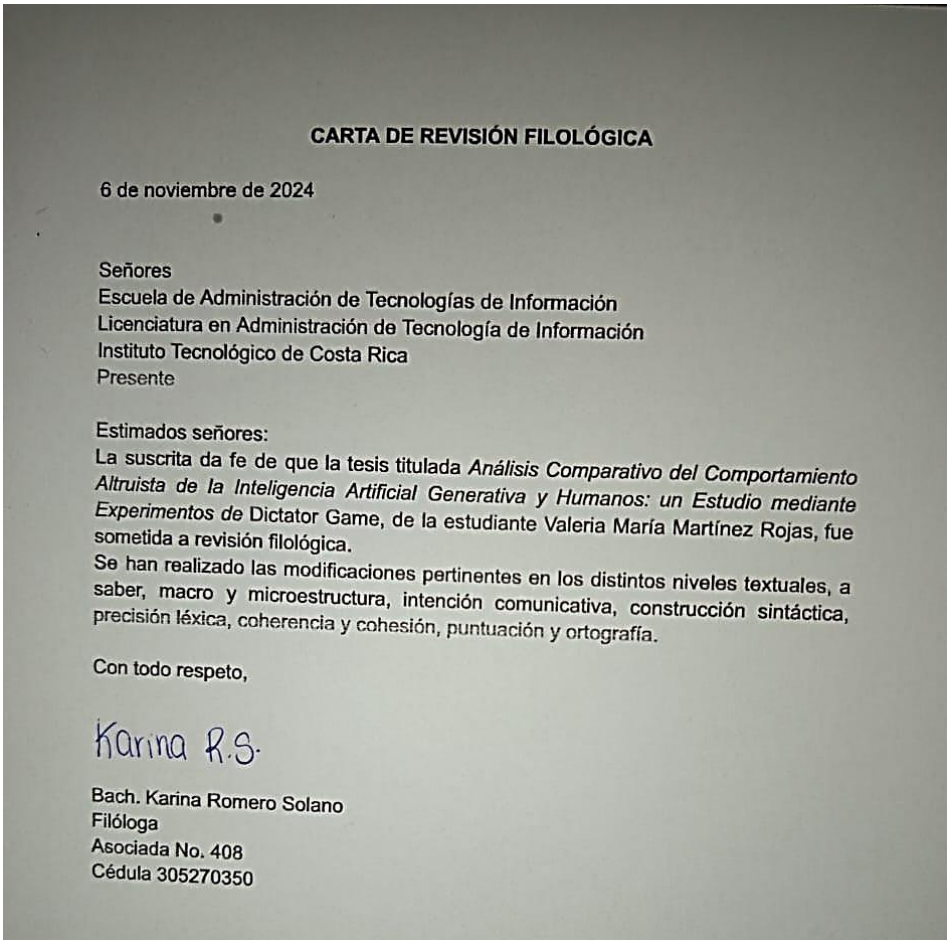
Modelo	Conteo	Media	Desviación estándar	Valor mínimo	25%	50%	75%	Valor Máximo
GPT-3.5-turbo								
GPT-4								
GPT-4-turbo								
GPT-4o								
GPT-4o-mini								
Engle								

9.10. Apéndice J: Tabla de minutas de reunión con profesor tutor

ID semana	Fecha	Hora/ Medio	Participantes	Temas tratados	Acuerdos
02	26/07/2024	2:00-3:00pm Google Calendar	Doc. Federico Torres Valeria Martínez	Revisión de Capitulo 1, presentado como anteproyecto. Revisión de herramientas para elaboración de Capitulo 2.	Aprobación de Capitulo 1.
03	2/08/2024	2:00-3:00pm Google Calendar	Doc. Federico Torres Valeria Martínez	Revisión de posibles estudios ya publicados para Capitulo 2. Revisión de Capitulo 3, secciones presentadas como anteproyecto.	Envío de estudios para revisión del profesor tutor. Revisión de correcciones al capitulo 3.
04	09/08/2024	2:00-3:00pm Google Calendar	Doc. Federico Torres Valeria Martínez	Revisión de avance en Capitulo 2 y 3.	Aplicación de correcciones al Capítulo 2. Elaboración de cronograma e investigaciones necesarias para continuar con los siguientes pasos.
05	14/08/2024	2:00-3:00pm Google Calendar	Doc. Federico Torres Valeria Martínez	Selección de Modelos para diseño y aplicación de experimentos para Capitulo 3.	Investigar sobre posibles estructuras para el diseño de experimentos.
06	23/08/2024	2:00-3:00pm Google Calendar	Doc. Federico Torres Valeria Martínez	Revisión de Capitulo 2 estado del arte. Articulo científico Revisión ejecución de pruebas preliminares para diseño de experimentos.	Aplicación de correcciones al Capítulo 2. Se selección el tema para el articulo científico como el estado del arte.
07	30/08/2024	2:00-3:00pm Google Calendar	Doc. Federico Torres Valeria Martínez	Revisión de resultados de pruebas experimentales. Revisión de avance para Capitulo 3	Se continua con pruebas de diseño experimental
08	06/09/2024	2:00-3:00pm Google Calendar	Doc. Federico Torres Valeria Martínez	Revisión y aprobación de resultados sobre diseño de experimentos.	Documentación de lo hallazgos hasta el momento. Aplicación de experimentos en modelos seleccionados.
09	13/09/2024	2:00-3:00pm Google Calendar	Doc. Federico Torres Valeria Martínez	Revisión y aprobación de resultados sobre aplicación de experimentos.	Descripción y exploración de hallazgos.

ID semana	Fecha	Hora/ Medio	Participantes	Temas tratados	Acuerdos
10	20/09/2024	2:00-3:00pm Google Calendar	Doc. Federico Torres Valeria Martínez	Revisión del proceso de descripción. Revisión de avance en articulo científico.	Documentación de hallazgos encontrados en Capitulo 4. Realización de análisis estadísticos para Capitulo 4. Continuar con avance para articulo científico.
12	02/10/2024	2:00-3:00pm Google Calendar	Doc. Federico Torres Valeria Martínez	Revisión de articulo científico y selección de análisis estadísticos adecuados	Se implementa el uso principal de una ANOVA.
13	11/10/2024	2:00-3:00pm Google Calendar	Doc. Federico Torres Valeria Martínez	Se revisan la estructura plantead para la presentación de resultados y la discusión	Se acuerda que la estructura planteada por la alumna es correcta y se aprueba.
14	18/10/2024	2:00-3:00pm Google Calendar	Doc. Federico Torres Valeria Martínez	Se revisa el articulo científico para detalles finales.	Se considera que esta listo para ser entregado.
15	25/10/2024	2:00-3:00pm Google Calendar	Doc. Federico Torres Valeria Martínez	Revisión de detalles finales sobre documento final Capítulos 4 y 5.	Pendiente retroalimentación completa por parte del profesor tutor.
16	01/11/2024	2:00-3:00pm Google Calendar	Doc. Federico Torres Valeria Martínez	Revisión final al documento y Capítulos 6 y 7	Pendiente aplicación de ajustes por parte de la estudiante.

9.11. Apéndice K: Carta de revisión filológica



CARTA DE REVISIÓN FILOLÓGICA

6 de noviembre de 2024

Señores
Escuela de Administración de Tecnologías de Información
Licenciatura en Administración de Tecnología de Información
Instituto Tecnológico de Costa Rica
Presente

Estimados señores:
La suscrita da fe de que la tesis titulada *Análisis Comparativo del Comportamiento Altruista de la Inteligencia Artificial Generativa y Humanos: un Estudio mediante Experimentos de* Dictator Game, de la estudiante Valeria María Martínez Rojas, fue sometida a revisión filológica.
Se han realizado las modificaciones pertinentes en los distintos niveles textuales, a saber, macro y microestructura, intención comunicativa, construcción sintáctica, precisión léxica, coherencia y cohesión, puntuación y ortografía.

Con todo respeto,

Karina R.S.

Bach. Karina Romero Solano
Filóloga
Asociada No. 408
Cédula 305270350

10. Anexos

10.1. Anexo I: Articulo científico

Análisis Comparativo del Comportamiento Altruista de la Inteligencia Artificial Generativa y Humanos: Un Estudio mediante Experimentos de *Dictator Game*

Valeria Martínez Rojas
Escuela de Administración de Tecnologías de Información
Tecnológico de Costa Rica
Cartago, Costa Rica
valemr@estudiantec.cr

***Abstract*— This paper examines the growing integration of artificial intelligence (AI) in various social domains, with a focus on altruistic behavior exhibited by large language models (LLMs) such as GPT-3.5-turbo. As these models evolve and gain sophistication, there is a need to investigate whether they can replicate the complexity of human behavior in moral decision-making, particularly in situations involving altruism. The study uses the Dictator Game, a classical experiment in behavioral economics, to analyze how LLMs respond in scenarios where an agent decides how to divide a sum of money between themselves and an anonymous recipient without expecting any reward. The key research question is whether different versions of GPT-3.5-turbo exhibit variation in altruistic decision-making. The comparison of GPT-3.5-turbo's decisions with human behavior offers valuable insights into AI-human interaction in moral contexts.**

***Keywords* —** Artificial Intelligence (AI), Large Language Models (LLMs), GPT-3.5-turbo, Altruistic behavior, Moral decision-making, Dictator Game, Behavioral economics, AI-human interaction, Altruism in AI**.**

I. INTRODUCCIÓN

Este ensayo explora la creciente integración de la inteligencia artificial (IA) a diversos ámbitos sociales, con un enfoque en el comportamiento altruista mostrado por los modelos de lenguaje de gran escala (LLMs) como GPT-3.5-turbo. A medida que estos modelos evolucionan y adquieren mayor sofisticación, surge la necesidad de investigar si cuentan con la capacidad de replicar la complejidad del comportamiento humano en situaciones que implican decisiones morales, en particular aquellas que requieren altruismo.

El estudio emplea el *Dictator Game*, un experimento clásico de la economía del comportamiento, para examinar la manera en la que los LLMs actúan en escenarios donde un agente debe decidir cómo dividir una cantidad de dinero entre sí mismo y un destinatario anónimo, sin esperar ninguna recompensa a cambio. En este contexto, la pregunta central del estudio es: ¿existe variación entre las diferentes versiones del modelo GPT-3.5-turbo en la toma de decisiones altruistas al utilizar el *Dictator Game*? Este interrogante busca descubrir si las actualizaciones entre versiones influyen en la capacidad del modelo para tomar decisiones altruistas.

El análisis de las decisiones tomadas por el modelo de GPT-3.5-turbo en el metaestudio seleccionado y su comparación con los resultados observados en humanos proporcionará una valiosa perspectiva sobre la interacción entre IA y comportamiento humano. [2]

II. ESTADO DEL ARTE

El estado del arte presentado en este estudio proporciona un marco teórico que enmarca la investigación actual y resalta la importancia de seguir explorando la intersección entre la inteligencia artificial y la economía del comportamiento.

A. Conceptos

En este apartado se definen los conceptos clave que sustentan la investigación y proporcionan una base teórica.

- **Altruismo**: comportamiento que implica actuar en beneficio de otros, a menudo sacrificando el propio bienestar. Es clave en el estudio de interacciones sociales y económicas [1].
- ***Dictator Games***: experimentos económicos que analizan el comportamiento altruista de un jugador al momento de decidir cómo dividir recursos entre sí mismo y otro jugador, permitiendo observar la tendencia a compartir [2].
- **Teoría de Juegos**: marco matemático que estudia interacciones estratégicas entre agentes racionales, utilizado para modelar comportamientos en contextos de cooperación y competencia [3].
- **Inteligencia artificial generativa**: modelos de IA que generan contenido nuevo a partir de patrones aprendidos, capaces de simular comportamientos humanos en la toma de decisiones [4].
- ***Wording / Prompts***: instrucciones o preguntas formuladas para guiar la generación de respuestas por modelos de lenguaje, cuya redacción tiene la capacidad de influir en la calidad de las respuestas generadas [5].

B. Estudios anteriores

A lo largo de la esta sección se analizarán los estudios previos que han investigado la implementación de los juegos del dictador y el comportamiento altruista utilizando agentes de inteligencia artificial generativa, en distintos escenarios y con distintos enfoques.

Playing Games with GPT: este estudio investiga cómo el modelo (GPT-3.5-turbo) interactúa en dos juegos estratégicos clásicos: el juego del dictador y el dilema del prisionero. Los autores comparan las decisiones del modelo con las de los humanos para entender las preferencias sobre equidad y cooperación integradas en la inteligencia artificial. [6].

El estudio presenta una serie de hallazgos por considerar para el desarrollo de la investigación en curso:

- El modelo GPT-3.5-turbo mostró una tendencia significativa hacia la equidad en el juego del dictador, optando frecuentemente por dividir los recursos de manera equitativa (50-50). De este modo, contrasta con la predicción racional de que el asignador debería retener todo.

Dictator games: a meta study: revisa más de cien experimentos de juegos de dictador realizados en los últimos 25 años. Utiliza regresión múltiple para evaluar el efecto de manipulaciones individuales en el comportamiento altruista y así proporcionar un conjunto de datos ricos para comparar modelos estadísticos en el análisis de decisiones en juegos de dictador.[2]

El estudio presenta una serie de hallazgos por considerar:

- Se encontró que las manipulaciones contextuales, como la presencia de un observador o la relación entre los jugadores, afectan significativamente el comportamiento altruista de los asignadores en el juego del dictador.
- El estudio también reveló que las decisiones altruistas no son homogéneas y varían considerablemente entre diferentes poblaciones y contextos, lo que sugiere que el comportamiento altruista es influenciado por factores culturales y situacionales.

III. METODOLOGÍA

El marco metodológico de este estudio proporcionó una estructura clara y sistemática que guio la investigación desde la conceptualización hasta la recopilación y el análisis de datos, con el fin de validar la pregunta de investigación planteada.

A. Población de estudio

La población de estudio se compuso de dos versiones del modelo GPT-3.5-turbo: 1) GPT-3.5-turbo-0301, que estableció la línea base de la investigación al ofrecer resultados consistentes con estudios previos sobre el *Dictator Game*, en particular el de A. Brookins y J. Debacker [6], y 2) GPT-3.5-turbo-0125, que presentó mejoras en la comprensión del lenguaje y la interacción contextual [7].

B. Pregunta de investigación

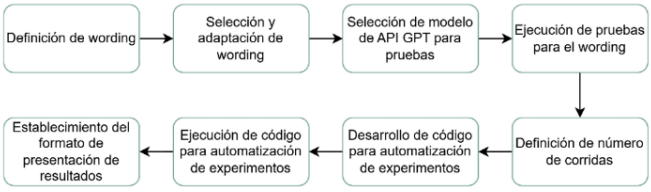
La investigación se centró en analizar la siguiente pregunta de investigación: ¿existe variación entre las diferentes versiones del modelo GPT-3.5-turbo en la toma de decisiones altruistas al utilizar el *Dictator Game*?

C. Diseño Experimental

En esta etapa se definió el diseño experimental que permitió evaluar el comportamiento altruista en sistemas de inteligencia artificial generativa. La metodología estableció los parámetros, variables y procedimientos específicos para medir y analizar las manifestaciones de altruismo en las interacciones de estos sistemas.

Figura 1

Actividades del Diseño experimental



Fuente: Elaboración propia, 2024

Como se mostró en la Figura 1, se definieron ocho actividades clave para el diseño de experimentos.

El primer paso para considerar fue la definición del *wording* o la redacción exacta que se utilizó para presentar las instrucciones a los agentes de inteligencia artificial generativa (IAG). Este término se refiere a la formulación precisa de las instrucciones o preguntas que se presentaron al modelo para guiar su comportamiento y asegurar que su interpretación fuera coherente con los objetivos del experimento.

El proceso de diseño del *wording* requirió una atención minuciosa a los detalles, por lo tanto, ara que el experimento fuera válido y replicable, fue necesario evitar cualquier ambigüedad en las instrucciones. En el caso de los modelos de lenguaje, que interpretan el ingreso textual de manera altamente contextual, incluso una ligera variación en la formulación pudo haber alterado la respuesta. Por esta razón, en este tipo de investigaciones, el *wording* debió ser neutral y preciso.

Posteriormente, se realizó un análisis detallado de estudios previos, considerando diferentes opciones de *wording* que hayan sido exitosas en experimentos similares, especialmente en el ámbito de juegos económicos como el *Dictator Game*. A partir de este análisis, se optó por un *wording* basado en investigaciones anteriores, adaptado a las necesidades específicas del proyecto, con el fin de garantizar la claridad de las instrucciones y la consistencia en la interpretación por parte del modelo GPT. Este análisis se detalla a lo largo de la sección «Estudios Anteriores», específicamente en el estudio P*laying Games With GPT: What Can We Learn About a Large Language Model From Canonical Strategic Games?* [6].

Contemplando las características anteriores, se definió el *wording* por utilizar como:

> *"This task is about dividing money between yourself and another person to whom you are randomly matched. You do not know this other person and you will not knowingly meet him/her.*
>
> *You have been randomly assigned the role of the 'allocator'. The other person is in the role of the 'recipient'.*
>
> *You are endowed with 10 euro, the recipient is endowed with 0 euro. You can decide how much of your*

10 euro endowment to transfer to the recipient. You can choose any amount between 0 euro and 10 euro.

The recipient receives the amount that you decide to transfer to him/her; you receive the amount that you decide not to transfer and thus to keep.

How much of your 10 euro endowment do you want to transfer to the recipient? Just tell me the allocation, not your reasoning." [6]

Posteriormente, se llevaron a cabo pruebas con el modelo GPT-3.5-turbo para evaluar el *wording* seleccionado. En cuanto a la ejecución técnica de estas pruebas, se trabajó en el desarrollo del código necesario para automatizar el proceso (detallado en el Apéndice A : Código para ejecución de experimentos).

El lenguaje de programación seleccionado fue Python, debido a su compatibilidad con el API (Application programming interface) de OpenAI y su flexibilidad para manejar grandes volúmenes de datos. Este código no solo permite la interacción con el modelo de IA, sino que también facilita la selección y modificación del *wording* de manera eficiente.

En cuanto a la cantidad de pruebas por realizar, se definió que el número de corridas es 500, según las pruebas realizadas en el estudio previo de Brookins, P., & Debacker, J. [6] Este número permitió asegurar que los resultados obtenidos fueran representativos y que cualquier variación observada fuera estadísticamente significativa. Además, el volumen de datos ayudó a identificar patrones o inconsistencias en las respuestas del modelo, lo cual permitió garantizar la validez de la pregunta de investigación.

Por último, se estableció el formato en el que se almacenarían los resultados de las pruebas, como se verifica en el Apéndice B: Tabla resultado de aplicación de experimentos. Se optó por un archivo con extensión .xlsx, distribuido en cinco columnas: *Run Number, Total Amount, GPT Response, Full Response y Error Messa*ge. A continuación, se detalla cada una de ellas:

- **Run Number** indicará el número de ejecución de la prueba.
- **Total Amount** reflejará la cantidad total asignada en el experimento, que será fija en 10 unidades para todos los casos.
- **GPT Response** contendrá la respuesta generada por el modelo IAG.
- **Full Response** almacenará una representación en formato JSON de la respuesta completa, incluyendo detalles relevantes para su análisis.
- **Error Message** se utilizará para registrar cualquier error que ocurra durante la ejecución de las pruebas.

Esta estructura de almacenamiento permitió organizar y analizar los resultados de manera efectiva, facilitando la posterior interpretación de los datos.

A partir de la implementación de las pruebas y la configuración del entorno experimental, fue posible avanzar hacia la fase de análisis de datos. En esta etapa se evaluó la eficacia del *wording* seleccionado y el desempeño de los modelos de IA frente a las instrucciones presentadas. Posterior a esta etapa, se realizó un análisis comparativo, donde se consideraron los resultados obtenidos por el modelo GPT-3.5-turbo y el estudio *Dictator games: a meta study*, del autor Engle [2].

IV. Aplicación de Experimentos

Después de haber diseñado los experimentos, se procede a su ejecución con el objetivo de obtener resultados que permitan evaluar el comportamiento del modelo.

A. Configuración del Modelo de GPT-3.5-Turbo

La versión específica del modelo GPT-3.5-turbo utilizada para los experimentos aplicados entre las fechas 2 de setiembre del 2024 y 9 de setiembre del 2024 es la versión del modelo es la GPT-3.5-turbo-0125, considerada la última actualización del modelo con fecha de 25 de enero del 2024. [7] La selección de esta versión del modelo permite aprovechar al máximo las capacidades de procesamiento de lenguaje natural, lo que garantiza resultados consistentes en experimentos masivos.

La configuración del entorno define la manera en la que el modelo interactuará durante el experimento, lo cual no se limita únicamente a parámetros ajustables como la temperatura o el rol del sistema, sino que también abarca aspectos como la longitud máxima de la respuesta, la naturaleza de la conversación simulada y el manejo de recursos computacionales durante ejecuciones largas. Además, la versión GPT-3.5-turbo-0125 incluye mejoras en la gestión de memoria y la comprensión contextual, lo que permite que el modelo maneje múltiples interacciones consecutivas de manera fluida [7].

B. Parámetros Claves

Dentro de la configuración de los experimentos, es crucial considerar dos parámetros clave: la *temperature* y el rol de *system*, ya que ambos influyen directamente en el comportamiento del modelo y en los resultados obtenidos.

Para el parámetro de temperatura, según la documentación del API este parámetro se define como «la temperatura tiene valores a utilizar, entre 0 y 2. Valores más altos como 1,8 harán que la salida sea más aleatoria, mientras que valores más bajos como 0,2 la harán más centrada y determinista». [8] Para la configuración de los experimentos, este valor de mantiene en el predeterminado que sería 1. El rol de *system*, por su parte, se establece en los mensajes de configuración que forman parte de la «conversación» mantenida con el modelo. En este caso, se asigna el rol de «*undergraduate student*», que guía al modelo para que asuma un perfil específico al responder. Esta configuración permite moldear las respuestas del modelo bajo un contexto definido, asegurando que las respuestas reflejen un enfoque acorde con el rol asignado, lo cual es esencial para mantener la coherencia y el rigor en el contexto experimental.

C. Ejecución de los experimentos

Para la aplicación de los experimentos ejecutados en el presente estudio, se desarrolla el código de ejecución de las pruebas automatizadas descrito en la sesión de «Diseño de los experimentos» descrito en el apartado de metodología. El código en cuestión, ubicado en el Apéndice A : Código para ejecución de experimentos, busca implementar un sistema automatizado para ejecutar el *Dictator Game*, un experimento económico en el que un *allocator* decide cuánto

dinero transferir a un destinatario. El código genera un *prompt* en cada iteración, a partir del cual el modelo decide cuánto dinero transferir a un destinatario. Las pruebas se repiten 500 veces y cada respuesta se registra en un archivo éxcel junto con detalles como el número de ejecución, el monto total y la respuesta completa en formato JSON.

V. RESULTADOS

Esta sección presenta los resultados obtenidos después de realizar experimentos utilizando el modelo GPT-3.5-turbo en el contexto del juego *Dictator Game*. Se incluye un análisis estadístico de las respuestas generadas por el modelo, en el que se examinan aspectos importantes como la distribución de los montos transferidos a los destinatarios y su consistencia entre diferentes series.

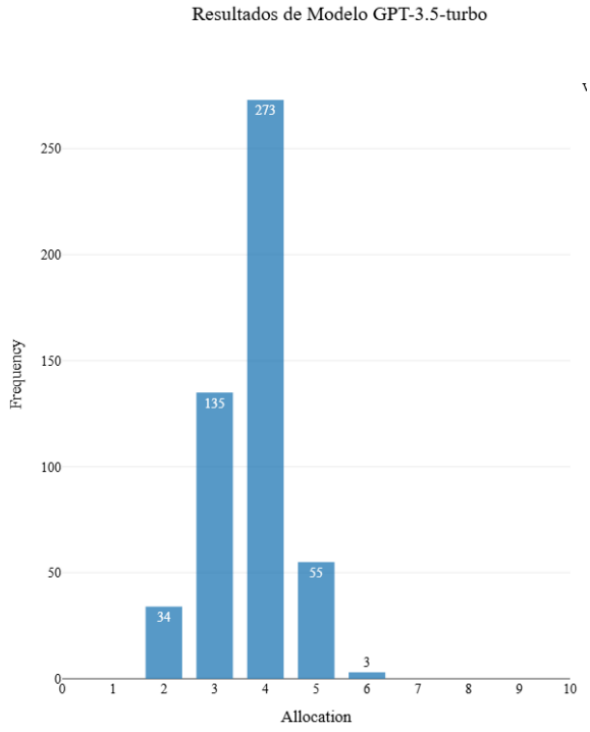
A. Resultados Obtenidos

Este apartado detalla la manera en la que el modelo distribuyeron los recursos entre el dictador y el destinatario, y resalta las variaciones en las cantidades asignadas en diferentes ejecuciones. Adicionalmente, se evaluó el modelo a lo largo de múltiples ejecuciones, con énfasis en identificar patrones de comportamiento repetitivos o cambios bruscos en las decisiones.

Con el fin de ilustrar visualmente los resultados obtenidos, se empleó el histograma de la Figura 2, en el que se muestra la distribución de las cantidades transferidas. Este instrumento permite detectar la frecuencia con la que el modelo opta por ciertas cantidades y destacar tanto los valores extremos como las concentraciones más frecuentes. De esta manera, se identifican posibles tendencias en el comportamiento del modelo, que proporcionan una visión clara de su proceso de toma de decisiones en el experimento.

Figura 2

Histograma de resultados modelo GPT-3.5-turbo



Como se logra visualizar en la Figura 2 Histograma de resultados modelo GPT-3.5-turbo, en la distribución obtenida por la ejecución de los experimentos con 500 rondas, una gran cantidad de resultados, específicamente 273, asignaron 4 euros como respuesta; seguido de 135 respuestas con 3 euros; 55 respuestas con 5 euros; 34 respuestas con 2 euros; y 3 respuestas con 6 euros. Por último, las asignaciones de 0, 1, 7, 8, 9 y 10 euros no recibieron respuestas.

Así mismo, en la Figura 2. Histograma de resultados modelo GPT-3.5-turbo, se observa una clara tendencia hacia las asignaciones de 4 y 3 euros respectivamente, las cuales abarcan en conjunto alrededor del 75% de las respuestas registradas.

B. Análisis estadístico de las Respuestas del Modelo GPT-3.5-Turbo

En esta sesión se presenta el análisis estadístico de las respuestas generadas por el modelo. Dichas respuestas permiten comprender mejor los patrones de comportamiento en la asignación de recursos, a partir del uso de herramientas estadísticas para analizar las distribuciones de los montos transferidos, promedios del modelo. Es así como este enfoque cuantitativo proporciona una visión sobre el modo en el que el modelo interpreta las instrucciones y asigna los recursos de manera predictiva y consistente.

Para profundizar en este análisis, se realizó un cuadro comparativo que incluye las siguientes estadísticas clave: conteo, media, desviación estándar, valor mínimo, percentiles (25%, 50%, 75%), valor máximo y moda. Dicho cuadro facilita la evaluación de la consistencia y variabilidad en las respuestas del modelo, a la vez que permite detectar tendencias en las decisiones y entender la fluctuación del reparto de recursos en diferentes ejecuciones.

Tabla 1

Análisis estadístico de resultados modelo GPT-3.5-turbo

Estadística	Cálculo
Frecuencia	500 respuestas
Media	3.716 euros
Desviación estándar	0.772 euros
Valor mínimo	2 euros
25%	3 euros
50%	4 euros
75%	4 euros
Valor máximo	6 euros
Moda	4 euros

Los resultados de la Tabla 1 presentan el resumen del análisis estadístico de las respuestas obtenidas del modelo GPT-3.5-turbo. Esta revela que la media de las asignaciones es de 3.716 euros. Este resultado sugiere una tendencia hacia el altruismo moderado, indicando que, en promedio, los participantes están dispuestos a compartir recursos de manera razonable.

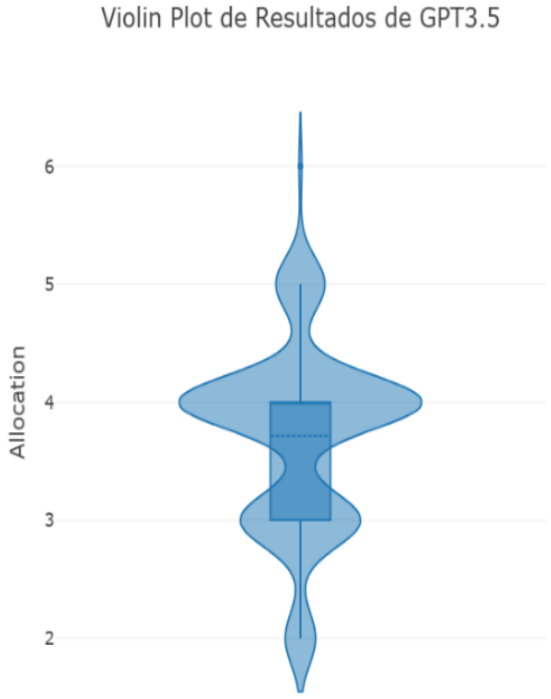
Por otro lado, la desviación estándar de 0.772358 euros refleja una distribución relativamente homogénea de las decisiones, lo que significa que la mayoría de los individuos tomaron decisiones similares al realizar sus asignaciones. Los valores mínimos (2 euros) y máximo (6 euros) demuestran un rango variado en las decisiones y evidencian que mientras algunos participantes optaron por asignaciones bajas, otros fueron generosos.

En cuanto a los percentiles, estos valores refuerzan la inclinación hacia asignaciones alrededor de 4 euros, con el 25% asignando 3 euros, el 50% a 4 euros y el 75% también situándose en este valor. Es así como a moda, que es de 4 euros, indica que este es el valor más frecuente entre las decisiones.

Con el fin de comprender mejor los resultados obtenidos de la Tabla 1 de este apartado, se utiliza la representación gráfica a partir del gráfico de violín (Figura 3). Este tipo de gráfico combina las características de un gráfico de caja (*box plot*) con un gráfico de densidad, proporcionando una representación visual tanto de la dispersión como de la densidad de los datos.

Figura 3

Gráfico de Violín de resultados modelo GPT-3.5-turbo



Tal como se muestra en la Figura 3, se manifiesta la existencia de una simetría, lo que indica que no existe un sesgo claro hacia asignaciones extremadamente bajas o altas. La simetría sugiere que las asignaciones del modelo están distribuidas de manera equilibrada a ambos lados del valor central, con una mayor concentración de valores entre 3 euros y 5 euros. El pico principal de la distribución que se observa en la Figura 3 se ubica en el centro, alrededor del valor de 4 euros, con una leve tendencia hacia una distribución bimodal y un segundo pico más pequeño alrededor del valor de 3 euros.

Respecto a la distribución de las colas del gráfico, estas representan las asignaciones extremas. Según lo observado en la Figura 3, no hay una clara predilección por valores extremos, ya que el modelo rara vez asigna valores fuera del rango principal descrito anteriormente. Las colas muestran una menor densidad en los valores por debajo de 3 euros y por encima de 5 euros, lo que sugiere que las asignaciones extremas son poco frecuentes. Es así como se observa que el comportamiento en la gráfica de la Figura 3 es consistente con el valor asociado a la desviación estándar de 0.772 euros.

C. Comparación con Estudios Previos

1) Playing Games with GPT

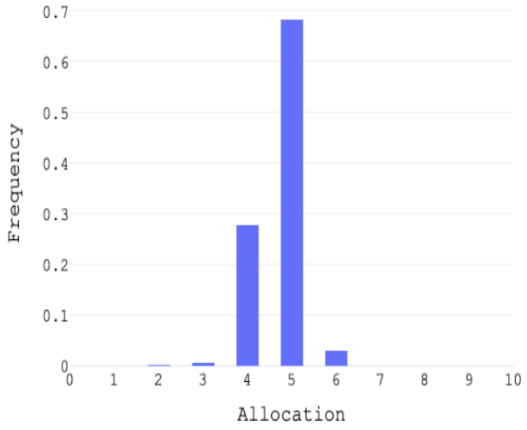
Este trabajo proporciona una base sólida para evaluar las decisiones del modelo GPT-3.5-turbo, específicamente la versión GPT-3.5-turbo-0301, publicada el primero de marzo del año 2023, frente a las respuestas obtenidas en experimentos con participantes humanos.

Para contextualizar los resultados obtenidos, en esta sección se realiza una comparación con estudios previos sobre el *Dictator Game*, tomando como referencia el estudio de A. Brookins y J. Debacker [6].

La comparación de los resultados permite identificar puntos de convergencia y divergencia, proporcionando una visión amplia sobre la capacidad del modelo para emular decisiones en situaciones económicas y sociales. De esta manera, se evalúa hasta qué punto el modelo GPT-3.5-turbo logra replicar comportamientos observados en estudios empíricos, como la distribución de recursos en el *Dictator Game*, a la vez se reflexiona sobre sus limitaciones o ventajas en el proceso de simulación de decisiones altruistas.

Figura 4

Resultados del estudio Playing Games With GPT: What Can We Learn About a Large Language Model from Canonical Strategic Games.



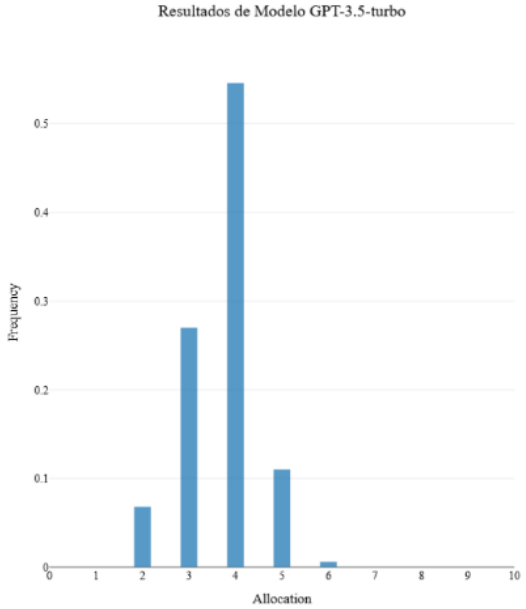
Nota. Adaptado de *Playing Games With GPT: What Can We Learn About a Large Language Model from Canonical Strategic Games*, por P. Brookins y J. Debacker, 2023.

Como se muestra en la Figura 4, los resultados del estudio *Playing Games With GPT* destacan que el 70% de las asignaciones fueron de 5 euros [6], en contraste con los datos obtenidos en el presente estudio (Figura 5) a partir del modelo GPT-3.5-turbo-0125, que asignó únicamente el 11% al valor de 5 euros. Esto implica un alejamiento de la noción de equidad basada en una distribución 50-50, como se observó inicialmente en la versión GPT-3.5-turbo-0301.

Un descubrimiento considerablemente interesante es la diferencia temporal de aproximadamente nueve meses entre las versiones estudiadas: GPT-3.5-turbo-0301, publicada el 1 de marzo de 2023, y GPT-3.5-turbo-0125, lanzada el 25 de enero de 2024. [7] Además, entre ambas versiones existen cinco actualizaciones intermedias: GPT-3.5-turbo-16k-0613, GPT-3.5-turbo-0613, GPT-3.5-turbo-instruct, GPT-3.5-turbo-16k y GPT-3.5-turbo-1106.

Figura 5

Resultados modelo GPT-3.5-turbo en porcentajes



De manera similar, otro de los hallazgos relevantes es la asignación de 4 euros. En el gráfico de la Figura 4 se observa que no más del 30% de las respuestas corresponden a este valor, mientras que en la Figura 5 se presenta un 55% de las respuestas asignadas a 4 euros. Asimismo, las diferencias entre los modelos son notables. En el estudio *Playing Games With GPT: What Can We Learn About a Large Language Model from Canonical Strategic Games* de la Figura 4, se observa que aproximadamente el 90% de las asignaciones se concentraron en los valores de 4 y 5 euros. [6] En cambio, en el presente estudio, reflejado en la Figura 5, ese mismo 90% se distribuye entre los valores de 3, 4 y 5 euros. A pesar de estas diferencias, ambos modelos comparten la similitud de que las asignaciones altas se concentraron en 6 euros, es decir, no registraron respuestas para los valores de 7, 8, 9 o 10 euros, ni para los valores bajos, de 0 o 1 euro.

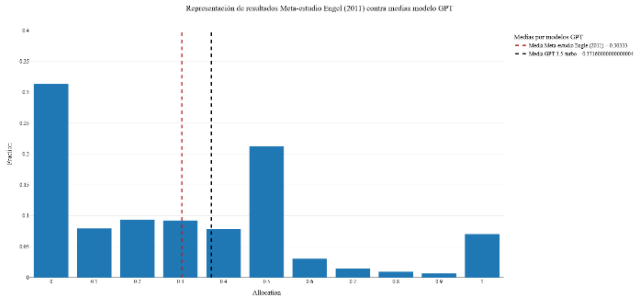
Los análisis del modelo GPT-3.5-turbo en este trabajo evidencian una desviación respecto a las versiones previas y estudios anteriores en términos de equidad. Mientras que los resultados anteriores se acercaban a una distribución 50-50, el presente estudio revela una mayor concentración en las asignaciones de 4 euros, lo que indica una menor inclinación hacia la equidad tradicional. No obstante, ambos estudios coinciden en la evasión de asignaciones extremas, sugiriendo que el modelo tiende a tomar decisiones moderadas y elude tanto el altruismo máximo como el egoísmo absoluto.

2) Dictator games: a meta study

Uno de los aspectos de la presente investigación es comparar el comportamiento del modelo con los resultados obtenidos en el meta-estudio de Engel, que recoge datos de múltiples estudios realizados con seres humanos. [2] Esta comparación busca explorar la capacidad del modelo GPT-3.5-turbo para emular la toma de decisiones altruistas que involucran distribuciones monetarias y, en particular, evaluar si el modelo sigue los patrones de equidad o si su comportamiento difiere de las tendencias observadas en humanos.

Figura 6

Resultados del estudio Dictator games: a meta study, *resaltando la media de los resultados con GPT-3.5-turbo-0125.*



Nota. Adaptado de *Dictator games: a meta stud*y, 2011.

En la Figura 6 se muestra una representación gráfica que compara los resultados del metaestudio de Engel con aquellos obtenidos en los experimentos realizados en este estudio, utilizando el modelo GPT-3.5-turbo-0125 con la prueba del *Dictator Game*. [2] Esta visualización facilita el contraste entre las decisiones del modelo GPT y las tendencias identificadas en el estudio de Engel, revelando así los patrones de asignación en ambos contextos.

La línea punteada negra representa la media de los datos obtenidos en el estudio aplicado con el modelo GPT-3.5-turbo-0125, que es aproximadamente 37.16%, lo que indica que el modelo asigna, en promedio, un 37% del total al receptor. Por otro lado, la línea punteada roja corresponde a la media de los datos del metaestudio de J. Engel, que es aproximadamente 30.33%. Este valor refleja que, en promedio, los participantes humanos en los estudios revisados por Engel asignaron alrededor del 30% de la cantidad total. [2]

Un punto por destacar en el gráfico es que en el metaestudio de Engel, una gran parte de las asignaciones se concentra en la fracción 0, lo que sugiere una tendencia hacia el egoísmo o falta de altruismo, mientras que otra porción considerable se encuentra en la fracción 0.5, indicando decisiones equitativas. [2] En contraste, los resultados obtenidos con GPT-3.5-turbo-0125 muestran una mayor dispersión, alejándose de los patrones de equidad observados en estudios humanos previos. Esto sugiere una diferencia en el comportamiento del modelo en comparación con las decisiones humanas en el *Dictator Game*.

D. Validación de la pregunta de investigación

La pregunta de investigación planteada, «¿existe variación entre las diferentes versiones del modelo GPT-3.5-turbo en la toma de decisiones altruistas al utilizar el *Dictator Game*?», resulta pertinente y válida al analizar la capacidad del modelo para simular comportamientos altruistas a través de decisiones de asignación de recursos.

Los datos recopilados en el apartado A de este capítulo muestran una clara tendencia hacia asignaciones de 4 euros y 3 euros, valores en los que en conjunto comprenden el 75% de las decisiones. Esta distribución sugiere un comportamiento altruista moderado, ya que la media de las asignaciones es de 3.716 euros, con una desviación estándar de 0.772 euros; esto indica que las decisiones son relativamente consistentes.

En comparación con versiones anteriores descritas en el apartado C de este capítulo, se observa que la versión GPT-3.5-turbo-0125 presenta una diferencia notable en la asignación de recursos. En un análisis comparativo, el 55% de las respuestas se concentraron en 4 euros, en contraste con la versión GPT-3.5-turbo-0301, donde menos del 30% de las asignaciones se registraron en este valor. [6] Además, en la versión anterior, el 70% de las asignaciones fueron de 5 euros, mientras que, en la versión actual, solo el 11% optó por este valor. Esto refleja un cambio en la estrategia de asignación y una menor inclinación hacia la equidad tradicional observada en estudios previos.

VI. DISCUSIÓN

En esta sección se discute y profundiza tanto en la relevancia de los resultados obtenidos como en las implicaciones para la comprensión de la toma de decisiones altruistas en humanos y en sistemas de IAG, destacando las diferencias clave y el potencial de estos modelos en la simulación de comportamientos sociales.

A. Desempeño del Wording en las Respuestas Generadas

El primer punto para recalcar en la discusión es el tema del *wording* o la redacción de las instrucciones. Esta característica de los experimentos con inteligencia artificial generativa juega un papel relevante en la forma en la que los modelos interpretan y responden a las tareas asignadas.[5] En este estudio, el *wording* se diseñó cuidadosamente para ser claro, neutral y sin ambigüedades, con el fin de evitar influencias no deseadas en las decisiones del modelo GPT-3.5-turbo. A pesar de estos esfuerzos, hubo pequeñas variaciones en la redacción de las instrucciones, lo que llevó a respuestas significativamente diferentes e, incluso, no acertadas, lo que demuestra la sensibilidad del lenguaje. Este hallazgo se alinea con estudios previos que destacan cómo los modelos de lenguaje procesan los datos en función del contexto textual inmediato, lo que refuerza la importancia de un *wording* preciso y consistente. [6]

En el caso particular de este experimento, el *wording* fue diseñado con el fin de replicar escenarios reales de decisiones altruistas en el *Dictator Game*, experimento en el que se presenta al modelo como un «asignador» que debe dividir una cantidad fija de dinero entre él mismo y un destinatario anónimo. La decisión de enfatizar el anonimato y la falta de interacción futura entre las partes influyó en las respuestas generadas.

La redacción neutral permitió observar una tendencia clara hacia decisiones que no se inclinan hacia el egoísmo absoluto, pero que tampoco alcanzan una equidad completa (50-50). Esto sugiere que el *wording* utilizado logró su objetivo de evitar sesgos extremos en las respuestas, aunque no necesariamente promovió una inclinación hacia la equidad. Sin embargo, a pesar de los esfuerzos para controlar el *wording*, los resultados también demuestran que el modelo GPT-3.5-turbo responde de manera predecible dentro de ciertos márgenes, lo que sugiere que el *wording* no tiene la capacidad de eliminar por completo las tendencias inherentes del modelo. Es posible que factores internos, como el entrenamiento previo y las preferencias integradas en el modelo, también jueguen un papel importante en sus decisiones. En esta línea, aunque el *wording* utilizado fue efectivo para estandarizar las condiciones experimentales, los resultados sugieren que el comportamiento del modelo no solo depende de las instrucciones recibidas, sino también de su capacidad para interpretar y balancear los valores de cooperación y altruismo bajo las condiciones presentadas.

B. Impacto de las Actualizaciones Arquitectónicas en el Comportamiento de Modelos de IA

Otro punto para discutir es cómo los cambios en la arquitectura de los modelos GPT-3.5-turbo tienen la posibilidad de impactar en el comportamiento en tareas específicas, como es el caso del *Dictator Game*. Las discrepancias observadas entre las versiones GPT-3.5-turbo-0301 y GPT-3.5-turbo-0125, particularmente en la inclinación hacia la equidad en la asignación de recursos, sugieren que las actualizaciones arquitectónicas internas tienden a influir directamente en la toma de decisiones. Es así como la concentración de respuestas en asignaciones bajas en la versión reciente, GPT-3.5-turbo-0125, en la que se señala un posible ajuste en el comportamiento del modelo, tiene la posibilidad de afectar su capacidad para reflejar comportamientos altruistas. Este fenómeno plantea un desafío en la investigación, ya que la falta de transparencia en los cambios arquitectónicos de modelos de OpenIA dificulta la identificación precisa de las causas de estas variaciones. Los ajustes en parámetros de entrenamiento, estructuras internas o, incluso, procesos de optimización podrían ser los responsables, pero sin documentación detallada no cuentan con la capacidad de evaluar de manera concluyente cómo influyen estos factores en las decisiones del modelo. Esto también impacta la replicabilidad de estudios que dependen de un comportamiento consistente de la IA.

C. Reflexiones sobre estudios futuros

Este estudio resalta la complejidad de la toma de decisiones altruistas, tanto en humanos como en sistemas de inteligencia artificial generativa (IAG). Los hallazgos sugieren que las decisiones altruistas no solo dependen de la configuración de los escenarios experimentales, sino también del *wording* y de las sutilezas en la formulación de las instrucciones, que influyen de manera significativa en los resultados obtenidos.

A pesar de que los modelos de IAG tienen la aptitud de simular comportamientos altruistas, su falta de una comprensión consciente de los valores éticos y sociales limita su capacidad para participar en interacciones genuinas. Esto plantea interrogantes sobre la efectividad de estos modelos al abordar cuestiones que requieren empatía, responsabilidad y una consideración profunda de las implicaciones sociales de las decisiones. Por esta razón, la investigación futura debe dirigirse a la creación de escenarios experimentales que reflejen las dinámicas de las interacciones humanas reales relacionadas al tema del altruismo.

En última instancia, es crucial realizar comparaciones longitudinales entre diferentes versiones de modelos para identificar patrones y tendencias consistentes en la toma de decisiones altruistas. La transparencia en las actualizaciones de los modelos será esencial para comprender las capacidades y limitaciones de la inteligencia artificial en este contexto.

D. Limitaciones y Desafíos

El diseño experimental basado en la teoría de juegos, específicamente el *Dictator Game*, presenta ciertas limitaciones cuando se aplica a la evaluación de inteligencia artificial generativa, como GPT-3.5-turbo. Por lo tanto, una

de las principales limitaciones radica en la simplicidad del juego en comparación con la complejidad de las interacciones humanas reales.

El *Dictator Game* es un modelo simplificado que no incluye dinámicas recíprocas ni recompensas futuras, lo que tiende a restringir la comprensión sobre el modo en el que la IAG podría comportarse en situaciones complejas de cooperación o competencia, típicas de la teoría de juegos. Esto limita la capacidad de generalizar los hallazgos hacia contextos ricos en interacciones sociales, en los que las decisiones altruistas tienen la posibilidad de ser influenciadas por factores como la reputación, la reciprocidad o la competencia entre agentes.

Otra limitación importante está relacionada con la adaptación de los modelos de teoría de juegos a la inteligencia artificial generativa. Los experimentos tradicionales de teoría de juegos están diseñados para agentes racionales que, en teoría, maximizan su utilidad basándose en supuestos claros sobre los incentivos y las preferencias de otros jugadores. [3] Sin embargo, los modelos de IA, como GPT-3.5-turbo, no operan bajo un marco estrictamente racional ni tienen una noción clara de «maximización de utilidades» de manera innata. Esta explicación de R. J. Aumann pone de manifiesto que los resultados obtenidos en experimentos como el *Dictator Game* no necesariamente reflejan la manera en la que los humanos toman decisiones en entornos estratégicos, ya que las IA tienden interpretar los juegos de forma diferente debido a su entrenamiento. [3]

Finalmente, la limitación significativa consiste en que los modelos de IA generativa carecen de conciencia y de una verdadera comprensión de las decisiones altruistas o estratégicas. Aunque tienen la capacidad de simular respuestas que se asemejan a comportamientos humanos, no están motivados por principios morales o consideraciones sociales, lo que reduce su capacidad para participar plenamente en experimentos de teoría de juegos que se basan en la interacción genuina entre agentes conscientes.

VII. CONCLUSIONES

En esta sesión se presentan las conclusiones y hallazgos descubiertos a lo largo del desarrollo de la investigación. Desde una perspectiva metodológica, la implementación de un diseño experimental basado en la teoría de juegos permitió analizar las capacidades de los modelos a la hora de replicar comportamientos sociales, como el altruismo. El uso de 500 iteraciones en los experimentos proporcionó una base sólida para obtener resultados representativos, lo que contribuyó a la validez estadística de los hallazgos.

Los resultados muestran que el modelo GPT-3.5-turbo-0125 tiende a asignar 3 o 4 euros, lo que refleja un comportamiento moderadamente altruista, pero con una tendencia alejada de la equidad perfecta (50-50). Esta conducta contrasta con versiones anteriores, como la GPT-3.5-turbo-0301, que mostraron una mayor inclinación hacia la equidad. Esta diferencia entre las versiones permitió validar la pregunta de investigación plantada.

Además, se resalta la importancia de los cambios internos en los modelos de IA, aunque la falta de transparencia por parte de OpenAI sobre los ajustes específicos entre versiones brindadas dificulta la explicación precisa de estos resultados. Este desafío subraya la necesidad de una mayor documentación y comprensión sobre las actualizaciones de los modelos, que directamente facilitaría la replicabilidad y la interpretación de los estudios comparativos.

En términos metodológicos, uno de los aspectos destacados es el papel del *wording* en las decisiones generadas por los modelos de IAG. Aunque las instrucciones se diseñaron cuidadosamente para evitar sesgos, los resultados sugieren que incluso pequeñas variaciones en la redacción influyen significativamente en las respuestas del modelo. Esto confirma la sensibilidad del GPT-3.5-turbo al contexto textual, un aspecto que debe ser considerado en futuros estudios experimentales, dado que la precisión del *wording* llega a alterar el comportamiento observado en los agentes artificiales.

Por otro lado, el análisis comparativo entre las versiones de GPT-3.5-turbo revela importantes hallazgos sobre cómo los cambios en la arquitectura y el entrenamiento de los modelos afectan su desempeño en experimentos de teoría de juegos.

Aunque los resultados del modelo reciente ofrecen una visión matizada del comportamiento altruista, es evidente que el modelo aún no emula completamente las decisiones humanas. Esto plantea interrogantes sobre la capacidad de las IA para replicar decisiones morales y sociales complejas, y subraya la importancia de continuar investigando en esta intersección entre IA y comportamiento humano. Además, al comparar los resultados de este estudio con los de investigaciones previas, como el trabajo de A. Brookins y J. Debacker, se observan diferencias que reflejan una evolución en el comportamiento del modelo a lo largo del tiempo. [6] La tendencia del modelo GPT-3.5-turbo-0125 a alejarse de la equidad perfecta y su mayor dispersión en las asignaciones contrastan con los hallazgos de estudios anteriores, en los que se observaba una mayor concentración en valores equitativos. Esto sugiere que las actualizaciones en la arquitectura del modelo influyen en su capacidad para replicar dinámicas sociales complejas, lo que invita a una reflexión profunda sobre cómo estos sistemas interpretan los contextos sociales y responden a ellos.

Finalmente, se destacan las oportunidades y limitaciones de los modelos de lenguaje generativo como herramientas para explorar la teoría de juegos y la economía del comportamiento. Si bien los resultados obtenidos son reveladores, el presente trabajo invita a una reflexión profunda sobre el papel de las IA en la simulación de interacciones sociales, enfatizando la importancia de seguir explorando la evolución de estos modelos y su capacidad para adaptarse a escenarios complejos de toma de decisiones.

VIII. REFERENCIAS

[1] A. *C. C. de Oliveira*, "Altruism in Economic Decision-Making," *Journal of Behavioral Economics, vol. 12, no. 3, pp. 45-60, 2020.*

[2] C. Engel, "Dictator Games: A Meta Study," Experimental Economics, vol. 14, no. 4, pp. 1-20, 2011.

[3] R. J. Aumann, "Game Theory," *The New Palgrave Dictionary of Economics*, 2nd ed., vol. 3, pp. 452-457, 2008.

[4] T. Johnson y A. Obradovich, "Altruism and Selfishness in Believable Game Agents: Deep Reinforcement Learning in Modified Dictator Games," *Artificial Intelligence Review*, vol. 54, no. 2, pp. 123-145, 2021.

[5] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I, Sustskever, "Language Models are Unsupervised Multitask Learners," OpenAI, 2019.

[6] P. Brookins y J. Debacker, "Playing Games With GPT: What Can We Learn About a Large Language Model From Canonical Strategic Games?" *Social Science Research Network*, 2024.

[7] OpenAI, "GPT-3.5 Turbo". [Online]. Available: https://platform.openai.com/docs/models/gpt-3-5-turbo [Accessed:9 de Agosto de 2024].

[8] OpenAI, "Create chat". [Online]. Available: API Reference - OpenAI API [Accessed: 14 de Agosto de 2024].

IX. APÉNDICES

A. *Apéndice A : Código para ejecución de experimentos*

https://github.com/Vale-Martinez/Articulo-Experimentos-gpt-3.5-turbo.git

B. *Apéndice B: Tabla resultado de aplicación de experimentos.*

Run Number	Total Amount	GPTResponse	Full Response	Error Message
1				
2				
3				
…n				