Facultad de Ciencias Económicas y Empresariales
ICADE

Patrones morales en ChatGPT: una aproximación exploratoria desde la Teoría de los Fundamentos Morales

Autor: Beatriz Martínez Zato
Director: Pablo Calvo Báscones

MADRID | Junio 2026

Índice

Índice de tablas

Índice de figuras

Resumen

El presente Trabajo de Fin de Grado analiza cómo ChatGPT estructura sus juicios morales ante situaciones cotidianas con contenido ético o emocional, con el objetivo de identificar posibles patrones estables o sesgos valorativos en sus respuestas. El estudio parte de la Teoría de los Fundamentos Morales y del Moral Foundations Questionnaire-2 como marco conceptual, pero no aplica el cuestionario de forma directa. En su lugar, los ítems del MFQ-2 se reformulan como dilemas cotidianos abiertos, con el fin de reducir el reconocimiento del instrumento original y aproximar el análisis a contextos más naturales de interacción con el modelo.

La metodología se desarrolla en varias fases: identificación y unificación de juicios morales, construcción de pares morales opuestos, evaluación de su aplicabilidad y análisis de los resultados mediante rankings de dominancia, estabilidad gana–neutro–pierde y frecuencia de aparición. Además, se comparan las valoraciones propias de ChatGPT con las respuestas que el modelo atribuye a la mayoría de las personas.

Los resultados muestran que ChatGPT no responde de forma completamente neutral ante dilemas morales. Fundamentos como Cuidado, Igualdad, Proporcionalidad y Autoridad presentan estructuras más marcadas, mientras que Lealtad y Pureza muestran patrones más moderados o ambiguos. En general, las respuestas atribuidas a la mayoría mantienen una estructura similar, aunque con menor intensidad y mayor neutralidad.

En conjunto, el trabajo concluye que la orientación moral de ChatGPT no es única ni homogénea, sino contextual: el modelo activa distintos juicios morales según el tipo de dilema planteado. La principal aportación del estudio es metodológica, al proponer una forma alternativa de analizar la dimensión moral de los modelos generativos a partir de situaciones cotidianas y juicios morales intermedios.

Palabras clave: ChatGPT, inteligencia artificial generativa, modelos de lenguaje, juicio moral, fundamento moral, Teoría de los Fundamentos Morales, Moral Foundations Questionnaire, sesgos valorativos.

Abstract

This Thesis analyzes how ChatGPT structures its moral judgments when faced with everyday situations involving ethical or emotional content, with the aim of identifying possible stable patterns or value-based biases in its responses. The study draws on Moral Foundations Theory and the Moral Foundations Questionnaire-2 as its conceptual framework, but it does not apply the questionnaire directly. Instead, the MFQ-2 items are reformulated as open-ended everyday dilemmas in order to reduce recognition of the original instrument and bring the analysis closer to more natural contexts of interaction with the model.

The methodology is developed in several phases: identifying and unifying moral judgments, constructing opposing moral pairs, evaluating their applicability, and analyzing the results through dominance rankings, win–neutral–lose stability, and frequency of occurrence. In addition, ChatGPT's own assessments are compared with the responses that the model attributes to the majority of people.

The results show that ChatGPT does not respond in a completely neutral way to moral dilemmas. Foundations such as Care, Equality, Proportionality, and Authority show more marked structures, whereas Loyalty and Purity display more moderate or ambiguous patterns. In general, the responses attributed to the majority maintain a similar structure, although with lower intensity and greater neutrality.

Overall, the study concludes that ChatGPT's moral orientation is neither unique nor homogeneous, but contextual: the model activates different moral judgments depending on the type of dilemma presented. The main contribution of this study is methodological, as it proposes an alternative way of analyzing the moral dimension of generative models through everyday situations and intermediate moral judgments.

Keywords: ChatGPT, generative artificial intelligence, language models, moral judgment, moral foundation, Moral Foundations Theory, Moral Foundations Questionnaire, value-based biases.

1. Introducción y motivación

En los últimos años, los modelos de lenguaje de gran tamaño han dejado de limitarse a tareas informativas o de generación de texto para empezar a ocupar también un lugar en contextos emocionalmente sensibles. En particular, ChatGPT ha comenzado a utilizarse como espacio de desahogo, validación emocional, búsqueda de consejo y apoyo en la toma de decisiones personales, en buena medida por su accesibilidad, disponibilidad constante y ausencia de juicio percibido por parte del sistema (Luo et al., 2025a; Parrilla, 2026).

Este uso, cada vez más visible también en el ámbito social y mediático, ha suscitado una preocupación creciente. Aunque algunos trabajos destacan su utilidad percibida para la reflexión personal, la orientación o la gestión del malestar cotidiano, la literatura advierte igualmente de limitaciones importantes, como la falta de comprensión emocional genuina, la posibilidad de ofrecer respuestas erróneas o excesivamente complacientes, los problemas de privacidad y la dificultad para afrontar situaciones de especial vulnerabilidad (Blease & Torous, 2023; Luo et al., 2025a; Sánchez Santamaría et al., 2025). A ello se añade la presencia potencial de sesgos derivados tanto de los datos de entrenamiento como de los procesos de supervisión humana, lo que resulta especialmente problemático en ámbitos como la salud mental (Blease & Torous, 2023).

En este contexto, no basta con preguntarse si las respuestas de ChatGPT resultan útiles o convincentes. También es necesario examinar qué criterios morales parecen subyacer a ellas, especialmente cuando el modelo se enfrenta a situaciones cotidianas con carga ética o emocional. Como advierten Bender et al. (2021), los modelos de lenguaje de gran escala pueden producir respuestas fluidas, persuasivas y aparentemente significativas sin que ello implique una comprensión real del mundo o del significado en sentido humano. La cuestión central de este trabajo es, por tanto, analizar qué valores prioriza ChatGPT en sus respuestas, qué oposiciones morales activa y hasta qué punto esas pautas pueden revelar tendencias relativamente estables o sesgos sistemáticos.

Con este fin, el presente Trabajo de Fin de Grado propone una metodología

exploratoria basada en el Moral Foundations Questionnaire-2 (MFQ-2), un cuestionario diseñado para medir distintos fundamentos morales. En lugar de aplicar el cuestionario de forma directa, sus ítems se reformulan como situaciones cotidianas abiertas, con el objetivo de reducir la posibilidad de que ChatGPT reconozca el instrumento original y responda en función de su estructura conocida. Esta reformulación permite observar de manera más natural qué criterios, valores o juicios morales aparecen en las respuestas del modelo ante dilemas con contenido ético o emocional. Posteriormente, las categorías identificadas se agrupan según su similitud semántica y se analizan atendiendo a tres aspectos principales: el peso relativo que adquieren frente a otras categorías, la frecuencia con la que aparecen y la consistencia con la que se repiten. A partir de este análisis, se busca identificar patrones estables, tensiones recurrentes y posibles sesgos valorativos en la forma en que ChatGPT articula sus juicios morales.

La relevancia del estudio es doble. Por un lado, contribuye al análisis crítico del uso creciente de modelos generativos en contextos emocionalmente sensibles, donde sus respuestas pueden influir en la interpretación que los usuarios hacen de sus propios problemas. Por otro, ofrece una propuesta metodológica para explorar empíricamente la arquitectura moral implícita en las respuestas de un modelo de lenguaje ampliamente utilizado. En un escenario en el que los sistemas basados en Lenguaje de Gran Escala (LLM) adquieren un papel cada vez más activo en la vida cotidiana, su influencia puede aumentar al tiempo que disminuye su transparencia, lo que refuerza la necesidad de examinar sus sesgos no solo desde una perspectiva técnica, sino también ética y social (Blease & Torous, 2023; Luo et al., 2025a, 2025b).

2. Marco teórico

La presente sección tiene como objetivo establecer las bases conceptuales y los marcos de referencia teóricos que soportan este estudio. Para analizar de qué manera un LLM articula sus juicios morales, resulta necesario, en primer lugar, revisar las principales teorías de la psicología moral que han configurado la comprensión contemporánea de la moralidad humana. En segundo lugar, se examinará la naturaleza tecnológica de estos modelos, prestando especial atención a cómo sus procesos de entrenamiento buscan incorporar, reproducir o alinearse con valores humanos. Por último, se revisarán estudios similares con el fin de obtener una visión general sobre la moralidad en los modelos de lenguaje y poder establecer una comparación al final del análisis.

2.1 Teoría de los Fundamentos Morales

Este apartado establece el marco teórico desde el que se analizará la dimensión moral de las respuestas generadas por ChatGPT. Para ello, en primer lugar, se expone la Moral Foundations Theory (MFT) como propuesta explicativa de la diversidad de juicios morales. En segundo lugar, se presenta el Moral Foundations Questionnaire (MFQ), en su versión revisada, como herramienta para medir los fundamentos morales propuestos por la MFT.

2.1.1 La MFT como marco de análisis moral

La comprensión de la moralidad ha evolucionado de manera significativa a lo largo de la historia. Durante buena parte del siglo XX predominaron los enfoques racionalistas, representados por autores como Piaget y Kohlberg, para quienes el juicio moral surgía fundamentalmente de la reflexión consciente y del razonamiento deliberativo. Desde esta perspectiva, una persona emitía juicios morales tras analizar una situación concreta y aplicar principios normativos de forma racional, mientras que las emociones y las intuiciones quedaban relegadas a un papel secundario o incluso perturbador del proceso moral (Gómez de Liaño, 2011).

A comienzos del siglo XXI, esta visión comenzó a ser ampliamente cuestionada.

En particular, Jonathan Haidt, a través de su modelo del intuicionismo social (2001), defendió que el juicio moral no se origina habitualmente en una deliberación racional pausada, sino en intuiciones rápidas e inmediatas, estrechamente vinculadas a respuestas emocionales, hábitos sociales y marcos culturales compartidos. En este planteamiento, el razonamiento moral no actuaría como el origen principal del juicio, sino como un proceso posterior mediante el cual las personas justifican, explican o racionalizan intuiciones morales ya formadas (Gómez de Liaño, 2011).

A partir de este giro intuicionista se desarrolla posteriormente la *Moral Foundations Theory* (MFT), formulada por Haidt, Graham y otros colaboradores. Esta teoría parte de la idea de que la moralidad humana no se organiza en torno a un único principio universal, sino a partir de varios fundamentos morales relativamente diferenciados. Su propósito es describir cómo operan los juicios morales en la práctica y explicar por qué personas, grupos sociales y culturas distintas pueden priorizar aspectos morales diferentes al interpretar una misma situación moral (Moral Foundations, 2024; van den Berg y Corrias, 2026).

En su formulación más reciente, la Teoría de los Fundamentos Morales distingue seis fundamentos básicos que estructuran la moralidad humana (Moral Foundations, 2024; Aksoy, M., 2024):

Fundamento Moral	**Definición**
Cuidado	Sensibilidad ante el sufrimiento ajeno y preocupación por proteger a otras personas del daño
Igualdad	Preocupación por que las personas reciban iguales oportunidades y recursos
Proporcionalidad	Idea de que las recompensas deben corresponder al mérito, contribución o al esfuerzo
Lealtad	Compromiso con el grupo y alineación del éxito grupal con el individual
Autoridad	Respeto al liderazgo, a la jerarquía y a las tradiciones

Pureza	Defensa de lo sagrado, lo limpio y lo moralmente correcto frente a la contaminación o la degradación

Tabla 1. Fundamentos morales de la MFT

La principal aportación de la MFT es que entiende la moralidad como algo plural, es decir, formada por varios valores. Según esta teoría, las personas no siempre discrepan porque unas sean más morales que otras, sino porque dan más importancia a unos valores que a otros.

Por ejemplo, ante un mismo tema, una persona puede fijarse sobre todo en el cuidado y el daño, mientras que otra puede dar más importancia a la autoridad y la lealtad. Así, ambas pueden estar razonando moralmente, pero desde prioridades distintas.

Por este motivo, la MFT se ha utilizado para estudiar diferencias ideológicas y políticas, ya que muestra que distintos grupos suelen apoyarse en fundamentos morales diferentes al formar sus opiniones (Graham et al., 2009).

2.1.2 El MFQ: estructura, dimensiones y aplicación

El Moral Foundations Questionnaire-2 (MFQ-2) es un instrumento diseñado para evaluar empíricamente la relevancia de los fundamentos morales de la Teoría de los Fundamentos Morales. Dentro del marco de esta teoría, este cuestionario permite medir de forma estructurada hasta qué punto cada uno de los fundamentos participa en la manera en que una persona interpreta situaciones y emite juicios morales constituyendo una herramienta útil para trasladar el marco teórico de los fundamentos morales a un formato de evaluación sistemática.

En su versión actualizada, el MFQ-2 está compuesto por 36 ítems distribuidos entre los seis fundamentos morales, de manera que cada uno de ellos cuenta con seis preguntas asociadas, como se muestra en la Figura 1. Esta estructura permite obtener una medida diferenciada para cada dimensión moral y facilita tanto la comparación entre fundamentos como el análisis de su peso relativo (Münker, 2025).

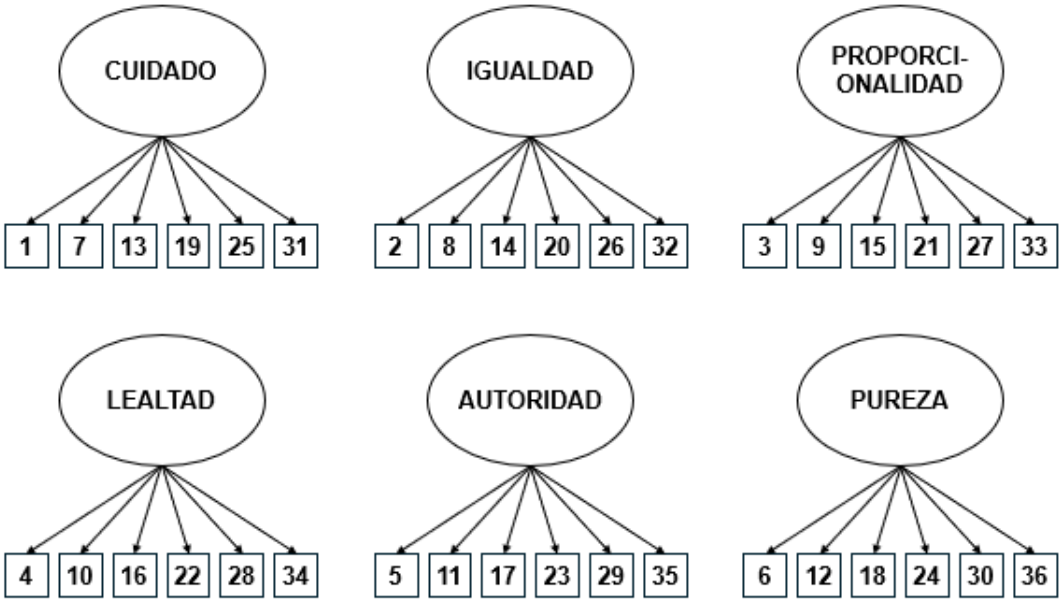


Figura 1. Distribución de los ítems del MFQ-2 por fundamento moral

La evaluación se realiza mediante una escala Likert de cinco puntos, en la que cada respuesta va de 1 (“no me describe en absoluto”) a 5 (“me describe extremadamente bien”). Este formato permite cuantificar el grado de identificación con cada afirmación y obtener posteriormente una puntuación para cada fundamento moral a partir de los ítems correspondientes.

Desde el punto de vista metodológico, el cuestionario ha sido utilizado principalmente en investigaciones con participantes humanos, especialmente en estudios sobre psicología moral, ideología política y diferencias culturales. Sin embargo, en la literatura reciente también ha comenzado a aplicarse a modelos de lenguaje, presentando cada ítem como un prompt y tratando la respuesta generada como si procediera de un participante sintético. Esta extensión del instrumento abre nuevas posibilidades de análisis, aunque también plantea problemas específicos de interpretación y validez.

2.2 Inteligencia artificial, modelos de lenguaje y ética

Tras comprender los fundamentos de la moralidad humana, es necesario analizar el vehículo tecnológico que protagoniza este estudio: los Modelos de Lenguaje de Gran Escala (LLMs). Esta sección explora la naturaleza técnica de

ChatGPT, las fuentes de información empleadas en su desarrollo, el modo en que se entrena y alinea su comportamiento, así como los mecanismos mediante los cuales se intenta orientar sus respuestas conforme a valores y criterios definidos por humanos.

2.2.1 Origen y naturaleza de los datos de entrenamiento

Para comprender adecuadamente las respuestas de ChatGPT, conviene comenzar por el origen de los datos utilizados en su desarrollo. OpenAI señala que sus modelos fundacionales se construyen a partir de tres grandes fuentes de información: (1) información públicamente disponible en internet, (2) información a la que accede mediante acuerdos con terceros y (3) información que los propios usuarios, entrenadores humanos e investigadores proporcionan o generan (OpenAI, s. f). Esta precisión resulta relevante porque permite descartar la idea de que el sistema opere sobre una fuente única, cerrada o completamente homogénea de conocimiento.

En relación con la primera de estas fuentes, OpenAI indica además que aplica filtros para excluir determinados materiales que no desea que sus modelos aprendan, entre ellos el discurso de odio, el contenido para adultos, los sitios que agregan información personal y el spam (OpenAI, s. f.), lo que demuestra que la recopilación de datos no consiste en una incorporación indiscriminada de contenidos, sino en una selección sometida a ciertos criterios previos.

En consecuencia, el origen de la información que sirve de base a ChatGPT debe entenderse como plural y heterogéneo. No se trata únicamente de textos tomados de internet, sino de un conjunto más amplio de materiales procedentes de distintas fuentes y sometidos a procesos de selección.

2.2.2 Entrenamiento, alineamiento y generación de respuestas

Una vez analizadas las fuentes de datos utilizadas para el desarrollo de los modelos de lenguaje, conviene aclarar que el hecho de que estos modelos se entrenen a partir de grandes volúmenes de información no significa que, al responder, consulten directamente una base de datos fija o un repositorio cerrado de conocimientos. Su funcionamiento no consiste en buscar una

respuesta previamente almacenada, sino en generar texto a partir de los patrones aprendidos durante el entrenamiento y el prompt del propio usuario.

Desde el punto de vista técnico, el desarrollo del modelo puede explicarse en tres momentos principales: el preentrenamiento, el alineamiento posterior y la generación de respuestas.

La primera fase es el preentrenamiento. OpenAI explica que GPT-4 es un modelo basado en la arquitectura Transformer y que fue preentrenado para predecir el siguiente token en un documento (OpenAI, 2023a). En términos generales, esto significa que el modelo aprende a identificar regularidades del lenguaje a partir de grandes cantidades de texto. La arquitectura Transformer permite procesar las relaciones entre los distintos elementos de una secuencia mediante mecanismos de atención, sin depender de sistemas secuenciales tradicionales basados en recurrencia. En particular, el mecanismo de self-attention permite valorar la relevancia de cada palabra o fragmento en relación con el resto del contexto, lo que facilita conectar elementos alejados dentro de una misma secuencia (Vaswani et al., 2017).

A través de este proceso, el modelo aprende patrones lingüísticos, relaciones estadísticas y formas probables de continuación textual. Por ello, su funcionamiento responde principalmente a una lógica de predicción probabilística del lenguaje. No debe entenderse, por tanto, como una comprensión humana del contenido, sino como la capacidad de producir respuestas coherentes a partir de regularidades aprendidas en los datos de entrenamiento (OpenAI, 2023a; Vaswani et al., 2017; Zapata-Ros, 2023).

La segunda fase es el alineamiento, también denominado ajuste posterior al preentrenamiento. Esta etapa tiene como finalidad orientar el comportamiento del modelo para que sus respuestas sean más útiles, seguras y adecuadas al contexto de uso. En este punto resulta especialmente relevante el aprendizaje por refuerzo a partir de retroalimentación humana, conocido como Reinforcement Learning from Human Feedback o RLHF. Esta técnica utiliza valoraciones humanas para enseñar al sistema a priorizar determinadas respuestas frente a otras, favoreciendo aquellas que se consideran más adecuadas, útiles o seguras

(Christiano et al., 2017). OpenAI señala, en esta misma línea, que GPT-4 fue afinado posteriormente mediante RLHF con el objetivo de mejorar su comportamiento y ajustarlo mejor a las expectativas humanas (OpenAI, 2023b).

A este proceso técnico se añaden las directrices de comportamiento definidas por OpenAI. El Model Spec establece que los modelos deben responder de forma útil, segura y alineada con las necesidades de usuarios y desarrolladores, así como una jerarquía de instrucciones para resolver posibles conflictos entre indicaciones (OpenAI, 2025). Por ello, el comportamiento final del modelo no depende solo del entrenamiento y del alineamiento, sino también de este marco operativo que orienta sus respuestas.

Finalmente, la tercera fase es la generación de respuestas. Cuando un usuario introduce una pregunta o instrucción, el modelo no recupera mecánicamente una respuesta almacenada, sino que genera una secuencia de tokens en función del contexto recibido, de los patrones aprendidos durante el preentrenamiento y de las restricciones incorporadas en las fases de alineamiento y comportamiento. En otras palabras, la respuesta final surge de la combinación entre aprendizaje estadístico, ajuste posterior e instrucciones de funcionamiento.

En suma, las respuestas de ChatGPT no proceden de una comprensión humana ni de la consulta directa a una base fija de conocimientos, sino de un aprendizaje estadístico posteriormente ajustado mediante técnicas de alineamiento y directrices de comportamiento. Por ello, el modelo puede generar respuestas coherentes y adaptadas al contexto, aunque dentro de los límites propios de un sistema de predicción lingüística. Esta precisión resulta especialmente relevante para el análisis posterior, ya que permite abordar sus respuestas ante dilemas morales no como manifestaciones de una moralidad propia, sino como producciones lingüísticas condicionadas por su entrenamiento, su alineamiento y el contexto de la interacción.

2.3 Estado del arte

El estudio de los sesgos en modelos generativos ha adquirido una relevancia creciente en la literatura reciente, especialmente a medida que sistemas como

ChatGPT se han extendido como herramientas de consulta, generación de contenido y apoyo a la toma de decisiones. En este contexto, distintos trabajos han analizado hasta qué punto las respuestas de estos modelos pueden reflejar orientaciones ideológicas, sesgos de representación social o patrones condicionados por su entrenamiento y alineamiento. A partir de esta base, este apartado revisa, en primer lugar, la evidencia empírica sobre sesgos e ideología en modelos generativos y, en segundo lugar, los estudios que han aplicado específicamente la Moral Foundations Theory (MFT) a ChatGPT.

2.3.1 Evidencia sobre sesgos e ideología en modelos generativos

La literatura reciente muestra que los modelos generativos no producen respuestas completamente neutrales cuando se enfrentan a cuestiones políticas, culturales o socialmente controvertidas, sino que tienden a reflejar patrones sistemáticos de sesgo y de representación social. En esta línea, Santurkar et al. (2023) señalan que los modelos de lenguaje no reproducen de forma homogénea las opiniones de la población general, sino que sus respuestas se aproximan más a determinados grupos demográficos e ideológicos que a otros. Esta idea resulta especialmente relevante, ya que sugiere que los modelos no solo generan texto, sino que también proyectan perfiles sociales reconocibles.

En el caso específico de ChatGPT, varios estudios han encontrado indicios consistentes de sesgo político. Motoki et al. (2024) concluyen que las respuestas por defecto del modelo se alinean de manera sistemática con posiciones de centroizquierda o izquierda, observando una mayor proximidad con perfiles demócratas en Estados Unidos, con Lula en Brasil y con el Labour Party en Reino Unido. En una línea similar, Rozado (2023) administró quince tests de orientación política a ChatGPT y halló que catorce de ellos situaban sus respuestas en posiciones de izquierda, a pesar de que el propio sistema afirmaba ser neutral cuando se le preguntaba directamente por su orientación política. En conjunto, ambos trabajos refuerzan la idea de que la aparente neutralidad declarada por el modelo no impide la presencia de sesgos ideológicos detectables empíricamente.

Junto al sesgo político, la investigación también ha documentado sesgos de representación cultural y social. Yuan et al. (2025) muestran que ChatGPT puede reproducir estereotipos y sesgos culturales en sus respuestas, aunque la intensidad de estos varía según la versión del modelo y la estrategia de prompting empleada. De forma complementaria, el informe de Fundar (Ortiz de Zárate et al., 2024) subraya que los modelos de lenguaje generativo tampoco representan de manera equilibrada al conjunto de la población, sino que tienden a aproximarse más a perfiles socialmente situados, en particular más masculinos, politizados y, en algunos casos, de mayor edad y nivel educativo. En resumen, esta evidencia muestra que diversos estudios han identificado posibles sesgos políticos, sociales y culturales en las respuestas de los modelos de lenguaje. Aunque estos trabajos no abordan siempre la dimensión moral de forma directa, permiten justificar la importancia de analizar también si los LLM presentan patrones sistemáticos en sus juicios morales. En este sentido, resulta pertinente recurrir a instrumentos ampliamente reconocidos, como el *Moral Foundations Questionnaire* (MFQ), cuya aplicación a ChatGPT y otros modelos de lenguaje se revisa a continuación.

2.3.2 Estudios específicos sobre el MFQ aplicado a ChatGPT

De forma adicional, este apartado tiene como objetivo analizar cuál ha sido hasta el momento la aplicación del Moral Foundations Questionnaire (MFQ), así como de otros instrumentos estrechamente vinculados a la Moral Foundations Theory (MFT), a ChatGPT y a otros instrumentos de la familia GPT. En particular, interesa examinar de qué manera se ha aplicado el instrumento, con qué objetivos se ha utilizado y qué resultados se han obtenido. En términos generales, la mayoría de los estudios recientes persigue comparar las respuestas de ChatGPT con las de los humanos para determinar hasta qué punto ambas se encuentran alineadas. Sin embargo, las metodologías empleadas difieren de forma considerable, tanto en el formato de administración como en el tipo de comparación realizada.

2.3.2.1 Aplicaciones directas del MFQ

En primer lugar, conviene destacar los trabajos que administran el MFQ de una

manera relativamente directa, es decir, manteniendo en gran medida la lógica original del cuestionario. En esta línea, Abdulhai et al. (2024) presentan cada pregunta del MFQ como un prompt independiente, solicitan una respuesta en escala 0–5, repiten cada ítem 50 veces y calculan después la puntuación final a partir de la clave estándar del instrumento. Posteriormente, comparan los perfiles obtenidos con distintos grupos humanos y analizan además si el prompting puede modificar el perfil moral resultante y afectar tareas posteriores. Desde el punto de vista metodológico, este trabajo resulta especialmente relevante porque reproduce de forma bastante fiel la estructura del cuestionario original, aunque adaptada al formato conversacional del modelo. Sus resultados muestran que el perfil por defecto de GPT se asemeja más al de participantes conservadores, aunque también ponen de manifiesto que dicho perfil puede modularse mediante prompting.

También dentro de esta lógica de aplicación relativamente directa del cuestionario, Abdurahman et al. (2024) administran el MFQ-2 a GPT-3.5 en 1.000 ocasiones y comparan posteriormente la distribución de respuestas obtenida con distribuciones humanas de 19 poblaciones culturales. En este caso, la metodología no se centra tanto en obtener un único perfil medio del modelo como en examinar si este puede comportarse como un “participante sintético” capaz de reflejar la diversidad moral humana. Este diseño permite observar no solo la media, sino también la varianza de las respuestas del modelo. Los autores concluyen que GPT-3.5 presenta una variabilidad muy inferior a la humana y, por tanto, no reproduce adecuadamente la diversidad moral entre individuos y culturas. Además, detectan un sesgo hacia mayores puntuaciones en care y proportionality, y menores en equality y loyalty.

También puede situarse en este grupo el trabajo de Bastos Costa et al. (2025), aunque con un objetivo diferente. En este caso, los autores utilizan el MFQ-30 para analizar la estabilidad del perfil moral de varios modelos GPT. Primero construyen un baseline sin role-play, en el que cada modelo responde repetidamente a las 30 preguntas del cuestionario, y después repiten el procedimiento haciendo que el modelo responda como si fuera 100 personas diferentes. A partir de este diseño calculan dos métricas, moral robustness y

moral susceptibility, que permiten observar hasta qué punto el perfil moral del modelo permanece estable o cambia cuando se modifica la identidad asumida. Sus resultados indican que, en la condición base, los modelos GPT puntúan más alto en Harm/Care y Fairness/Reciprocity, y más bajo en Loyalty, Authority y especialmente Purity/Sanctity. Además, muestran una robustez relativamente alta, aunque no completa, ya que el role-play sigue introduciendo variaciones en el perfil.

2.3.2.2 Aplicaciones indirectas de la MFT

Junto a estos estudios, nos encontramos con estudios que, aunque continúan trabajando dentro del marco de la MFT, modifican el instrumento o cambian de manera más clara la forma de administración. Un primer ejemplo es MoralBench, donde Abdulhai et al. (2024) no administran el MFQ-30 tal y como fue diseñado para humanos, sino que lo adaptan a dos tareas más manejables para los modelos: respuestas binarias Agree/Disagree ante enunciados morales y elecciones comparativas entre dos enunciados para decidir cuál es “más moral”. Estas respuestas se comparan después con las valoraciones humanas medias. Metodológicamente, este enfoque transforma el cuestionario en una tarea de clasificación y comparación, más cercana al tipo de evaluación habitual en los modelos de lenguaje. Sus resultados muestran una alineación considerable con respuestas humanas, aunque variable según la tarea, lo que sugiere que el comportamiento moral del modelo depende en parte del formato concreto de evaluación.

Un enfoque diferente, aunque todavía dentro de la Moral Foundations Theory, consiste en recurrir a instrumentos alternativos al MFQ clásico. En esta línea se sitúa el trabajo de Kirgis (2025), que utiliza las Moral Foundations Vignettes (MFV) en lugar del cuestionario de autoinforme. En este diseño, se administran 116 viñetas morales y se pide a los modelos que valoren la gravedad de cada infracción en una escala de 0 a 4. Cada viñeta se presenta en llamadas independientes para evitar efectos de orden, y las puntuaciones se comparan posteriormente con una línea base humana nacionalmente representativa. Desde el punto de vista metodológico, este enfoque desplaza la evaluación

desde el plano del autoinforme abstracto al juicio sobre situaciones concretas. Los resultados muestran que los modelos de OpenAI puntúan más alto que los humanos en care, fairness y liberty, y más bajo en loyalty, authority y sanctity, interpretándose este patrón como una cierta inclinación liberal dentro del marco de la MFT.

En una dirección parcialmente similar, aunque incorporando además una dimensión lingüística y cultural, Aksoy et al. (2024) aplican el MFQ-2 en ocho idiomas y comparan las respuestas de modelos GPT con respuestas humanas en esas mismas lenguas. La metodología combina, por tanto, dos niveles de análisis: por un lado, la comparación entre humanos y modelos y, por otro, la observación de cómo varía el perfil moral según el idioma en que se administra el instrumento. Este planteamiento permite estudiar si los modelos mantienen un perfil moral relativamente estable entre lenguas o si, por el contrario, reproducen sesgos lingüístico-culturales. Los resultados indican que GPT-3.5 y GPT-4o-mini se encuentran entre los modelos más alineados con respuestas humanas, aunque siguen mostrando diferencias asociadas al idioma y al contexto cultural.

Por último, otro planteamiento metodológico relevante es el de Nunes et al. (2024), quienes combinan MFQ y MFV en un mismo estudio para comprobar si el modelo mantiene coherencia entre los valores morales abstractos que expresa en el cuestionario y los juicios que emite ante casos concretos. Para ello, evalúan primero la consistencia interna de las respuestas dentro de cada instrumento y, después, analizan la relación entre ambos mediante regresiones lineales. Esta estrategia permite examinar no solo qué fundamentos aparecen más altos o bajos, sino también si existe una estructura moral coherente entre el nivel abstracto y el aplicado. Los autores concluyen que GPT-4 resulta relativamente consistente dentro de cada instrumento por separado, pero no entre ambos, lo que interpretan como una forma de hipocresía moral.

En conjunto, estos estudios muestran que no existe una única forma consolidada de aplicar el MFQ o la MFT a ChatGPT. Mientras algunos trabajos mantienen la estructura original del cuestionario, otros lo adaptan a formatos binarios, comparativos, multilingües o basados en viñetas. Por ello, los resultados

disponibles indican que ChatGPT puede reproducir parcialmente ciertos patrones morales humanos, pero que estos dependen en gran medida del instrumento y de la metodología utilizados.

3. Alcance del trabajo

En este capítulo se delimita el alcance de la investigación, concretando los objetivos, hipótesis, asunciones y restricciones que orientan el trabajo. Estos elementos resultan fundamentales para enmarcar la metodología empleada y para interpretar posteriormente los resultados obtenidos en el análisis de las respuestas de ChatGPT a partir de los fundamentos morales del MFQ-2.

3.1 Objetivos

El objetivo principal de este trabajo es explorar cómo ChatGPT estructura sus juicios morales ante dilemas cotidianos y analizar si en dicha estructuración pueden identificarse patrones consistentes o posibles sesgos valorativos.

Para alcanzar este objetivo principal, se establecen los siguientes objetivos específicos:

a) Revisar la literatura relevante sobre inteligencia artificial generativa, ChatGPT, juicio moral, Teoría de los Fundamentos Morales y cuestionarios de evaluación moral, con el fin de construir un marco conceptual sólido para el estudio.

b) Profundizar en la Teoría de los Fundamentos Morales como marco conceptual de referencia para interpretar la dimensión moral presente en las respuestas de ChatGPT.

c) Adaptar los ítems del MFQ-2 a situaciones cotidianas abiertas, con el fin de reducir el reconocimiento directo del cuestionario y aproximar el análisis a contextos más naturales de interacción.

d) Identificar los principales juicios o criterios morales que aparecen en las respuestas generadas por ChatGPT ante dichas situaciones.

e) Analizar la relación entre esos juicios morales y los fundamentos de la MFT tomados como referencia en el estudio

f) Explorar la posible existencia de patrones recurrentes, tendencias estables o sesgos valorativos en la forma en que el modelo articula sus respuestas.

3.2 Hipótesis

Con el fin de analizar si las respuestas de ChatGPT muestran una tendencia sistemática hacia alguno de los juicios morales evaluados, se plantean las siguientes hipótesis:

1. **Hipótesis nula (H_0):** ChatGPT no presenta una tendencia moral sistemática en sus respuestas.

   Esto implicaría que los distintos juicios morales tenderían a distribuirse de forma relativamente equilibrada, sin que algunos aparezcan de manera claramente dominante, estable y recurrente frente a otros. En este caso, cabría esperar pesos normalizados próximos a una distribución equilibrada, porcentajes de ganancia moderados y una aplicabilidad no concentrada en categorías concretas.

2. **Hipótesis alternativa (H_1):** ChatGPT presenta una tendencia moral sistemática en sus respuestas.

   Esto implicaría que determinados juicios morales adquieren mayor peso relativo, ganan con mayor frecuencia frente a sus opuestos y aparecen de forma recurrente en las situaciones analizadas.

Dado el carácter exploratorio del trabajo, estas hipótesis no se contrastan mediante pruebas estadísticas formales. Se utilizan como una guía para interpretar los resultados y valorar si las respuestas de ChatGPT se aproximan a la neutralidad o, por el contrario, muestran tendencias hacia determinados juicios morales.

Por tanto, el objetivo no es demostrar estadísticamente la existencia de un sesgo, sino identificar patrones relevantes en las respuestas del modelo a partir de un análisis exploratorio.

3.3 Asunciones

Durante el desarrollo de la investigación se han considerado una serie de asunciones que permiten sostener la coherencia del estudio y orientar la interpretación de los resultados. Estas asunciones se dividen en dos grupos: las relacionadas con el modelo analizado y las vinculadas a la metodología empleada.

Asunciones relacionadas con el LLM:

1. Se asume que las respuestas generadas por ChatGPT en el contexto de este trabajo no están condicionadas por un conocimiento previo del usuario ni por conversaciones anteriores, de modo que dependen fundamentalmente del contenido de los prompts planteados.

Asunciones relacionadas con la metodología:

2. Se presupone que las categorías morales emergentes identificadas a partir de las respuestas del modelo pueden ser agrupadas semánticamente de forma coherente en un conjunto de juicios morales unificados, sin perder por ello su valor analítico.
3. Se asume que las puntuaciones obtenidas en la escala Likert asociada a pares de juicios morales opuestos permiten aproximarse de forma razonable al grado de inclinación del modelo hacia uno u otro extremo moral.

3.4 Restricciones

El presente estudio presenta algunas restricciones que deben tenerse en cuenta al interpretar sus resultados y como han tratado de resolverse:

Restricciones técnicas y metodológicas

1. Falta de control técnico completo: el uso de la interfaz general de ChatGPT, en lugar de la API, limita el control sobre determinados parámetros técnicos y dificulta la replicabilidad exacta del procedimiento. Entre estos parámetros se encuentra la temperatura, que regula el grado

de aleatoriedad o variabilidad de las respuestas generadas. Al no poder fijar este valor de forma explícita, las respuestas pueden verse afectadas por pequeñas variaciones en la generación del modelo, lo que debe tenerse en cuenta al interpretar la estabilidad de los resultados.

2. Dependencia del modelo y del momento de aplicación: los resultados corresponden a una versión concreta de ChatGPT y podrían cambiar con futuras actualizaciones, ajustes en el alineamiento o modificaciones del sistema.
3. Sensibilidad al prompting: los resultados pueden variar según la formulación exacta de las preguntas, el contexto introducido o las instrucciones dadas al modelo, lo que plantea dudas sobre la estabilidad de los perfiles obtenidos (Binz & Schulz, 2023; Abdulhai et al., 2024). En este trabajo, esta limitación se ha intentado reducir mediante el uso de prompts homogéneos, con una misma estructura formal, de manera que las diferencias observadas respondan en mayor medida al contenido de las situaciones que a variaciones en la forma de preguntar.

Restricciones interpretativas

4. Posible reconocimiento del cuestionario: aunque los ítems se reformulen, ChatGPT podría reconocer la lógica del MFQ o asociar las situaciones con fundamentos morales concretos. Esta limitación se relaciona con los problemas de contaminación de datos en la evaluación de LLM, especialmente cuando se emplean cuestionarios o benchmarks previamente conocidos (Yang et al., 2024; Balloccu et al., 2024; Wang et al., 2026). Para reducir este riesgo, el estudio adopta un enfoque sintético, reformulando los ítems originales como situaciones cotidianas abiertas y menos reconocibles como parte del cuestionario.
5. Dificultad para distinguir entre moralidad real y respuesta aparente: que las respuestas parezcan coherentes desde el punto de vista moral no significa que el modelo tenga una moralidad propia, sino que puede estar generando una respuesta adecuada al contexto (Bender et al., 2021; La Fontaine, 2024).

6. Alcance exploratorio: el trabajo no busca demostrar si ChatGPT “tiene” moralidad, sino observar posibles patrones en sus respuestas. Por ello, los resultados deben entenderse como tendencias, no como pruebas de una moralidad propia del modelo.

4. Metodología

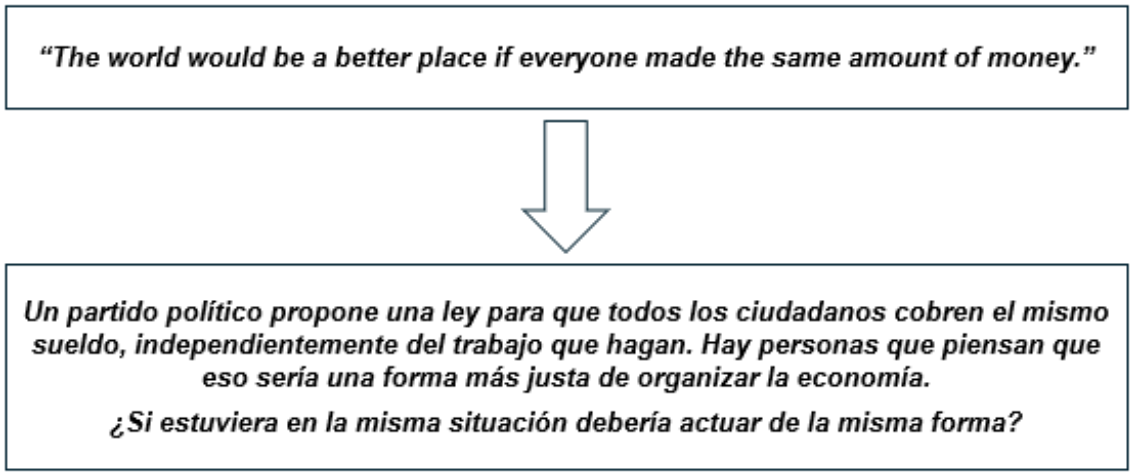
En esta sección se describe el procedimiento seguido para analizar cómo un modelo de lenguaje estructura sus juicios morales ante situaciones cotidianas. Para ello, se partió de la adaptación de cuestionarios de valores a un formato abierto y se desarrolló un proceso en varias fases que permitió identificar, estructurar y analizar los juicios morales generados por el modelo.

4.1 Fase 1: Diseño y adaptación del instrumento

Una vez seleccionado el MFQ-2 como cuestionario base para la evaluación del modelo dada su relevancia en el ámbito de la psicología, se realizó una exploración preliminar con el objetivo de comprobar hasta qué punto el modelo era capaz de reconocer dicho instrumento. Los resultados mostraron que el modelo identificaba el cuestionario incluso cuando se presentaban fragmentos parciales o versiones traducidas, lo que indicaba que sus respuestas podían estar condicionadas por su entrenamiento previo y no únicamente por el contenido de las preguntas.

Para evitar este problema y reducir la influencia del conocimiento previo del cuestionario, se procedió a su reformulación. En un primer momento, se elaboraron tres versiones alternativas en las que los ítems originales se transformaron en situaciones cotidianas, formuladas de manera moralmente ambigua y en lenguaje natural, pero manteniendo el fundamento moral correspondiente del MFQ-2.

Por ejemplo, uno de los ítems originales del MFQ-2 vinculado al fundamento de Igualdad se formulaba del siguiente modo:



De este modo, el contenido moral del ítem original se mantiene, pero se transforma en una situación cotidiana y contextualizada, reduciendo la posibilidad de reconocimiento directo del cuestionario por parte del modelo.

El objetivo de esta reformulación fue doble. Por un lado, minimizar el efecto del reconocimiento del cuestionario original por parte del modelo. Por otro, generar situaciones más cercanas al tipo de consultas que una persona podría plantear a ChatGPT en un contexto aplicado, como el uso del modelo en un rol similar al de un psicólogo, tal y como se planteaba en la introducción.

4.2 Fase 2: Identificación y unificación de juicios morales

Una vez reformulado el cuestionario en situaciones cotidianas, se procedió a una segunda fase orientada a identificar los juicios morales que emergían en las situaciones morales sintéticamente generadas del modelo. El objetivo de esta fase era observar qué categorías valorativas utilizaba ChatGPT para interpretar las situaciones planteadas, sin imponer previamente una clasificación cerrada basada en los fundamentos del MFQ-2.

Para ello, cada uno de los cuestionarios reformulados se presentó en tres ocasiones distintas de manera abierta, es decir, sin proporcionar al modelo una lista previa de categorías entre las que elegir. El objetivo de esta fase era observar qué categorías morales aparecían de forma espontánea ante cada una de las situaciones planteadas.

En esta fase, el modelo recibió una instrucción orientada a identificar los juicios morales presentes en cada situación. De forma resumida, el enunciado utilizado fue el siguiente:

> ***"Quiero que me respondas a estas preguntas con una escala según creas que será tu respuesta si algún usuario te plantea la situación sobre cual debería de ser su forma de pensar, y en otra columna lo que crees que piensa la mayoría. En la tercera columna el juicio moral (categoría) que se interpreta con un 5 (para ti), y en la cuarta columna el juicio moral (categoría) que se interpreta con un 1 (para ti), las categorías 1 y 5 tienen que estar vinculadas al tema que se cuestiona y deben ambas positivas y excluyentes, por ejemplo 5=empatía y 1=productividad.:"***

A partir de estas tres aplicaciones iniciales, se recopilaron las categorías morales que el modelo utilizaba para interpretar las distintas situaciones. Posteriormente, dichas categorías fueron comparadas y agrupadas en función de su similitud semántica, con el fin de reducir redundancias y unificar aquellas que remitían a significados próximos. Este proceso permitió pasar de un conjunto amplio y disperso de etiquetas iniciales a una estructura más ordenada y manejable.

A efectos de este estudio, se denomina **juicios morales** a estas categorías unificadas resultantes del proceso de agrupación. Por tanto, no se entienden como juicios morales en sentido humano o deliberativo, sino como etiquetas analíticas que permiten describir los criterios morales que aparecen en las respuestas generadas por ChatGPT.

Como resultado de esta fase, se definió un conjunto de dieciséis juicios morales unificados, que se recogen en la Tabla 2. En ella se presenta cada juicio moral, algunas de las categorías iniciales que engloba y una breve definición operativa elaborada para orientar su interpretación en las fases posteriores del análisis.

Juicio moral	**Categorías iniciales que engloba**	**Definición**
Apertura	• Respeto por convicciones • Apertura cosmopolita • Apertura respetuosa	Disposición al respeto, el pluralismo y la aceptación de distintas convicciones e identidades

Autoridad	• Autoridad funcional • Autoridad pedagógica • Liderazgo firme institucional	Reconocimiento del liderazgo, la guía y las normas dentro de una estructura
Autonomía	• Autocuidado corporal y consciente • Autonomía y consentimiento • Respeto y límites	Valoración del respeto por las decisiones, los límites personales y el cuidado consciente de uno mismo
Comunidad	• Pertenencia cívica • Pertenencia comunitaria • Civismo y pertenencia	Sentido de pertenencia, lealtad e identidad compartida dentro de un grupo o entorno común
Compasión	• Compasión activa • Compasión global • Atención a la vulnerabilidad	Defensa del cuidado, la ayuda y la sensibilidad ante el sufrimiento y la vulnerabilidad de los demás
Deber	• Deber cívico solidario • Deber de auxilio • Lealtad cívica / mínima	Sentido de obligación moral, cívica y solidaria hacia los demás
Elección	• Libertad de trayectorias • Decoro y respeto del espacio • Atención a la accesibilidad	Capacidad de decidir libremente dentro del respeto al espacio común y a los demás
Flexibilidad	• Armonía de grupo • Alegría por la justicia • Apertura respetuosa	Apertura a la adaptación, la espontaneidad y el equilibrio sin rigidez
Horizontalidad	• Horizontalidad participativa • Valor de la experiencia	Importancia de la participación, la igualdad de voces y el respeto mutuo

	• Respeto a la experiencia	
Igualitarismo	• Igualdad salarial • Igualdad material • Igualitarismo laboral	Defensa de la igualdad justa entre las personas en recursos, oportunidades y trato
Individualismo	• Individualismo abierto • Libertad identitaria • Preparación individual	Prioridad de la autonomía, la identidad y la responsabilidad personal por encima de lo colectivo
Meritocracia	• Meritocracia justa • Justicia meritocrática • Equidad meritocrática	Defensa del reconocimiento y la recompensa según el esfuerzo, el mérito y la capacidad
Orden	• Orden y estructura • Orden tradicional • Orden firme democrático	Importancia de la disciplina, la estructura y el respeto por normas que sostienen la convivencia
Responsabilidad	• Responsabilidad y esfuerzo • Responsabilidad y consecuencias • Responsabilidad y normas	Valoración del compromiso personal, la coherencia y la asunción de las consecuencias de los propios actos
Solidaridad	• Responsabilidad de ayuda • Solidaridad espontánea • Responsabilidad compasiva	Valoración de la ayuda y el apoyo a los demás como responsabilidad compartida
Tradición	• Respeto a la tradición • Orden tradicional • Continuidad familiar	Importancia de la continuidad, las costumbres y el respeto por lo heredado

Tabla 2. Juicios morales unificados y definición operativa

En conjunto, estos juicios morales unificados constituyen el marco categorial a partir del cual se analizó posteriormente la estructura moral de las respuestas del modelo.

4.3 Fase 3: Construcción y saturación de pares morales

Una vez definido el sistema de dieciséis juicios morales unificados, se procedió a una tercera fase orientada a construir pares morales opuestos. Estos pares consisten en combinaciones de dos juicios morales que representan posiciones contrapuestas dentro de una misma situación. Por ejemplo, ante una situación vinculada a la ayuda a otra persona, el modelo podía interpretar que existía una tensión entre Solidaridad e Individualismo, o entre Compasión y Responsabilidad.

El objetivo de esta fase no era todavía evaluar qué juicio moral predominaba, sino identificar qué oposiciones morales eran relevantes para cada situación reformulada. Para ello, se elaboraron nuevos cuestionarios con nuevas situaciones y se pidió al modelo que, para cada ítem, seleccionara dos juicios morales contrapuestos que representaran los extremos de una misma dimensión valorativa. En esta fase, el modelo debía operar exclusivamente dentro del sistema categorial definido en la fase anterior, es decir, solo podía utilizar los dieciséis juicios morales unificados y no podía generar nuevas categorías.

De este modo, cada respuesta permitía asociar una situación concreta con un par moral determinado. A medida que se administraban nuevos cuestionarios, se iban registrando los pares generados por el modelo y comprobando si aparecían combinaciones nuevas o si, por el contrario, se repetían pares ya identificados. El propósito era determinar cuándo el procedimiento alcanzaba un punto de saturación.

En este estudio, se entiende por saturación el momento en que la incorporación de nuevos cuestionarios deja de aportar un número significativo de pares morales novedosos. Es decir, cuando las nuevas aplicaciones empiezan a repetir mayoritariamente combinaciones ya observadas, puede considerarse que el repertorio de oposiciones morales está suficientemente cubierto para continuar

con el análisis.

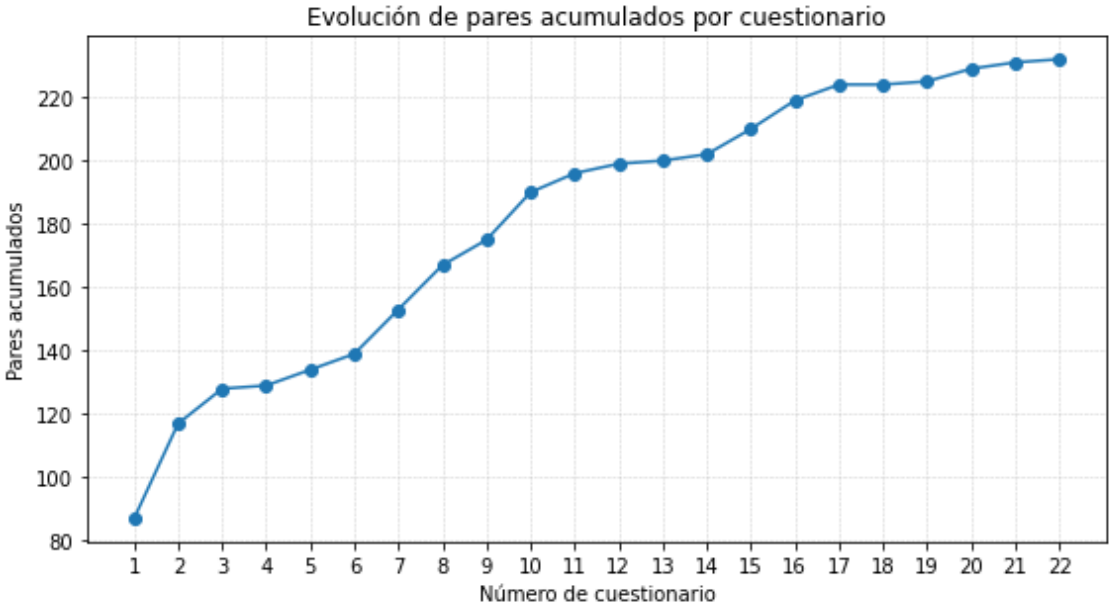


Figura 2. Evolución de pares morales acumulados por cuestionario

La Figura 2 muestra la evolución del número total de pares acumulados a medida que se incorporaban los distintos cuestionarios. Como puede observarse, el crecimiento fue especialmente intenso en las primeras aplicaciones, lo que indica que en esta etapa todavía emergían numerosas combinaciones nuevas. Sin embargo, a medida que aumentó el número de cuestionarios administrados, la pendiente de la curva comenzó a reducirse progresivamente hasta aproximarse a una fase de estabilización en los últimos casos. Esta tendencia respaldó la decisión de detener la construcción de nuevos cuestionarios al alcanzarse un nivel de saturación considerado suficiente para el análisis.

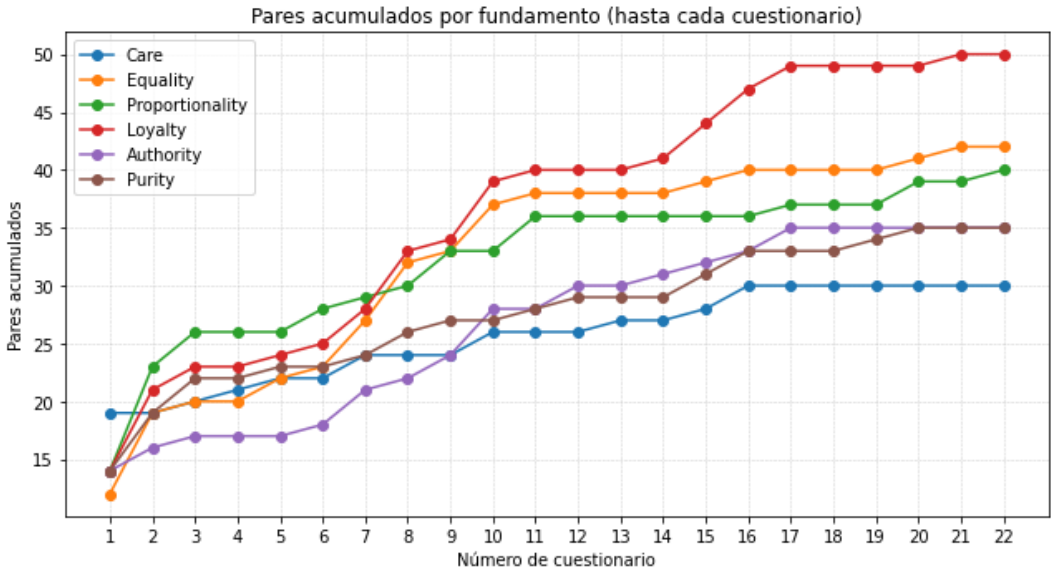


Figura 3. Evolución acumulada de pares morales por fundamento

Por su parte, la Figura 3 presenta la evolución de los pares acumulados desagregada por fundamento moral del MFQ-2. Los resultados muestran que la saturación no se produjo de manera homogénea en todos los fundamentos. Algunos, como Lealtad e Igualdad, continuaron incorporando nuevas combinaciones durante un mayor número de cuestionarios, mientras que otros, como Cuidado, mostraron una estabilización más temprana. Fundamentos como Proporcionalidad, Autoridad o Pureza presentaron trayectorias intermedias, con incrementos progresivos seguidos de fases de mayor estabilidad. Esta variabilidad sugiere que algunos fundamentos morales generan una mayor diversidad de oposiciones que otros en las respuestas del modelo.

En conjunto, esta fase permitió obtener un conjunto suficientemente amplio de pares morales para continuar con el análisis. La evolución de los resultados mostró que, tras varias aplicaciones, empezaban a aparecer cada vez menos combinaciones nuevas y se repetían con mayor frecuencia pares ya identificados. Por ello, se consideró que el repertorio de oposiciones morales estaba suficientemente cubierto. Estos pares sirvieron como base para la fase siguiente, en la que se evaluó su aplicabilidad y peso relativo en las respuestas del modelo.

4.4 Fase 4: Evaluación de los pares morales y su aplicabilidad

Una vez identificados los pares morales asociados a cada fundamento del MFQ-2, se realizó una cuarta fase para comprobar hasta qué punto esos pares servían para interpretar las respuestas del modelo. En esta etapa ya no se buscaban nuevos pares, sino valorar los que habían aparecido previamente.

El procedimiento se aplicó por separado para cada fundamento moral. En las fases anteriores, cada fundamento había dado lugar a un número distinto de pares morales posibles. Por ello, en esta cuarta fase se pidió a ChatGPT que volviera a evaluar las situaciones asociadas a cada fundamento, pero utilizando únicamente los pares morales previamente identificados para ese fundamento. Por ejemplo, si en Lealtad se habían detectado 50 pares morales, el modelo debía valorar las situaciones de Lealtad solo a partir de esos 50 pares.

En esta fase, ChatGPT no podía crear nuevas categorías ni proponer nuevos pares. Su tarea era valorar si cada par moral era útil para interpretar cada situación. Para ello, debía indicar si el par era aplicable o no. Cuando el par sí era aplicable, el modelo asignaba una puntuación que mostraba hacia cuál de los dos polos del par se inclinaba la situación.

Además, esta valoración se realizó en dos modalidades. En primer lugar, se pidió a ChatGPT que respondiera según su propia valoración de la situación. En segundo lugar, se le pidió que respondiera según cómo creía que lo haría la mayoría de las personas. De este modo, fue posible comparar la estructura moral propia del modelo con la estructura moral que el propio ChatGPT atribuye a la mayoría.

La valoración se hizo mediante una escala ordinal de cinco puntos. Un extremo indicaba que la situación favorecía claramente la primera categoría del par; el otro extremo, que favorecía claramente la segunda. El punto medio representaba una posición equilibrada entre ambas. Además, se añadió la opción “x” para los casos en los que el par no tenía una relación clara con la situación.

Para aumentar la estabilidad de los resultados, este proceso se repitió en 10 iteraciones independientes sobre un conjunto de 10 cuestionarios. Así, se pudo registrar tanto la frecuencia con la que cada par era considerado aplicable como

la dirección moral predominante en cada fundamento.

En conjunto, esta fase permitió convertir los pares morales identificados previamente en una base de datos cuantificable. Esto hizo posible analizar qué oposiciones morales aparecían con mayor frecuencia, cuáles eran más consistentes y qué peso tenían dentro de las respuestas generadas por el modelo.

4.5 Fase 5: Análisis de resultados por juicio moral

Tras finalizar la fase de evaluación, se analizaron las respuestas obtenidas en los 10 cuestionarios y sus 10 iteraciones independientes. El análisis se realizó en tres niveles, tanto para cada fundamento moral del MFQ-2 —Cuidado, Igualdad, Proporcionalidad, Lealtad, Autoridad y Pureza— como para el conjunto total de los resultados.

Los tres niveles fueron los siguientes:

- **Nivel 1:** análisis del peso relativo de los juicios morales mediante rankings de dominancia normalizados
- **Nivel 2:** análisis de la estabilidad de las respuestas entre cuestionarios e iteraciones
- **Nivel 3:** análisis de la frecuencia de aplicabilidad de cada juicio moral

Estos tres niveles se relacionan con la hipótesis nula del estudio, según la cual ChatGPT no presentaría una tendencia moral sistemática. En ese caso, sus respuestas deberían mantenerse relativamente equilibradas, sin favorecer de forma clara, estable y frecuente unos juicios morales frente a otros. Por el contrario, si determinados juicios aparecen con mayor peso, ganan con frecuencia alta y son aplicables en muchas situaciones, esto podría indicar la existencia de patrones morales en las respuestas del modelo.

Por último, los resultados obtenidos en la valoración propia de ChatGPT se compararon con los resultados atribuidos a la mayoría. Esta comparación permitió analizar si el modelo proyectaba sobre la mayoría una estructura moral

similar a la suya o si, por el contrario, atribuía a la mayoría una orientación diferente, más moderada o cercana a la neutralidad.

A continuación, se desarrolla cada uno de estos niveles de análisis, explicando el procedimiento seguido y los indicadores utilizados en cada caso.

Nivel 1: peso relativo de los juicios morales

El primer nivel de análisis tuvo como objetivo identificar qué juicios morales adquirían mayor peso relativo en las respuestas de ChatGPT. Para ello, se elaboraron rankings de dominancia, que permitieron ordenar los juicios morales según su importancia dentro de cada fundamento moral y en el conjunto total de los resultados.

Este procedimiento se basó en el Proceso Analítico Jerárquico (AHP, por sus siglas en inglés), un método de comparación por pares que permite estimar la importancia relativa de distintos elementos. En este estudio, el AHP se aplicó sobre los pares morales que habían surgido en las fases anteriores, con el objetivo de comparar los juicios morales entre sí y determinar cuáles tendían a predominar en las respuestas del modelo.

Dado que este procedimiento se basa en el uso de la media geométrica, antes de aplicarla fue necesario transformar las puntuaciones originales de la escala ordinal a valores compatibles con el AHP. Para ello, las respuestas se recodificaron del siguiente modo:

Valor original	Valor AHP
1	9
2	3
3	1
4	1/3
5	1/9

Tabla 3. Recodificación de la escala ordinal a valores AHP

De este modo, tras la recodificación los valores superiores a 1 indicaban

predominio del primer juicio moral del par, el valor 1 indicaba equilibrio entre ambos, y los valores inferiores a 1 indicaban predominio del segundo juicio moral. Los casos marcados con “x”, al no ser aplicables, se trataron como valores ausentes y no se incluyeron en el cálculo.

Una vez transformados los valores, se aplicó la media geométrica para agregar las respuestas obtenidas. La fórmula empleada fue:

$$\overline{a_{ij}} = \left(\prod_{k=1}^{n} a_{ij}^{k} \right)^{\frac{1}{n}}$$

donde a_{ij}^{k} representa cada valor válido obtenido para la comparación entre los juicios i y j, y n el número de valores válidos disponibles.

El proceso de agregación se realizó en tres pasos. En primer lugar, para cada cuestionario y cada pregunta, se agregaron mediante media geométrica las diez iteraciones de los pares morales previamente identificados, obteniendo una matriz $CxPy$. En segundo lugar, se agregaron las matrices correspondientes a los distintos cuestionarios de una misma pregunta, generando una matriz Py. Finalmente, se agregaron las matrices de las preguntas pertenecientes a un mismo fundamento moral, obteniendo una matriz final por fundamento.

Una vez obtenida la matriz final de cada fundamento, el peso geométrico observado de cada juicio moral se calculó mediante la media geométrica de los valores de cada fila:

$$w_i = \left(\prod_{j=1}^{m} a_{ij} \right)^{\frac{1}{m}}$$

donde w_i representa el peso geométrico observado del juicio moral i, a_{ij} representa el valor de comparación entre el juicio i y el juicio j, y m el número de comparaciones válidas disponibles para ese juicio.

Posteriormente, los pesos geométricos observados se normalizaron dividiendo cada peso entre la suma total de pesos del fundamento:

$$w_i^{norm} = \frac{w_i}{\sum_{r=1}^{m} w_r}$$

De este modo, los pesos normalizados de cada fundamento suman 1 y permiten interpretar qué proporción del peso total corresponde a cada juicio moral.

Finalmente, los valores normalizados se ordenaron de mayor a menor para construir el ranking de dominancia de cada fundamento moral y del análisis conjunto. En esta interpretación, una distribución completamente equilibrada entre los 16 juicios morales asignaría a cada categoría un peso de 0.0625. Por tanto, valores superiores a 0.0625 indican una presencia relativa mayor de la esperada bajo neutralidad, mientras que valores inferiores reflejan una menor presencia relativa.

Este mismo procedimiento sirvió también como base para el análisis transversal desarrollado posteriormente en la Fase 6. En dicha fase, los pesos normalizados obtenidos en cada fundamento se agregaron por juicio moral, con el objetivo de identificar qué categorías mantenían mayor peso medio en el conjunto del estudio, más allá de su comportamiento dentro de un fundamento concreto.

Nivel 2: estabilidad de las respuestas

El segundo nivel tuvo como objetivo comprobar la estabilidad de las respuestas, observando cómo variaba la dirección moral entre cuestionarios e iteraciones. Es decir, no se buscaba medir únicamente si las puntuaciones cambiaban mucho o poco, sino comprobar si esos cambios implicaban un desplazamiento hacia el juicio opuesto del par moral.

Por esa razón, se utilizó un análisis de tipo gana–neutro–pierde, en lugar de la desviación estándar. La desviación estándar permite medir dispersión, pero no indica claramente si el modelo cambia de orientación moral. Por ejemplo, no es lo mismo variar entre 1, 2 y 3, donde la respuesta sigue cercana al primer juicio o al equilibrio, que variar entre 2, 3 y 4, donde ya aparece un desplazamiento hacia el juicio opuesto.

Este análisis permitió identificar si cada juicio moral tendía a ganar, mantenerse neutral o perder frente a su opuesto. Así, fue posible comprobar si las respuestas del modelo seguían una dirección moral estable o si cambiaban entre los distintos

cuestionarios e iteraciones.

Nivel 3: frecuencia de aplicabilidad

Por último, de forma adicional, se realizó un análisis de la frecuencia de aplicabilidad de cada juicio moral. Para ello, se calculó el porcentaje de ocasiones en que cada categoría aparecía en pares considerados aplicables respecto al total de veces en que podía aparecer. Es decir, se tuvieron en cuenta los casos en los que el par moral no era marcado con “x”.

Este análisis permitió comprobar cuántas veces el modelo consideraba que los pares asociados a cada juicio moral eran relevantes para interpretar las situaciones evaluadas. En este sentido, la frecuencia de aplicabilidad funcionó como una medida adicional de fiabilidad: si un par era considerado aplicable en muchas situaciones y, además, sus resultados se mantenían estables, la tendencia observada podía interpretarse como más sólida.

De este modo, este análisis permitió diferenciar entre juicios morales cuyos pares eran aplicables en muchas situaciones y, por tanto, tenían una presencia más amplia y recurrente, y juicios cuyos pares solo resultaban aplicables en situaciones más concretas o puntuales.

A continuación, se presentan los resultados obtenidos a partir de estos tres niveles de análisis, primero para cada uno de los fundamentos morales del MFQ-2 y, posteriormente, para el conjunto total de los resultados.

4.5.1 Cuidado

El análisis del fundamento de Cuidado muestra que ChatGPT organiza estas situaciones principalmente alrededor de tres juicios morales: Compasión, Comunidad y Solidaridad. Estas categorías reflejan una orientación hacia la empatía, el apoyo colectivo y la preocupación por los demás.

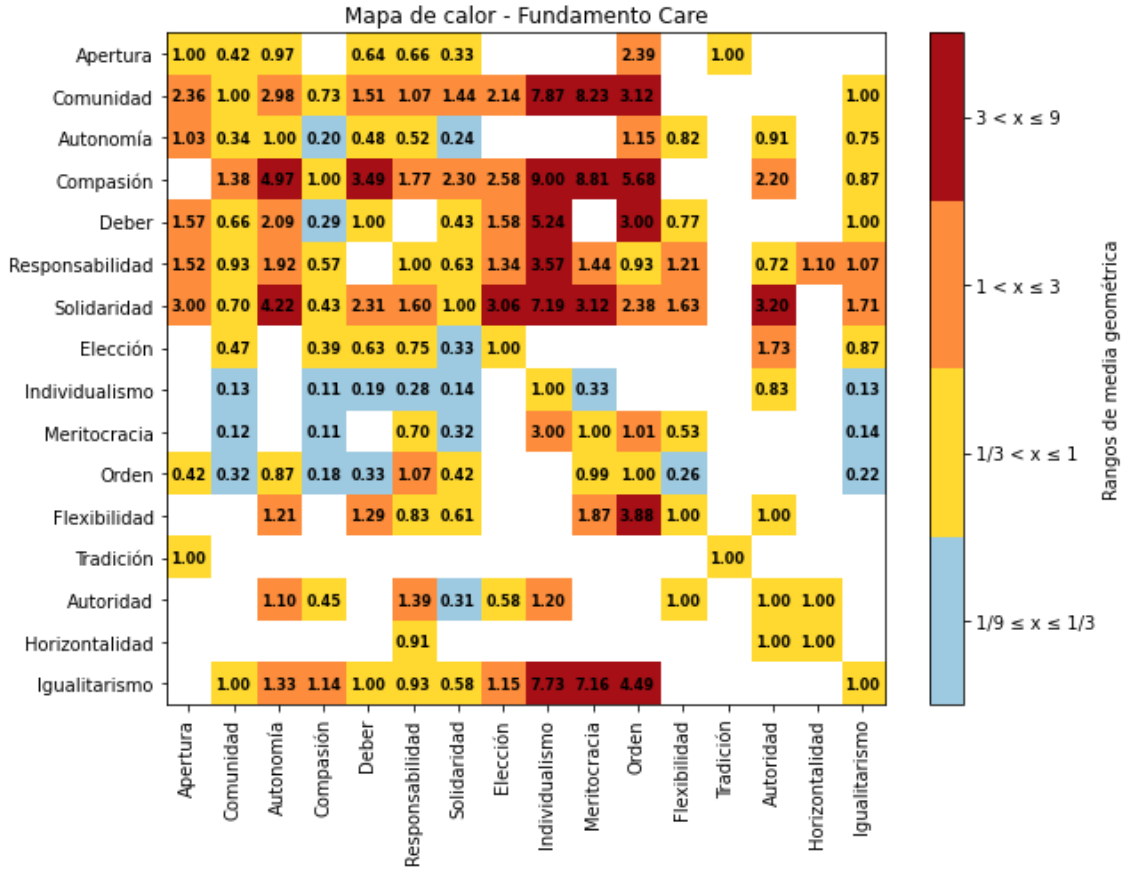


Figura 4. Mapa de calor de dominancia observada entre juicios morales en el fundamento de Cuidado — valoración propia de ChatGPT

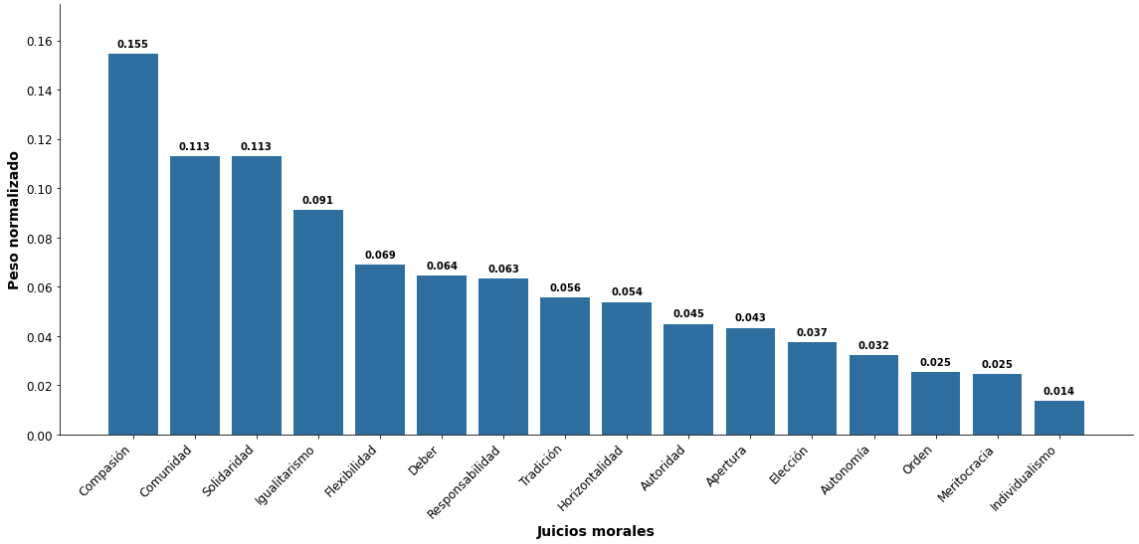


Figura 5. Ranking de pesos geométricos normalizados en el fundamento de Cuidado — valoración propia de ChatGPT

Como se observa en el mapa de calor de la Figura 4 y en el ranking de pesos geométricos normalizados de la Figura 5, estas tres categorías son las más relevantes dentro del fundamento. Compasión ocupa la primera posición, con un peso normalizado de 0.155, seguida de Comunidad y Solidaridad, ambas con un peso normalizado de 0.113. En conjunto, estas tres categorías concentran aproximadamente el 40% del peso total normalizado del fundamento de Cuidado, lo que indica que ChatGPT concentra una parte importante del peso relativo en categorías vinculadas a la sensibilidad ante el sufrimiento, la ayuda y el vínculo con otras personas. En cambio, categorías como Individualismo, Meritocracia u Orden quedan claramente alejadas del núcleo principal.

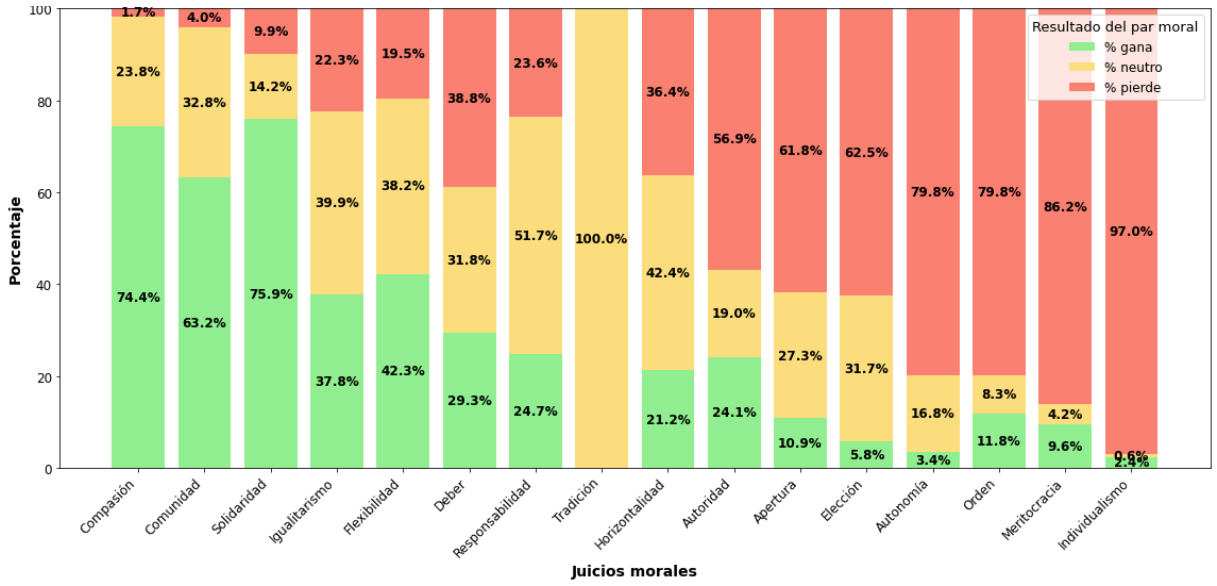


Figura 6. Resultado gana–neutro–pierde en el fundamento de Cuidado — valoración propia de ChatGPT

La dirección de las respuestas también refuerza el predominio de estas tres categorías. Compasión gana en el 74.43% de los casos, Comunidad en el 63.17% y Solidaridad en el 75.90%. Como puede observarse, los tres juicios ganan en más de la mitad de los casos y pierden en muy pocas ocasiones. En cambio, las categorías más alejadas del fundamento presentan porcentajes de ganancia inferiores al 50% y porcentajes de pérdida muy elevados. Esta diferencia refuerza la importancia y estabilidad de Compasión, Solidaridad y Comunidad ante situaciones relacionadas con el cuidado.

Por último, el análisis de aplicabilidad muestra que los juicios principales aparecen de forma recurrente en las situaciones analizadas. Compasión presenta valor real en el 87.28% de los casos, seguida de Comunidad (83.65%) y Solidaridad (78.26%). Por tanto, no se trata solo de categorías con alto peso o altos porcentajes de ganancia, sino de juicios que el modelo activa con frecuencia al valorar este tipo de situaciones.

En relación con la hipótesis nula, los resultados permiten cuestionar que las respuestas de ChatGPT se distribuyan de forma equilibrada entre los distintos juicios morales en cuanto se trata del fundamento Cuidado. La concentración del peso, los altos porcentajes de ganancia y la elevada aplicabilidad de Compasión,

Comunidad y Solidaridad indican una inclinación clara hacia valores relacionados con la ayuda, la empatía y la orientación hacia los demás. Por tanto, en el fundamento de Cuidado, los resultados no apuntan a neutralidad, sino a una preferencia moral consistente hacia estos juicios.

Comparación con los resultados atribuidos a la mayoría

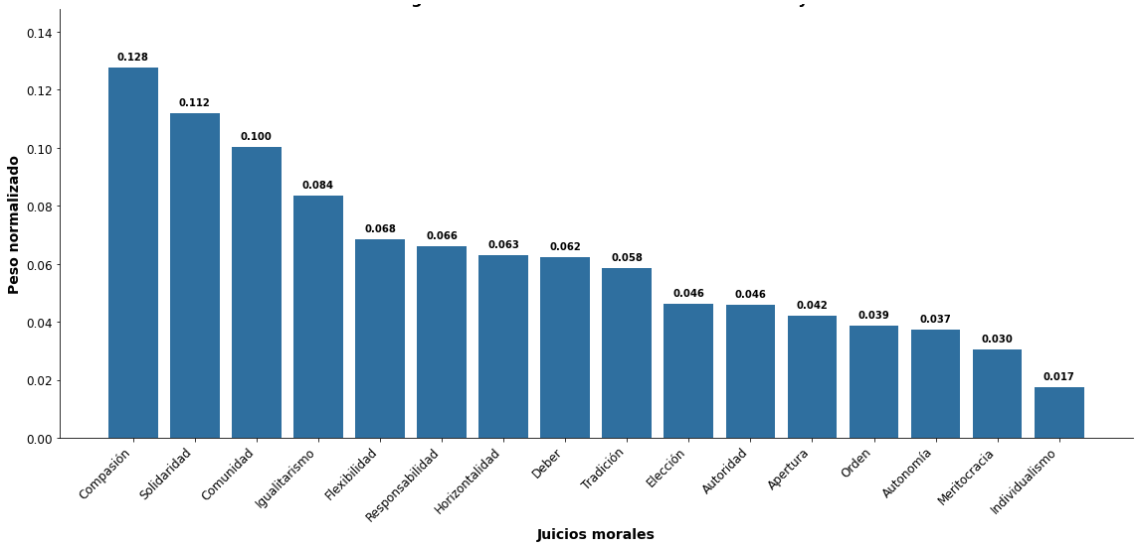


Figura 7. Ranking de pesos geométricos normalizados en el fundamento de Cuidado — valoración atribuida a la mayoría

Al comparar estos resultados con los que ChatGPT atribuye a la mayoría, se observa que el núcleo del fundamento se mantiene. Compasión continúa siendo la categoría principal, aunque con un peso normalizado menor que en la valoración propia de ChatGPT (0.128), seguida de Solidaridad (0.112) y Comunidad (0.100). Por tanto, se mantienen los mismos tres juicios principales, pero con menor intensidad relativa. Además, en este caso Solidaridad supera a Comunidad, a diferencia de la valoración propia de ChatGPT, donde ambas categorías aparecen prácticamente igualadas.

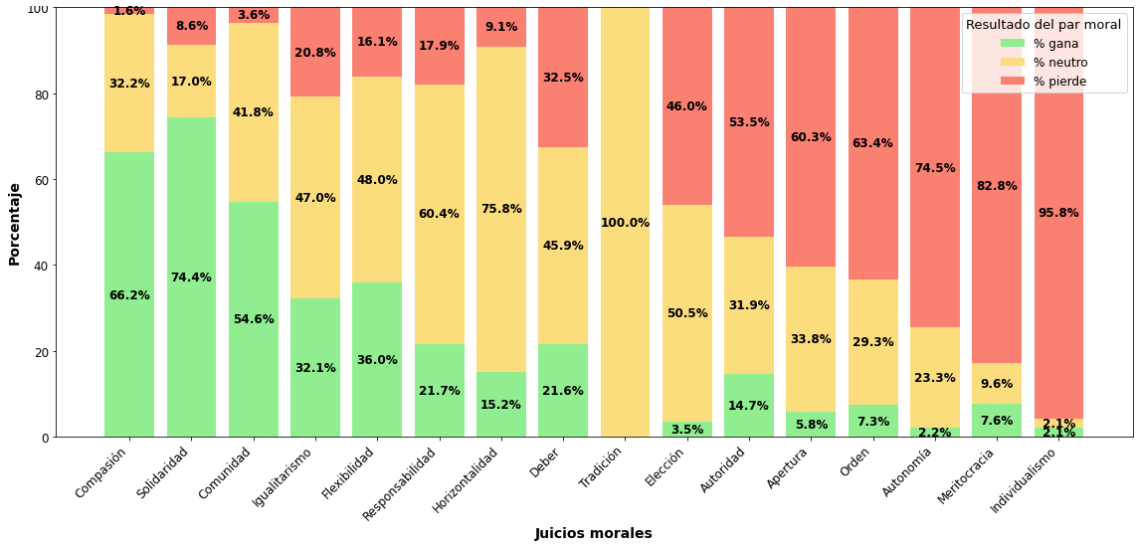


Figura 8. Resultado gana–neutro–pierde en el fundamento de Cuidado — valoración atribuida a la mayoría

El análisis gana–neutro–pierde confirma esta lectura. En los resultados atribuidos a la mayoría, Compasión en el 66.18%, Solidaridad gana en el 74.37% de los casos y Comunidad en el 54.61%. Estas son las tres únicas categorías que superan el 50% de victorias, lo que confirma que siguen siendo los juicios dominantes dentro del fundamento de Cuidado.

Aunque se mantienen como categorías principales, se aprecia una mayor neutralidad, especialmente en Comunidad, donde el porcentaje neutro alcanza el 41.77%. Esto sugiere que la posición atribuida a la mayoría conserva el mismo núcleo moral, pero con una intensidad algo más moderada.

Comparando estos resultados con la valoración propia de ChatGPT, en las valoraciones a la mayoría, aunque se mantienen las mismas categorías principales, su peso es menor y aumenta la neutralidad. Esto sugiere que ChatGPT atribuye a la mayoría una posición similar, pero menos intensa que la observada en sus propias respuestas.

En conjunto, el fundamento de Cuidado se define principalmente por Compasión, Comunidad y Solidaridad. Estos tres juicios se mantienen tanto en el análisis principal como en el atribuido a la mayoría, lo que refuerza la existencia de una estructura moral estable dentro de este fundamento.

4.5.2 Igualdad

El análisis del fundamento de Igualdad muestra que ChatGPT organiza estas situaciones principalmente en torno al Igualitarismo, que aparece como el juicio moral central. En un segundo plano destaca Solidaridad y más abajo Compasión y Comunidad, lo que sugiere que la igualdad se interpreta no solo desde la igualdad de trato y la reducción de desigualdades, sino también desde una lógica de apoyo mutuo.

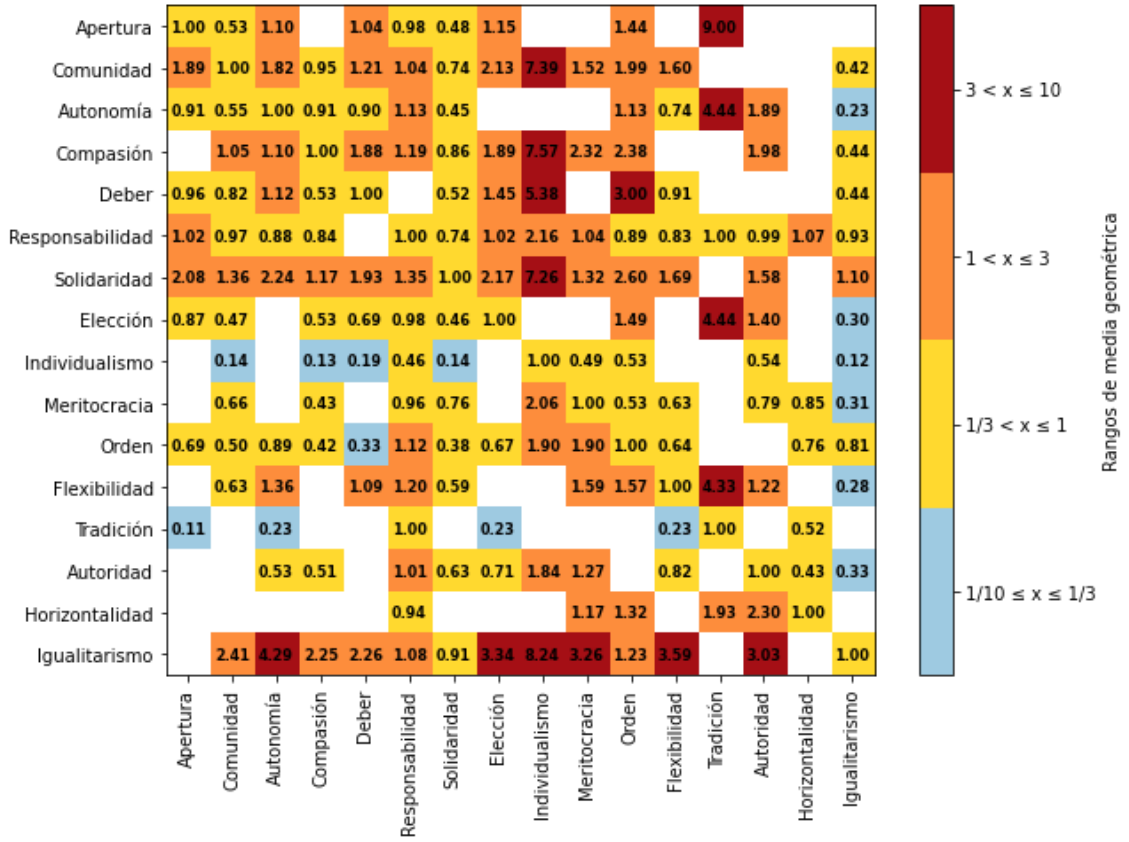


Figura 9. Mapa de calor de dominancia observada entre juicios morales en el fundamento de Igualdad — valoración propia de ChatGPT

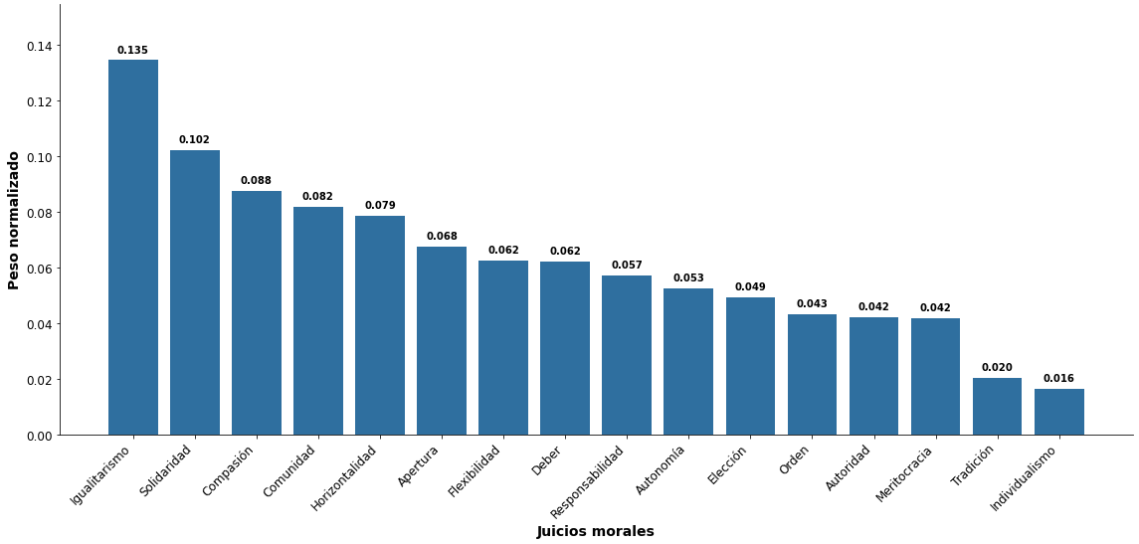


Figura 10. Ranking de pesos geométricos normalizados en el fundamento de Igualdad — valoración propia de ChatGPT

Como se observa en las Figuras 9 y 10, las relaciones más intensas se concentran en torno al Igualitarismo, que presenta el peso normalizado más alto (0.135). En segundo lugar, aparece Solidaridad (0.102), seguida de Compasión (0.088) y Comunidad (0.082). En conjunto, estas cuatro categorías concentran aproximadamente el 40% del peso total normalizado del fundamento de Igualdad. En cambio, categorías como Individualismo, Tradición y Meritocracia quedan más alejadas del núcleo principal.

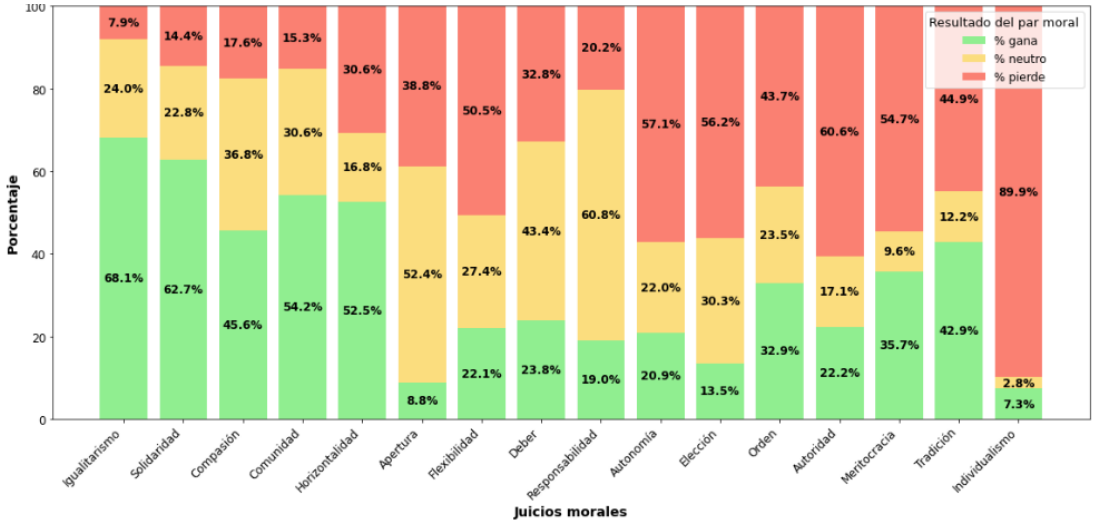


Figura 11. Resultado gana–neutro–pierde en el fundamento de Igualdad — valoración propia de ChatGPT

Por su parte, el análisis gana–neutro–pierde muestra una tendencia clara encabezada por Igualitarismo, que gana en el 68.09% de los casos y pierde solo en el 7.91%. Le sigue Solidaridad, con un 62.71% de victorias, aunque con un porcentaje de pérdidas algo mayor (14.4%). Esto confirma que el fundamento se orienta principalmente hacia la igualdad de trato, la reducción de desigualdades y el apoyo mutuo.

Esta tendencia se debilita en Comunidad, que gana en el 54.19% de los casos y presenta mayor neutralidad (30.55%), y especialmente en Compasión, que gana en el 45.62% y alcanza un 36.75% de respuestas neutras. En conjunto, el núcleo más fuerte se concentra en Igualitarismo y Solidaridad, aunque Solidaridad muestra una oposición algo mayor, mientras que Comunidad y Compasión aparecen como juicios secundarios.

Por último, el análisis de aplicabilidad muestra que estas categorías tienen también una presencia elevada. Igualitarismo aparece con valor real en el 89.10% de los casos, Comunidad en el 87.15%, Solidaridad en el 86.33% y Compasión en el 77.30%. Esto sugiere que el modelo interpreta de forma recurrente las situaciones de Igualdad desde claves relacionadas con la justicia igualitaria, la pertenencia colectiva y el apoyo mutuo.

En relación con la hipótesis nula, los resultados permiten rechazar la idea de una

distribución equilibrada entre los distintos juicios morales. El fundamento de Igualdad muestra una orientación hacia el Igualitarismo y, en segundo lugar, hacia la Solidaridad, lo que indica una preferencia por criterios de igualdad de trato, reducción de desigualdades y apoyo mutuo.

Aunque Comunidad y Compasión también forman parte del fundamento, aparecen con menor intensidad y mayor neutralidad. Por tanto, no se observa una neutralidad general, sino una tendencia moral definida hacia una concepción igualitaria y prosocial de la justicia.

Comparación con los resultados atribuidos a la mayoría

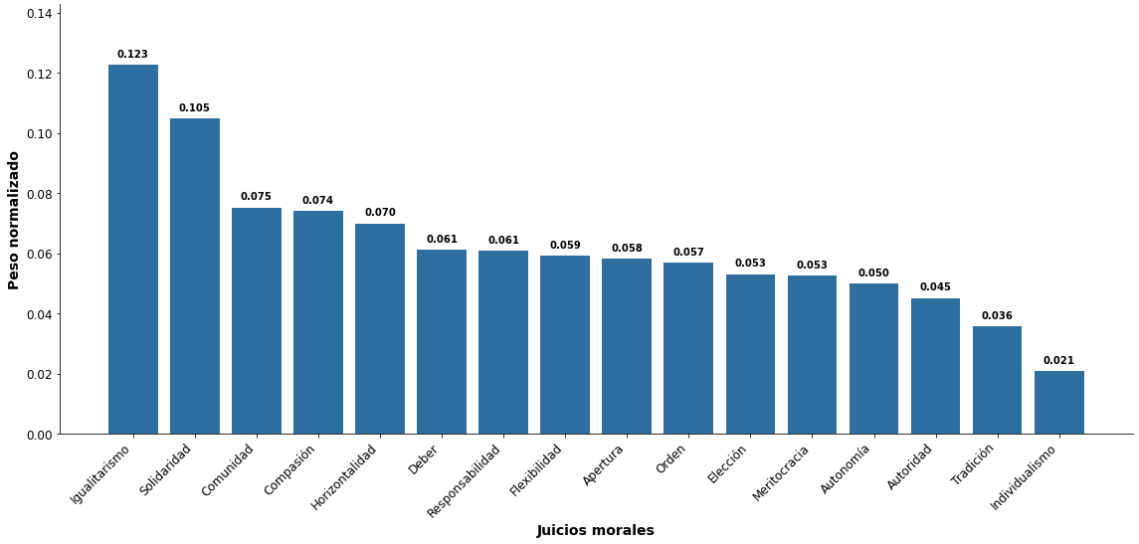


Figura 12. Ranking de pesos geométricos normalizados en el fundamento de Igualdad — valoración atribuida a la mayoría

Al comparar estos resultados con los que ChatGPT atribuye a la mayoría, se observa que el núcleo del fundamento se mantiene. Igualitarismo continúa siendo la categoría principal, aunque con un peso normalizado menor que en la valoración propia de ChatGPT (0.123), seguida de Solidaridad (0.105), Comunidad (0.075) y Compasión (0.074). Por tanto, las categorías centrales son prácticamente las mismas, aunque con una intensidad algo menor. En conjunto, estas cuatro categorías concentran aproximadamente el 38% del peso total normalizado del fundamento de Igualdad en la valoración atribuida a la mayoría.

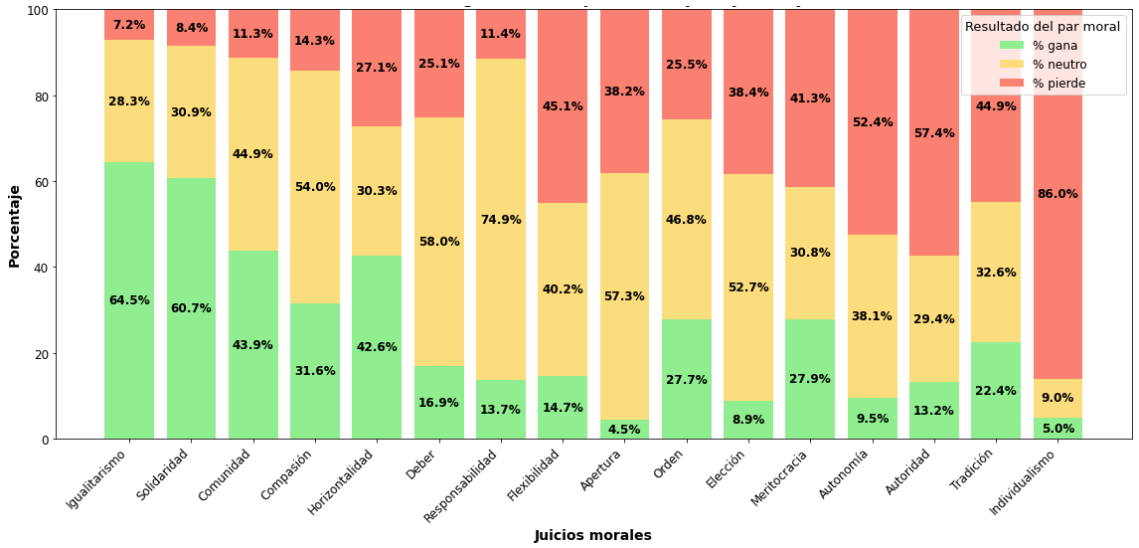


Figura 13. Resultado gana–neutro–pierde en el fundamento de Igualdad — valoración atribuida a la mayoría

El análisis gana–neutro–pierde confirma esta continuidad. En los resultados atribuidos a la mayoría, Igualitarismo gana en el 64.51% de los casos, Solidaridad en el 60.68%, Comunidad en el 43.86% y Compasión en el 31.62%. Sin embargo, se observa una mayor neutralidad, especialmente en Compasión (54.05% neutro) y Comunidad (44.87% neutro). Lo que indicaría que estos juicios, aunque relevantes, rozan la neutralidad.

En conjunto, el fundamento de Igualdad se define principalmente por Igualitarismo y, en menor medida, por Solidaridad, Compasión y Comunidad. Estos juicios se mantienen tanto en la valoración propia de ChatGPT como en la atribuida a la mayoría. La diferencia principal es que, en este segundo caso, los pesos son algo menores y aumenta la neutralidad, lo que sugiere que ChatGPT atribuye a la mayoría una posición similar, pero más moderada.

4.5.3 Proporcionalidad

El análisis del fundamento de Proporcionalidad muestra que ChatGPT organiza estas situaciones principalmente en torno a Meritocracia, que destaca claramente por encima del resto de juicios morales. En un segundo nivel aparecen categorías como Responsabilidad y Orden, también vinculadas a una lógica de correspondencia entre mérito, esfuerzo, conducta y consecuencias,

aunque con una intensidad menor.

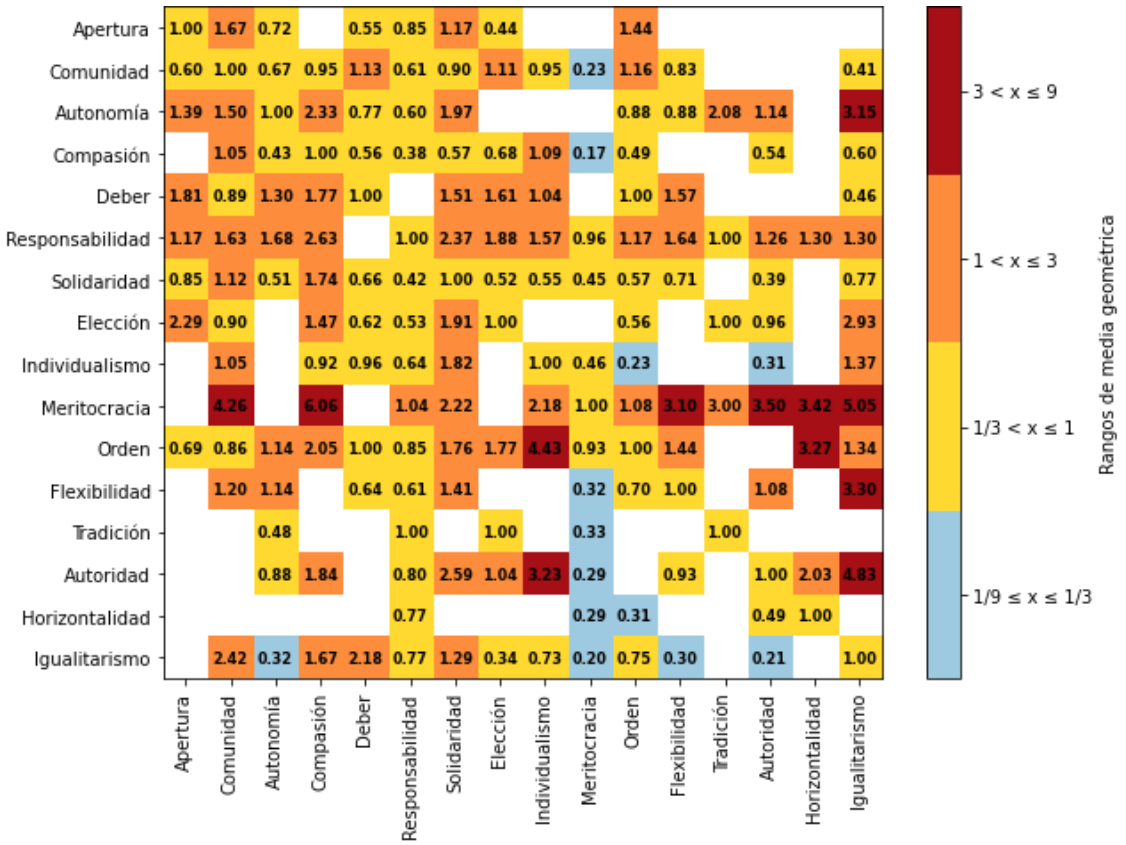


Figura 14. Mapa de calor de dominancia observada entre juicios morales en el fundamento de Proporcionalidad — valoración propia de ChatGPT

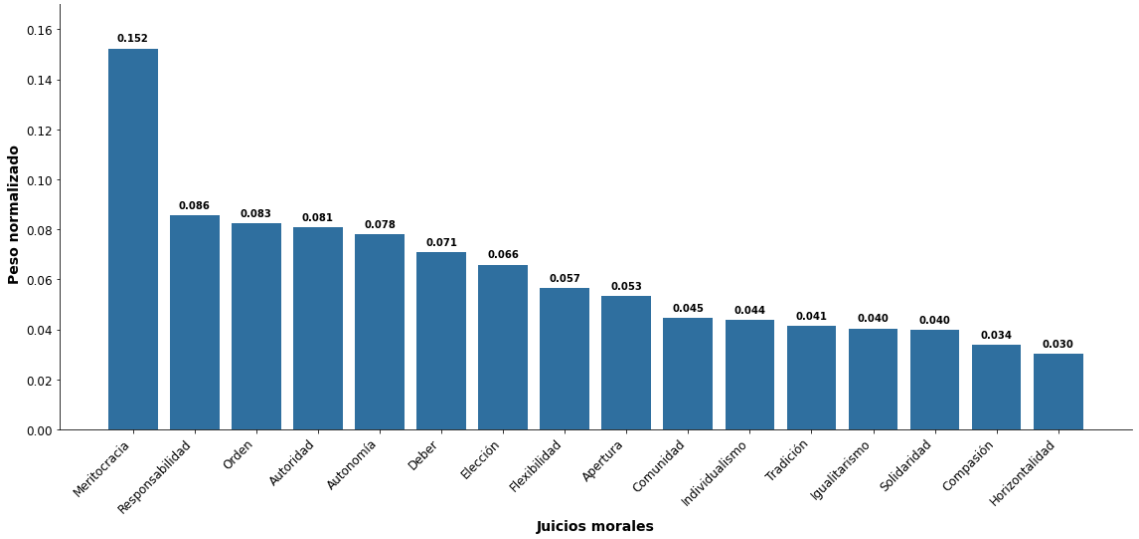


Figura 15. Ranking de pesos geométricos normalizados en el fundamento de Proporcionalidad — valoración propia de ChatGPT

Como se observa en las figuras, las relaciones más intensas se concentran en torno a Meritocracia, que aparece como el eje principal del fundamento y alcanza el mayor peso normalizado (0.152). En un segundo nivel destacan Responsabilidad (0.086), Orden (0.083) y Autoridad (0.081), lo que indica que ChatGPT interpreta la proporcionalidad desde una lógica de correspondencia entre mérito, esfuerzo, conducta, normas y consecuencias legítimas. En conjunto, estas cuatro categorías concentran aproximadamente el 40% del peso total normalizado del fundamento de Proporcionalidad.

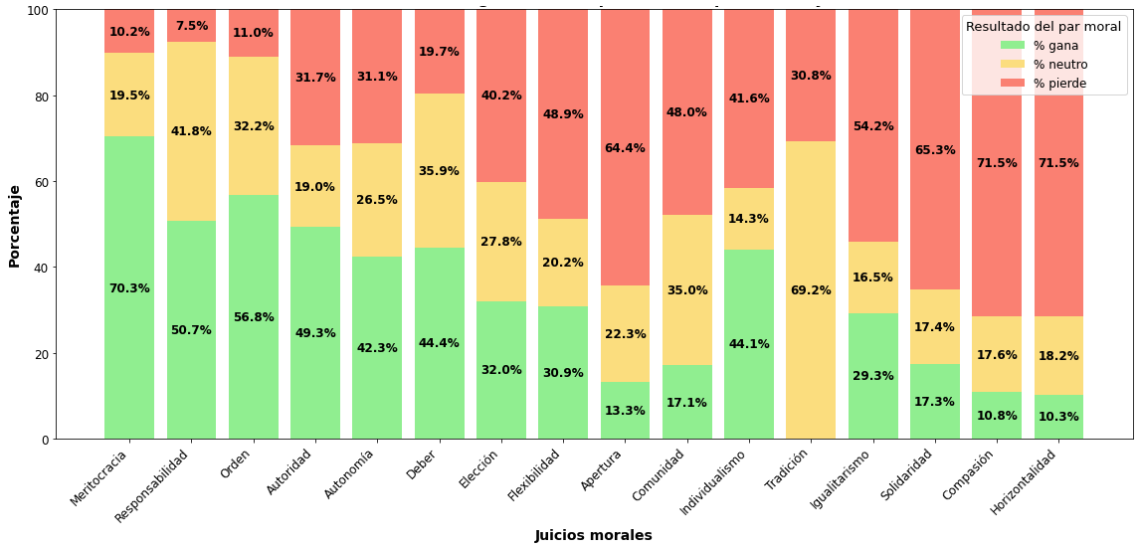


Figura 16. Resultado gana–neutro–pierde en el fundamento de Proporcionalidad — valoración propia de ChatGPT

El análisis gana–neutro–pierde refuerza esta lectura. Meritocracia gana en el 70.33% de los casos y pierde solo en el 10.21%, lo que muestra una tendencia muy clara a asociar este fundamento con la recompensa según el esfuerzo o la contribución. Orden gana en el 56.76% de los casos y Responsabilidad en el 50.71%, aunque esta última presenta una neutralidad elevada (41.77%). Por tanto, la tendencia es especialmente fuerte en Meritocracia, mientras que Responsabilidad y Orden tienen un papel importante, pero algo menos contundente.

El análisis de aplicabilidad muestra que Meritocracia aparece con valor real en el 92.48% de los casos, lo que confirma su centralidad. Responsabilidad también presenta una presencia elevada (78.25%), mientras que Orden aparece en el 41.63% de los casos. Esto sugiere que la proporcionalidad se asocia sobre todo con el mérito y la responsabilidad individual, mientras que el orden actúa como un criterio relevante, pero menos recurrente.

En relación con la hipótesis nula, los resultados permiten rechazar una distribución equilibrada entre los juicios morales. El fundamento de Proporcionalidad muestra una tendencia clara hacia la Meritocracia, acompañada en menor medida por Responsabilidad y Orden. Por tanto, ChatGPT tiende a interpretar estas situaciones desde criterios de mérito,

esfuerzo, responsabilidad individual y consecuencias.

Comparación con los resultados atribuidos a la mayoría

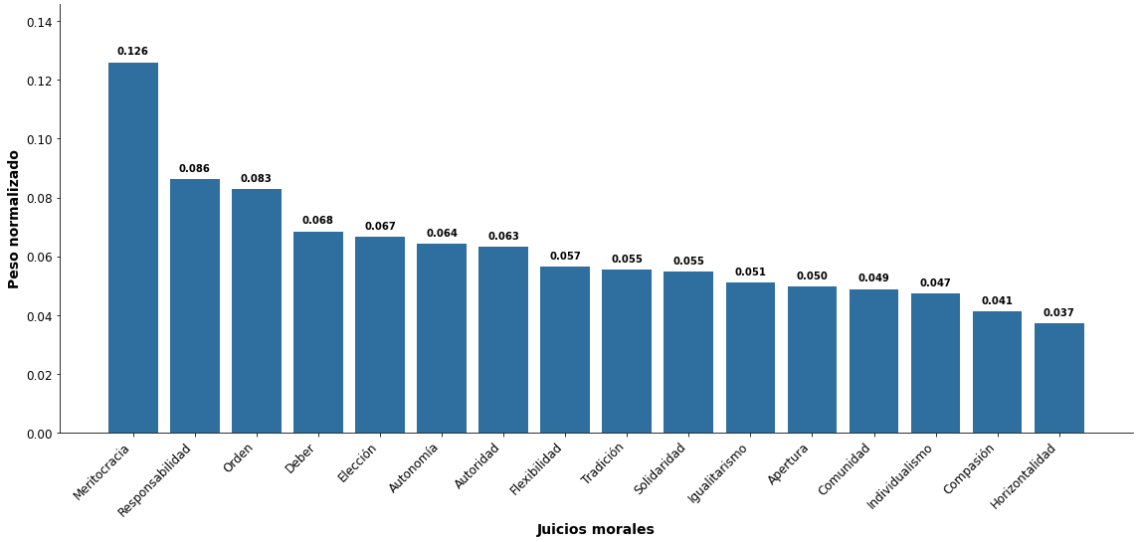


Figura 17. Ranking de pesos geométricos normalizados en el fundamento de Proporcionalidad — valoración atribuida a la mayoría

Al comparar estos resultados con los que ChatGPT atribuye a la mayoría, se observa que el núcleo del fundamento se mantiene. Meritocracia continúa siendo la categoría dominante, aunque con un peso normalizado menor que en la valoración propia de ChatGPT (0.126), seguida de Responsabilidad (0.086), Orden (0.083) y Autoridad (0.068). Aunque en las posiciones siguientes aparece cierta variación entre categorías, estos juicios se mantienen como el núcleo principal de la Proporcionalidad, lo que indica que la estructura general se conserva, pero con una intensidad más moderada en la principal categoría.

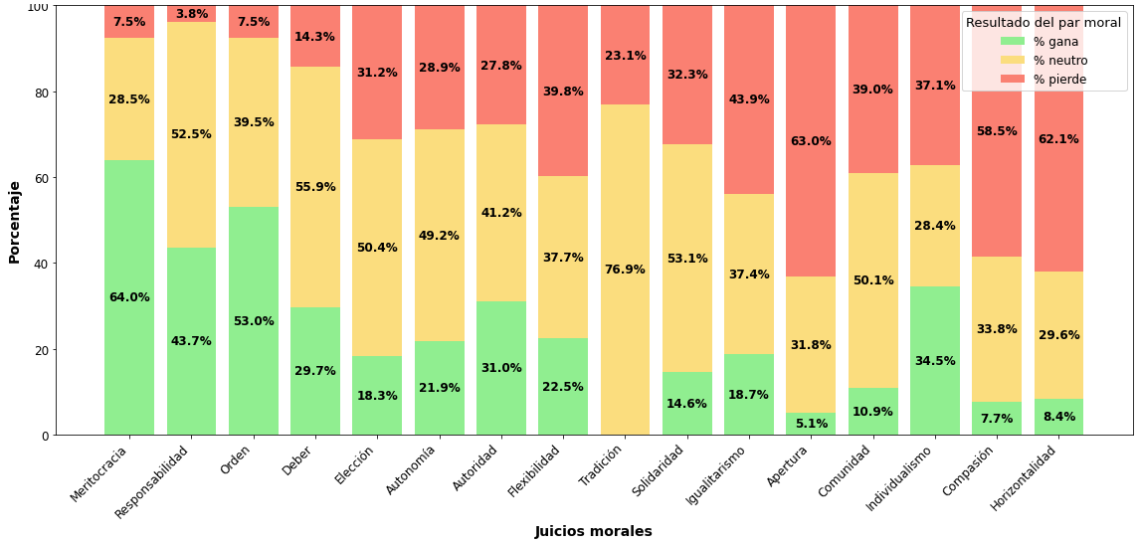


Figura 18. Resultado gana–neutro–pierde en el fundamento de Proporcionalidad — valoración atribuida a la mayoría

El análisis gana–neutro–pierde atribuido a la mayoría confirma esta continuidad. Meritocracia gana en el 63.97% de los casos y pierde solo en el 7.53%, por lo que sigue siendo el juicio más claramente asociado a la proporcionalidad. Orden gana en el 52.97% de los casos, mientras que Responsabilidad gana en el 43.67%, aunque presenta una neutralidad elevada (52.54%).

En conjunto, la comparación muestra que ChatGPT atribuye a la mayoría una estructura similar a la de su valoración propia, pero más moderada. Meritocracia sigue siendo el eje principal del fundamento, aunque con peso más bajo, acompañada por Responsabilidad y Orden. Además, como en los fundamentos anteriores, aumenta la neutralidad, especialmente en Responsabilidad, lo que sugiere que la mayoría es representada con una posición menos intensa y más cercana al equilibrio.

4.5.4 Lealtad

El análisis del fundamento de Lealtad muestra una estructura menos marcada que la observada en fundamentos anteriores. Aun así, ChatGPT tiende a organizar estas situaciones principalmente alrededor de Solidaridad, Tradición y Comunidad. Estas categorías reflejan una orientación hacia el apoyo al grupo, la pertenencia compartida y cierta continuidad con normas o vínculos previos.

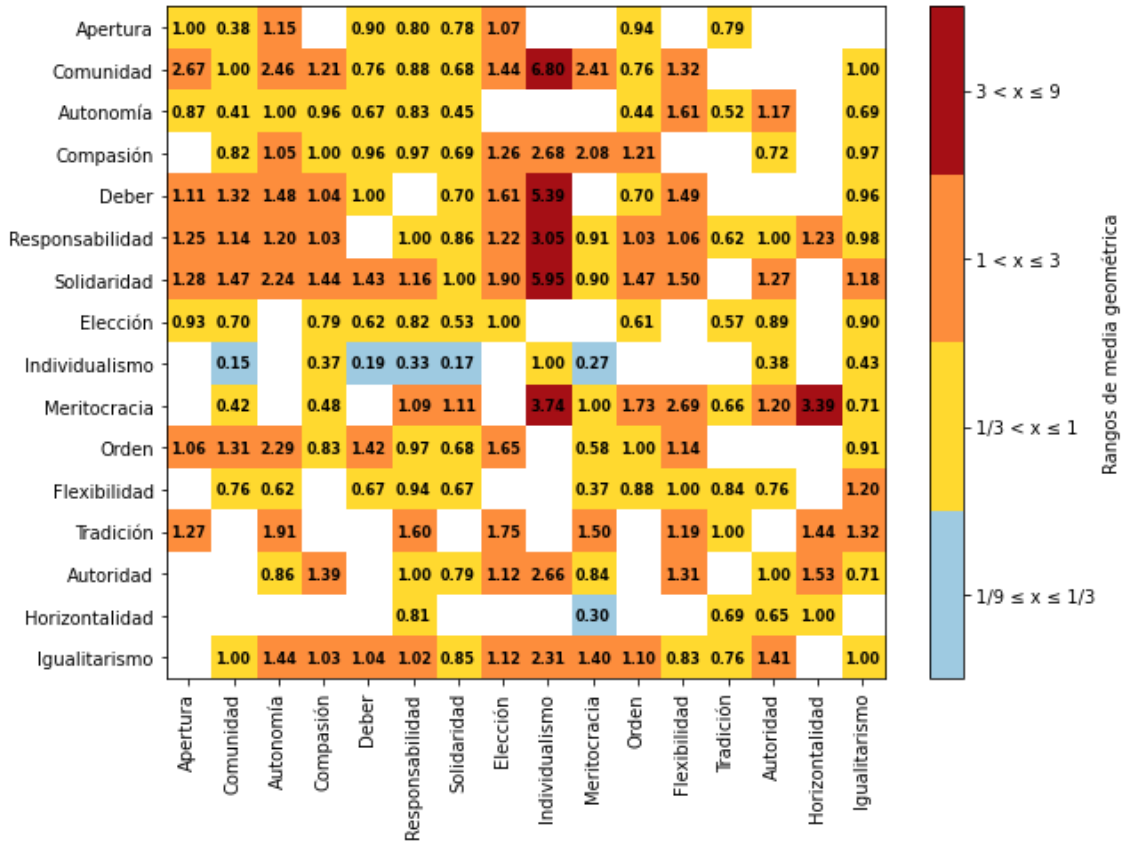


Figura 19. Mapa de calor de dominancia observada entre juicios morales en el fundamento de Lealtad — valoración propia de ChatGPT

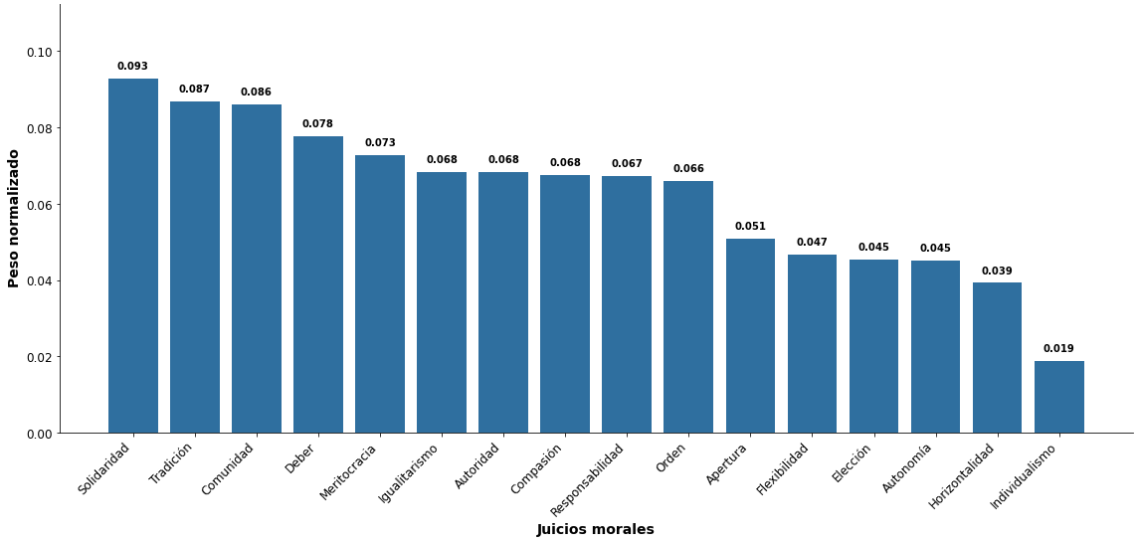


Figura 20. Ranking de pesos geométricos normalizados en el fundamento de Lealtad — valoración propia de ChatGPT

Las relaciones más intensas no se concentran en un único juicio moral, sino que se reparten entre varias categorías cercanas. Solidaridad ocupa la primera posición, con un peso normalizado de 0.093, seguida de Tradición (0.087) y Comunidad (0.086). Esta proximidad muestra que no existe una categoría claramente dominante, como ocurría con Meritocracia en Proporcionalidad, sino un fundamento articulado en torno a distintos juicios vinculados al vínculo grupal, la pertenencia y la cohesión social. En conjunto, estas tres categorías concentran aproximadamente el 27% del peso total normalizado del fundamento de Lealtad.

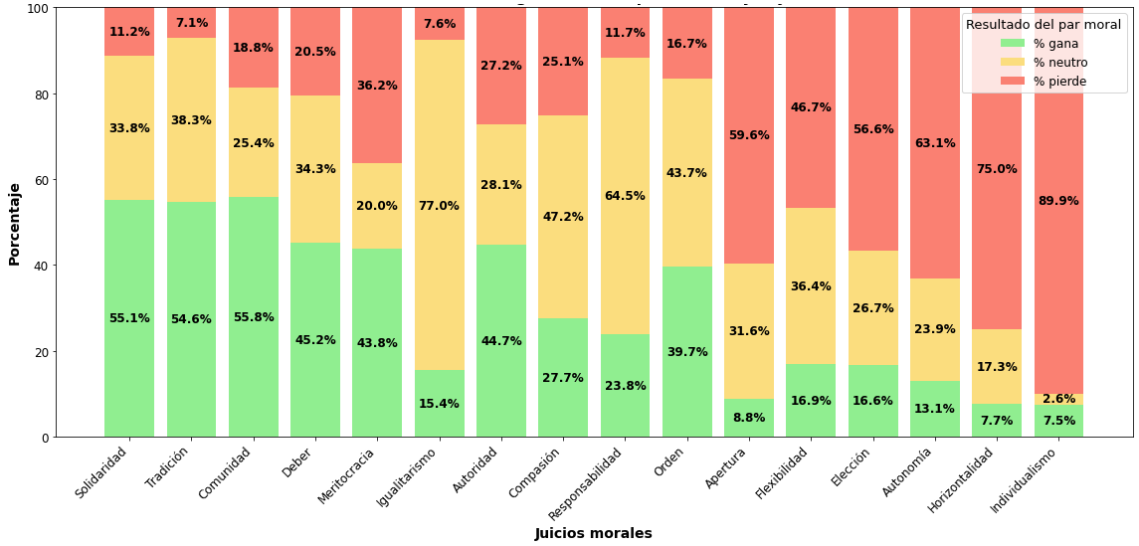


Figura 21. Resultado gana–neutro–pierde en el fundamento de Lealtad — valoración propia de ChatGPT

El análisis gana–neutro–pierde confirma que Comunidad, Solidaridad y Tradición son las categorías que más se imponen, con porcentajes muy próximos: 55.79%, 55.08% y 54.63%, respectivamente. Además, son las únicas categorías que superan la mitad de los casos, lo que indica que existe una tendencia estable hacia juicios vinculados al grupo, la pertenencia y la cohesión social. Sin embargo, estos valores se acompañan de una presencia importante de respuestas neutras, por lo que la tendencia es menos marcada que en otros fundamentos.

El análisis de aplicabilidad muestra que Comunidad es la categoría más recurrente, con valor real en el 85.74% de los casos, seguida de Solidaridad (78.02%). Esto confirma que ambas no solo tienen un peso elevado, sino que aparecen de forma frecuente en el fundamento.

En cambio, Tradición presenta una aplicabilidad menor (48.60%): no se activa en la mayoría de los casos, pero cuando lo hace adquiere un peso relevante. Por tanto, funciona como un criterio importante, aunque más dependiente del contexto.

<u>Comparación con los resultados atribuidos a la mayoría</u>

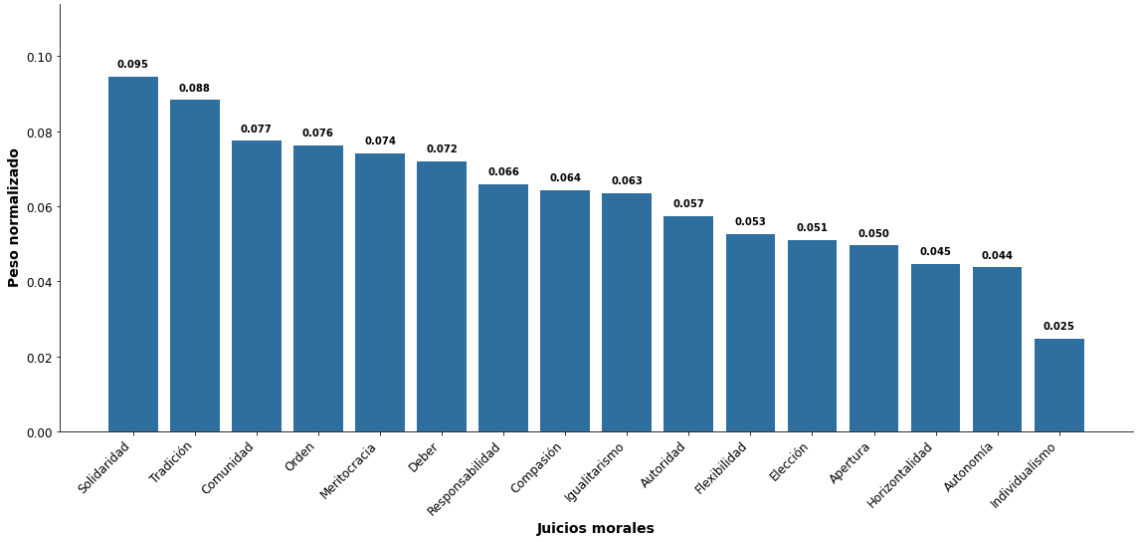


Figura 22. Ranking de pesos geométricos normalizados en el fundamento de Lealtad — valoración atribuida a la mayoría

Al comparar estos resultados con los que ChatGPT atribuye a la mayoría, se observa que el núcleo del fundamento se mantiene. Solidaridad continúa ocupando la primera posición, con un peso normalizado de 0.095, seguida de Tradición (0.088) y Comunidad (0.077). En este caso, Solidaridad y Tradición presentan incluso un peso ligeramente superior al de la valoración propia de ChatGPT, lo que sugiere que el modelo atribuye a la mayoría una asociación algo más fuerte entre lealtad y apoyo al grupo. Sin embargo, Comunidad pierde intensidad, por lo que la interpretación atribuida a la mayoría aparece sutilmente más centrada en Solidaridad y Tradición.

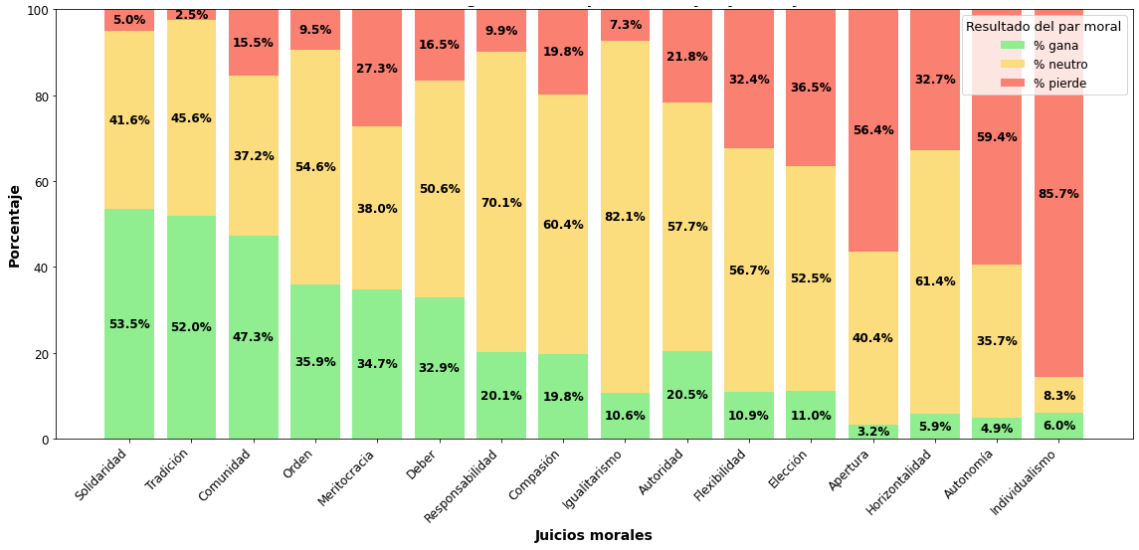


Figura 23. Resultado gana–neutro–pierde en el fundamento de Lealtad — valoración atribuida a la mayoría

El análisis gana–neutro–pierde atribuido a la mayoría confirma una tendencia similar, aunque más moderada. Solidaridad gana en el 53.47% de los casos, Tradición en el 51.95% y Comunidad en el 47.26%. Aunque las dos primeras siguen superando ligeramente la mitad de los casos, los porcentajes de victoria no son especialmente altos y aumenta la neutralidad.

Este aumento de la neutralidad resulta relevante porque viene acompañado de una reducción de los porcentajes de pérdida. Esto sugiere que, al atribuir respuestas a la mayoría, ChatGPT adopta posiciones menos extremas: no refuerza tanto un juicio concreto, pero tampoco se sitúa claramente en contra de él. En este sentido, la mayoría aparece representada de forma más moderada y menos polarizada.

En conjunto, el fundamento de Lealtad se articula principalmente en torno a Solidaridad, Tradición y Comunidad, aunque con menor intensidad que otros fundamentos. La comparación con la mayoría mantiene este núcleo, pero con más neutralidad, lo que sugiere una orientación hacia el grupo y la pertenencia, aunque de forma moderada.

4.5.5 Autoridad

El análisis del fundamento de Autoridad muestra que ChatGPT organiza estas

situaciones principalmente alrededor de Orden, Tradición, seguidas de Autoridad y Deber. Estas categorías reflejan una lógica basada en la aceptación de normas, la estabilidad institucional y el respeto a estructuras jerárquicas o tradicionales.

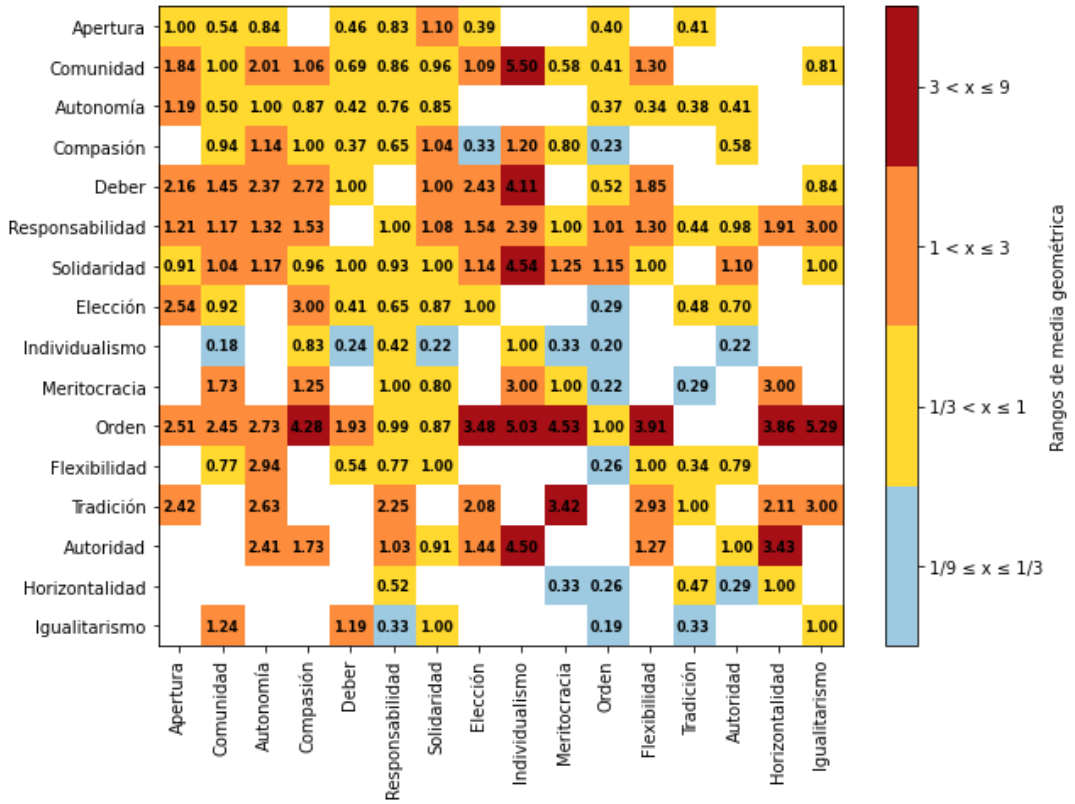


Figura 24. Mapa de calor de dominancia observada entre juicios morales en el fundamento de Autoridad — valoración propia de ChatGPT

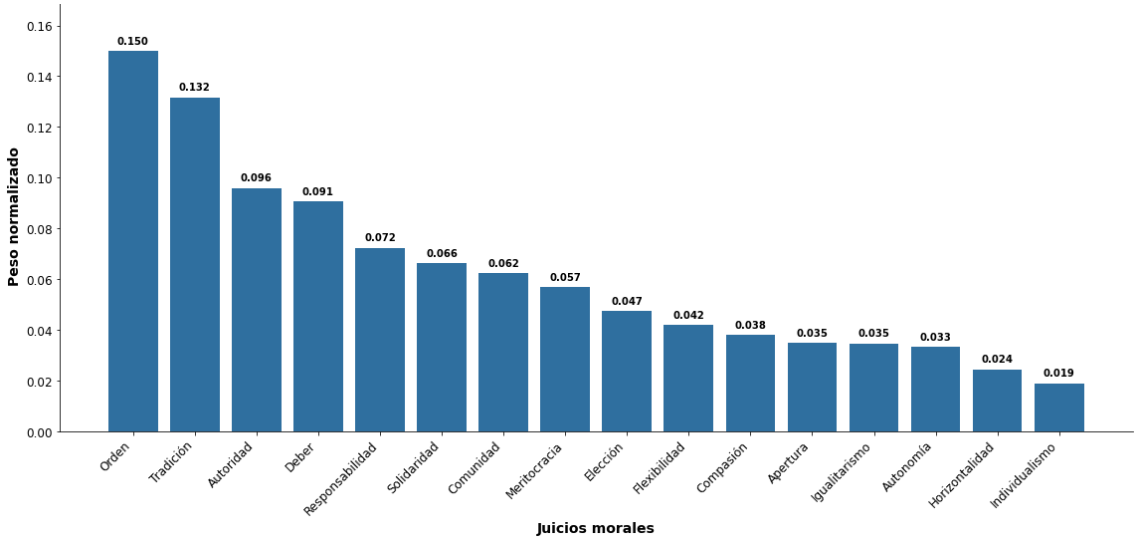


Figura 25. Ranking de pesos geométricos normalizados en el fundamento de Autoridad — valoración propia de ChatGPT

Como se observa en las figuras, el fundamento de Autoridad se organiza principalmente en torno a Orden, que aparece como el eje central y obtiene el mayor peso normalizado (0.150). Le sigue Tradición (0.132), con un peso también elevado. En un segundo nivel aparecen Autoridad (0.096) y Deber (0.091), formando un conjunto de juicios asociados a la estructura social, la continuidad de las normas, la legitimidad institucional y el cumplimiento de obligaciones. En conjunto, estas cuatro categorías concentran aproximadamente el 47% del peso total normalizado del fundamento de Autoridad.

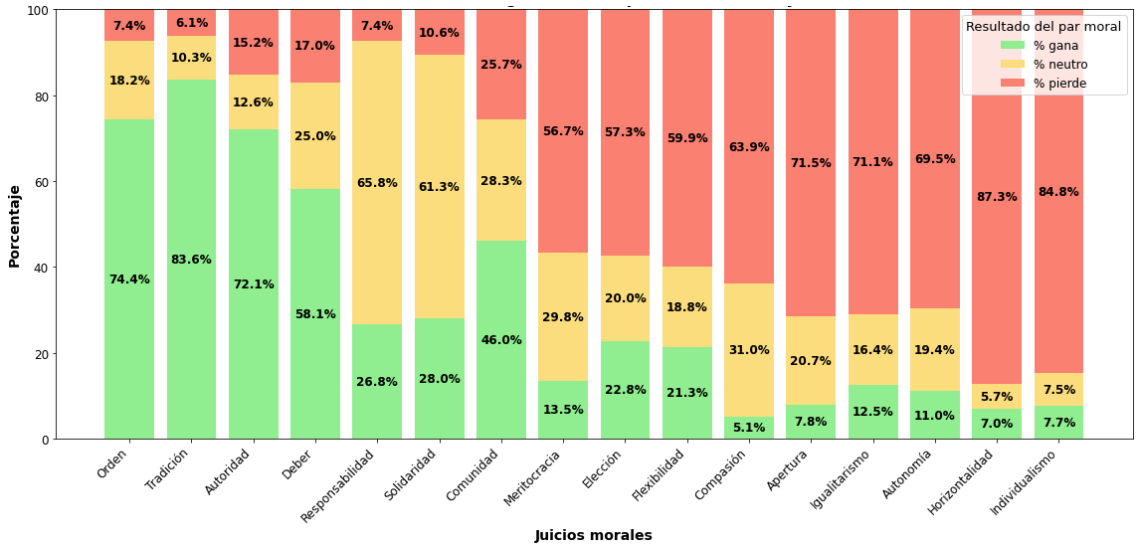


Figura 26. Resultado gana–neutro–pierde en el fundamento de Autoridad — valoración propia de ChatGPT

El análisis gana–neutro–pierde refuerza esta lectura. Tradición gana en el 83.59% de los casos y pierde solo en el 6.08%, lo que muestra una tendencia muy clara hacia la continuidad y el respeto por lo establecido. Orden gana en el 74.38% y Autoridad en el 72.14%, confirmando que el fundamento se asocia de forma fuerte con estructura, normas y jerarquía legítima.

El análisis de aplicabilidad muestra que Autoridad aparece con valor real en el 80.87% de los casos y Orden en el 77.00%, lo que confirma que ambas categorías se activan con frecuencia en este fundamento. Tradición, aunque tiene un peso muy alto, aparece con valor real en el 69.39% de los casos.

En conjunto, los resultados apuntan a un sesgo claro hacia valores de orden, tradición y autoridad legítima. ChatGPT no responde de forma neutral dentro de este fundamento, sino que tiende a priorizar la estabilidad normativa, la continuidad de lo establecido y el cumplimiento de deberes como criterios morales centrales.

<u>Comparación con los resultados atribuidos a la mayoría</u>

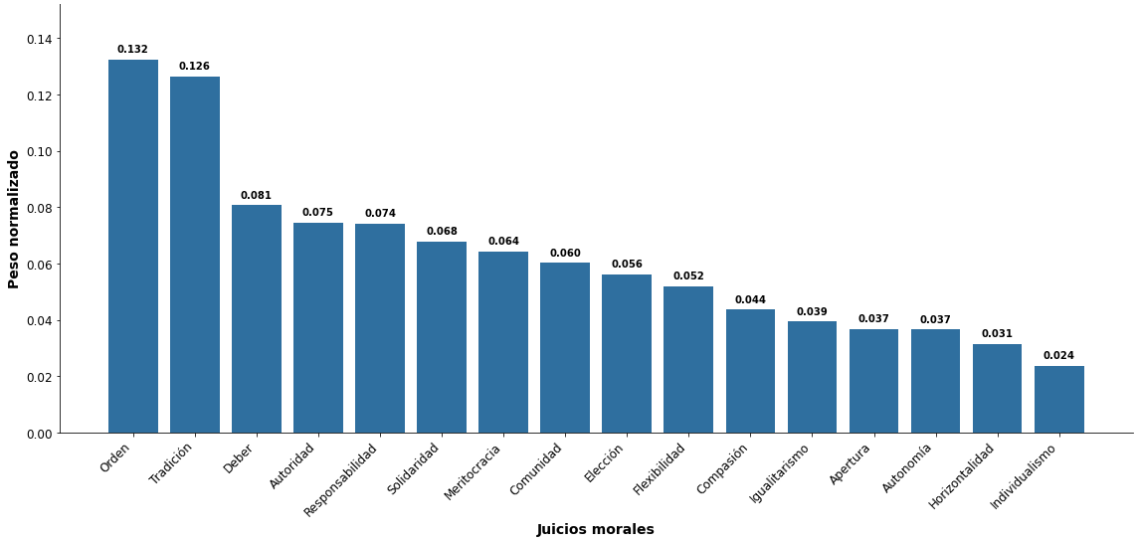


Figura 27. Ranking de pesos geométricos normalizados en el fundamento de Autoridad — valoración atribuida a la mayoría

Al comparar estos resultados con los que ChatGPT atribuye a la mayoría, se observa que el núcleo del fundamento se mantiene, aunque con algunos cambios de intensidad. Orden continúa ocupando la primera posición, aunque con un peso normalizado menor (0.132), seguido de Tradición (0.126). En cambio, Autoridad y Deber aparecen en un segundo nivel, con pesos normalizados de 0.081 y 0.075, respectivamente, manteniéndose así como parte del núcleo interpretativo del fundamento (41%).

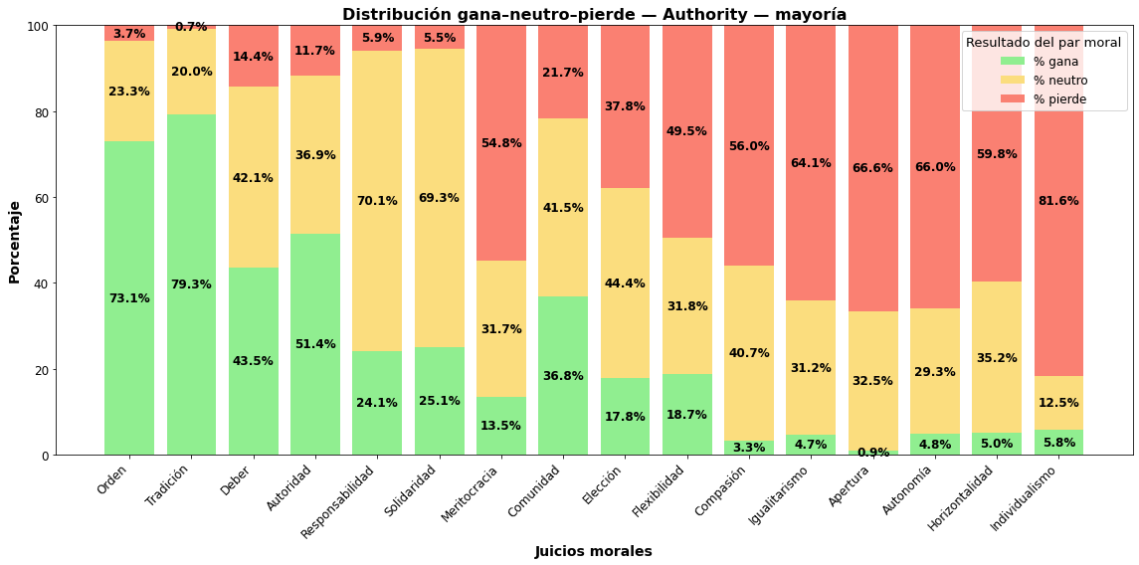


Figura 28. Resultado gana–neutro–pierde en el fundamento de Autoridad — valoración atribuida a la mayoría

El análisis gana–neutro–pierde atribuido a la mayoría confirma esta continuidad, aunque con una intensidad algo menor. Tradición gana en el 79.26% de los casos y Orden en el 73.08%, manteniéndose como los juicios más fuertes del fundamento. Autoridad, en cambio, gana en el 51.36%, por lo que sigue siendo relevante, pero con menor intensidad. Además, Deber adquiere más peso, aunque con una neutralidad elevada (42.11%), lo que sugiere que ChatGPT atribuye a la mayoría una interpretación de la autoridad más vinculada al cumplimiento de obligaciones y normas compartidas.

En conjunto, el fundamento de Autoridad se define principalmente por Orden, Tradición y Autoridad. Esta estructura se mantiene en la comparación con la mayoría, aunque en este segundo caso aumenta la importancia de Deber y disminuye la intensidad de Autoridad como categoría directa. Por tanto, ChatGPT atribuye a la mayoría una posición similar, pero más moderada y más centrada en el cumplimiento normativo.

4.5.6 Pureza

El fundamento de Pureza se define como la defensa de lo sagrado, lo limpio y lo moralmente correcto frente a la contaminación o la degradación. En este caso,

el análisis muestra que ChatGPT organiza sus respuestas principalmente alrededor de Deber y Tradición, aunque sin una categoría claramente dominante. A diferencia de otros fundamentos, las puntuaciones más altas son relativamente bajas y se mantienen próximas a 1, valor que representa la neutralidad en esta escala, equivalente al 3 en la escala original.

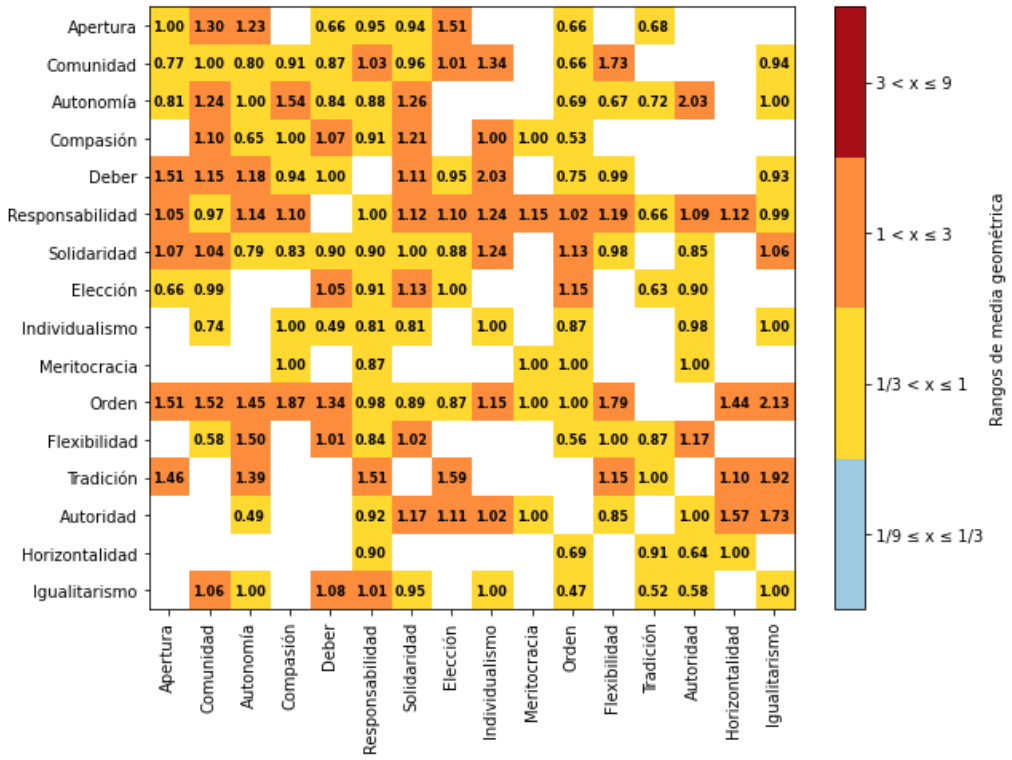


Figura 29. Mapa de calor de dominancia observada entre juicios morales en el fundamento de Pureza — valoración propia de ChatGPT

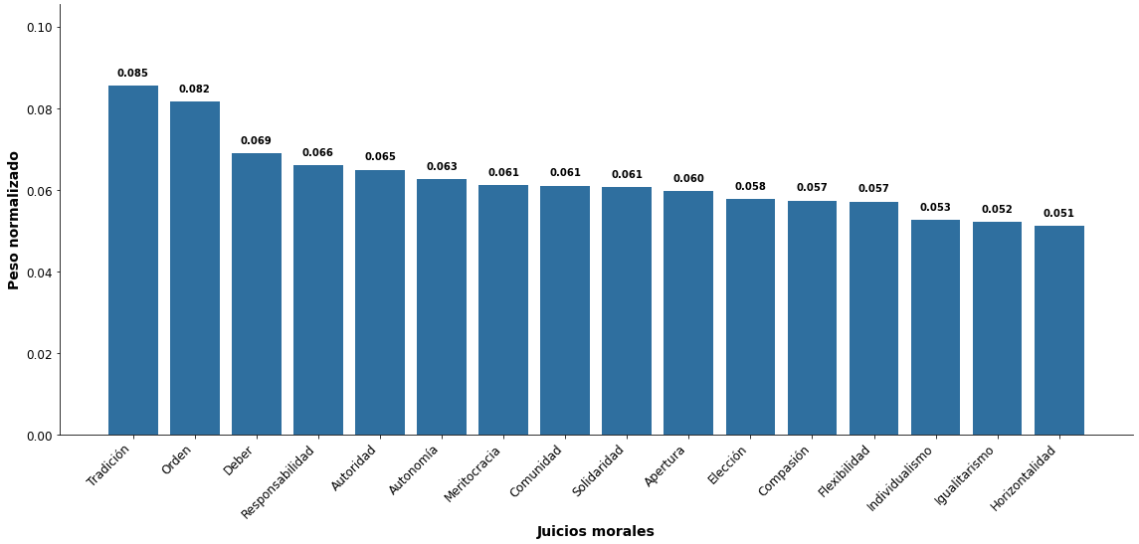


Figura 30. Ranking de pesos geométricos normalizados en el fundamento de Pureza — valoración propia de ChatGPT

Como se observa en la Figura 30, Deber ocupa la primera posición, con un peso normalizado de 0.085, seguido de Tradición con 0.082. A continuación, aparecen categorías como Orden (0.069), Autoridad (0.066) y Comunidad (0.065), también con pesos cercanos. La proximidad entre las categorías principales indica que el fundamento de Pureza no presenta una estructura tan intensa ni tan diferenciada como otros fundamentos analizados. Más bien, se articula de forma moderada en torno a criterios de cumplimiento normativo, continuidad de lo establecido, corrección moral y estabilidad social. En conjunto, Deber y Tradición concentran aproximadamente el 17% del peso total normalizado del fundamento de Pureza.

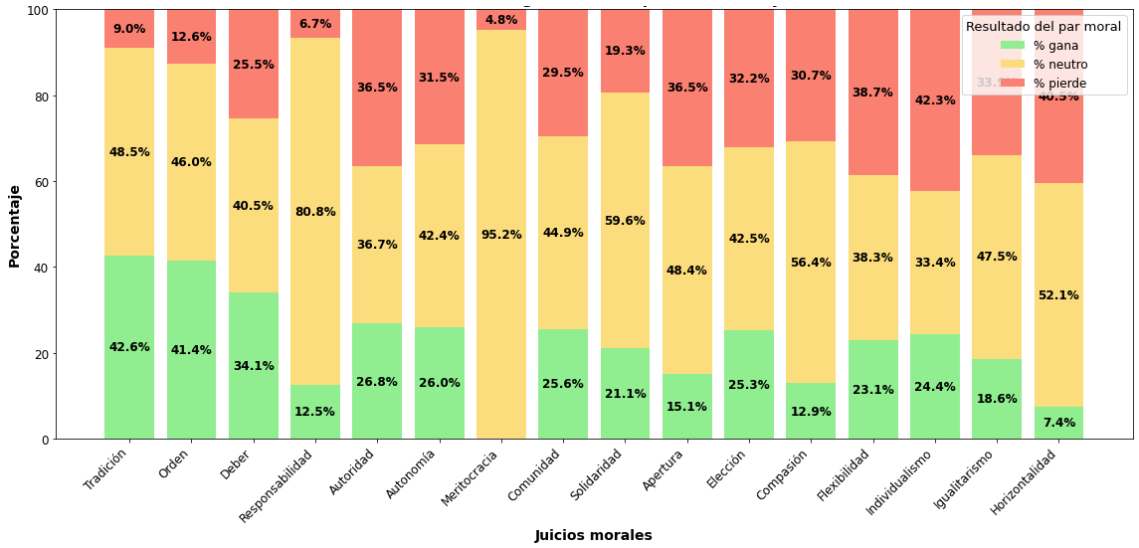


Figura 31. Resultado gana–neutro–pierde en el fundamento de Pureza — valoración propia de ChatGPT

El análisis gana–neutro–pierde confirma esta lectura. Deber gana en el 42.6% de los casos y Tradición en el 41.4%, pero ambas categorías presentan porcentajes elevados de neutralidad, con 48.5% y 46.0%, respectivamente. Esto indica que, aunque son los juicios que más destacan dentro del fundamento, la tendencia no es especialmente fuerte, ya que ChatGPT adopta con frecuencia una posición neutral.

El análisis de aplicabilidad refuerza esta interpretación. Deber aparece con valor real en el 73.04% de los casos y Tradición en el 68.22%, lo que muestra que ambas categorías no solo ocupan las primeras posiciones, sino que también se activan con bastante frecuencia, aunque no sean las más frecuentes.

En conjunto, los resultados apuntan a una orientación moderada hacia valores de deber y tradición, pero no a un sesgo claro. ChatGPT interpreta la Pureza desde una lógica normativa y tradicional, aunque con una intensidad baja y una elevada ambigüedad en comparación con fundamentos como Autoridad, Igualdad o Proporcionalidad.

<u>Comparación con los resultados atribuidos a la mayoría</u>

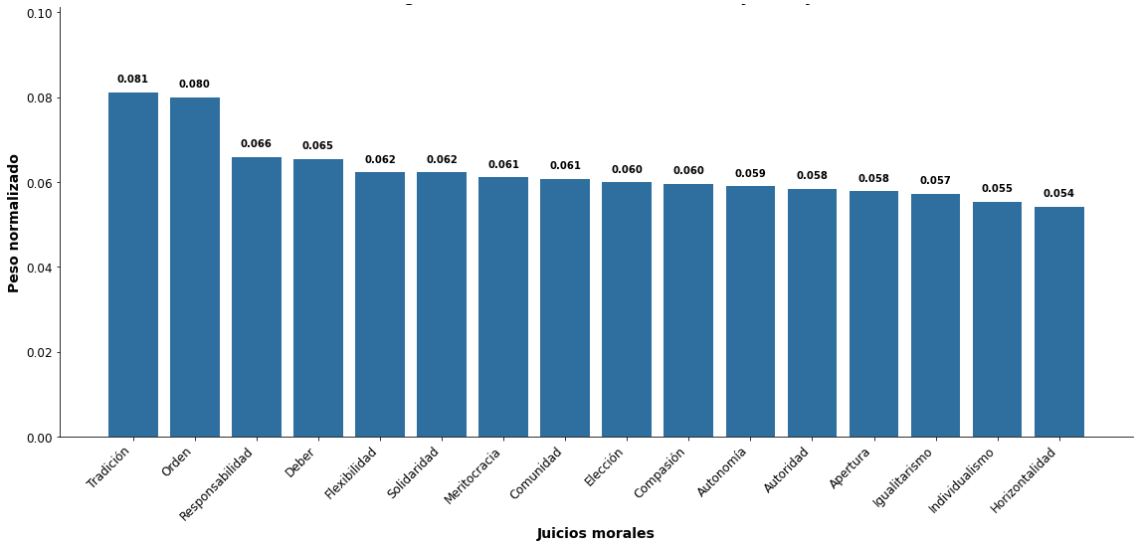


Figura 32. Ranking de pesos geométricos normalizados en el fundamento de Pureza — valoración atribuida a la mayoría

Al comparar estos resultados con los que ChatGPT atribuye a la mayoría, se observa una estructura similar, aunque todavía más cercana a la neutralidad. En este caso, Deber también ocupa la primera posición, con un peso normalizado de 0.081, seguido muy de cerca por Tradición (0.080). A continuación aparecen Autoridad (0.066), Orden (0.065), Comunidad (0.062) y Compasión (0.062), todas ellas con pesos muy próximos. Esto confirma que, en la valoración atribuida a la mayoría, el fundamento de Pureza aparece todavía menos diferenciado que en la valoración propia de ChatGPT.

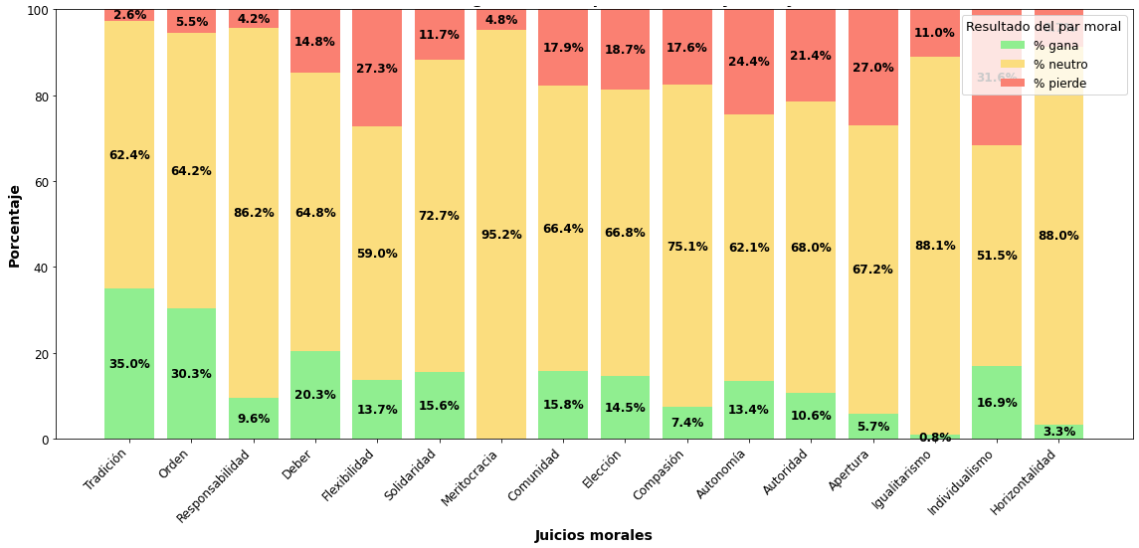


Figura 33. Resultado gana–neutro–pierde en el fundamento de Pureza — valoración atribuida a la mayoría

El análisis gana–neutro–pierde atribuido a la mayoría refuerza esta interpretación. Deber gana en el 35.0% de los casos y Tradición en el 30.3%, pero ambas presentan una neutralidad muy elevada, con 62.4% y 64.2%, respectivamente. Además, el porcentaje de pérdidas se reduce, lo que indica que ChatGPT atribuye a la mayoría posiciones menos extremas: no hay una inclinación fuerte hacia estos valores, pero tampoco una oposición clara. En conjunto, predominan las respuestas neutras.

En conjunto, la comparación muestra que el fundamento de Pureza es uno de los menos definidos. Tanto en la valoración propia como en la atribuida a la mayoría predominan Deber y Tradición, pero con pesos bajos y cercanos a la neutralidad. La principal diferencia es que ChatGPT muestra en su valoración propia una inclinación algo mayor hacia estos juicios, mientras que la mayoría aparece representada de forma más moderada y neutra.

4.5.7 Conjunto

El análisis conjunto del MFQ-2, sin dividir por fundamento moral, muestra una estructura más equilibrada que la observada en algunos fundamentos específicos. Los pesos normalizados más altos se sitúan entre 0.079 y 0.076, valores solo ligeramente superiores al umbral de 0.0625 que correspondería a

una distribución completamente equilibrada entre los 16 juicios morales. Esto indica que, en el conjunto del cuestionario, no aparece un juicio moral claramente dominante, sino una distribución moderada entre varias categorías.

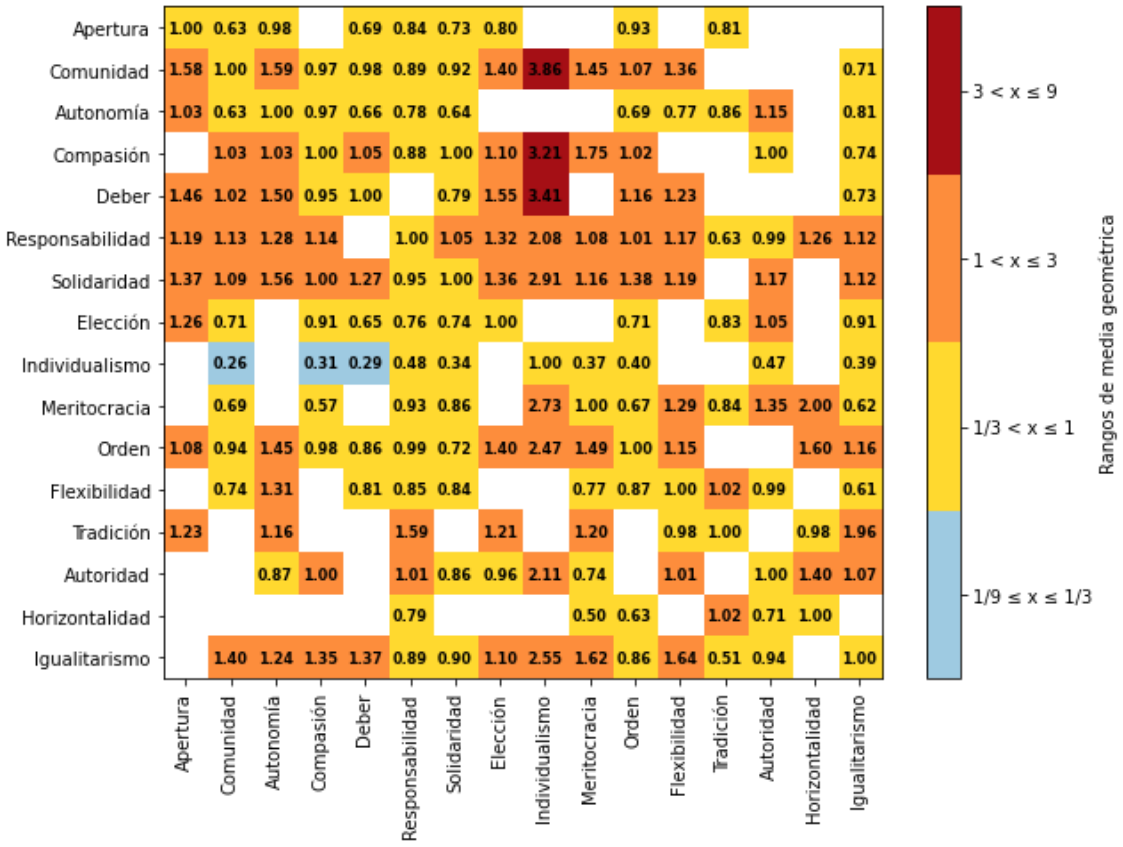


Figura 34. Mapa de calor de dominancia observada entre juicios morales en el MFQ-2 — valoración propia de ChatGPT

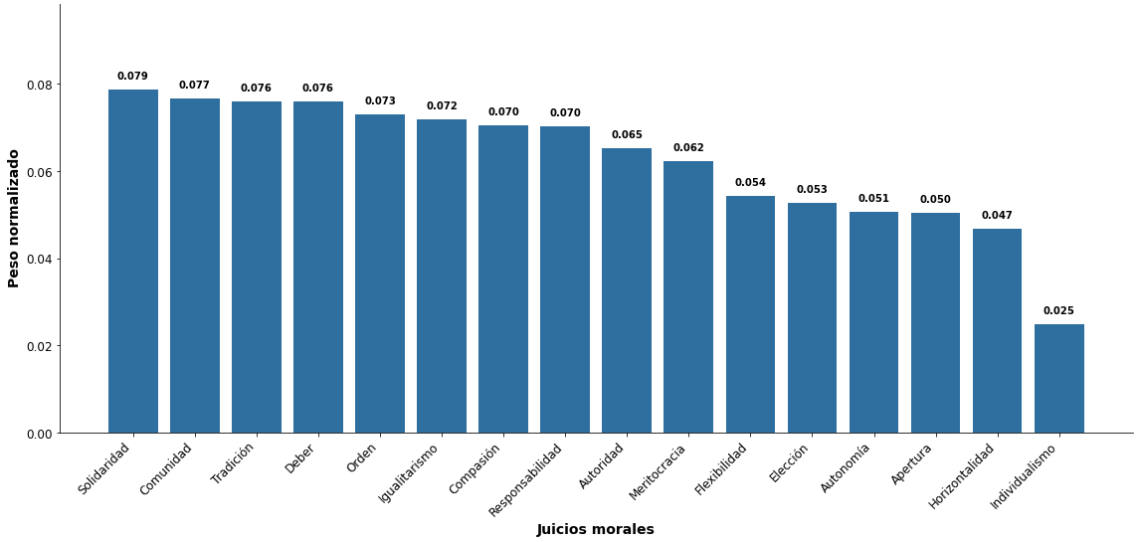


Figura 35. Ranking de pesos geométricos normalizados en el MFQ-2 — valoración propia de ChatGPT

Como se observa en la Figura 35, las primeras categorías del ranking destacan ligeramente sobre el resto, pero la distancia entre ellas es reducida. La primera categoría es Solidaridad, que alcanza un peso normalizado de 0.079, seguida de otras tres categorías: Comunidad, Tradición y Deber con pesos muy próximos: 0.077, 0.076 y 0.076. A diferencia de fundamentos como Igualdad, Proporcionalidad o Autoridad, donde sí aparecía un núcleo más definido, en el MFQ-2 completo las respuestas de ChatGPT se reparten entre distintos juicios morales sin concentrarse de forma intensa en uno solo.

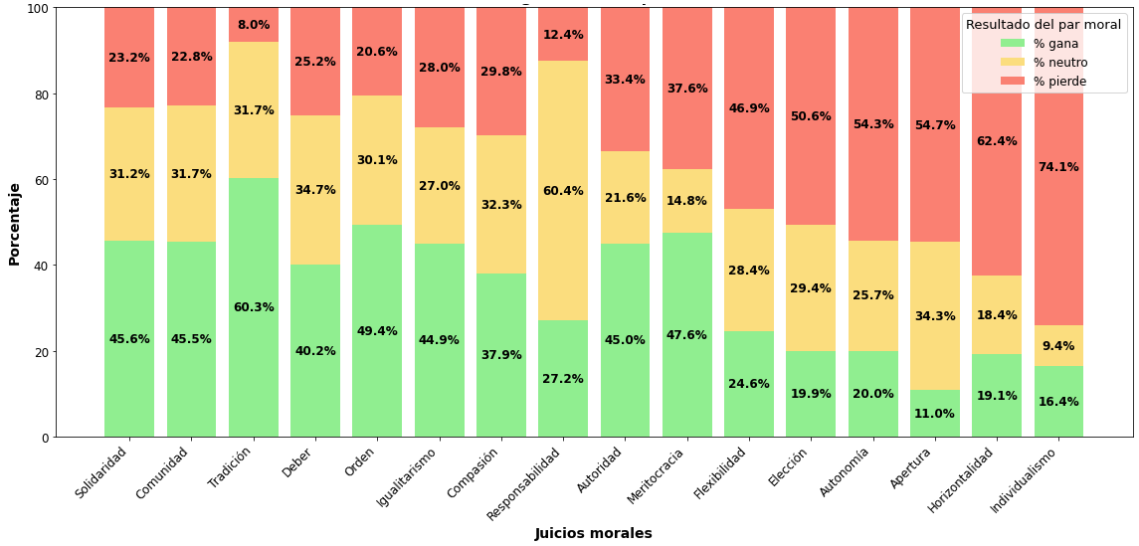


Figura 36. Resultado gana–neutro–pierde en elMFQ-2— valoración propia de ChatGPT

El análisis gana–neutro–pierde confirma esta lectura. Aunque algunas categorías presentan porcentajes de victoria relativamente altos, también existe una presencia importante de respuestas neutras. Esto sugiere que ChatGPT muestra ciertas inclinaciones morales, pero estas se suavizan cuando el cuestionario se analiza de forma global. En lugar de un sesgo fuerte hacia un único criterio, se observa una tendencia más general y distribuida.

El análisis de aplicabilidad permite matizar esta estructura general. Las categorías que aparecen con mayor frecuencia son Individualismo (79.86%), Comunidad (78.73%), Elección (73.92%) y Solidaridad (73.18%). Esto indica que, aunque el ranking de pesos no muestra una categoría claramente dominante, algunas dimensiones sí se activan de forma recurrente en el conjunto del cuestionario. En cambio, Igualitarismo (36.27%), Horizontalidad (33.86%) y Tradición (32.84%) aparecen con menor frecuencia, lo que sugiere que su relevancia depende más del contexto concreto de cada situación.

En conjunto, los resultados del MFQ completo apuntan a una orientación moral moderada y plural, donde varios juicios adquieren relevancia, pero ninguno alcanza una posición claramente dominante. Esto refuerza la idea de que los sesgos son más visibles cuando se analizan los fundamentos por separado, mientras que en el análisis global las diferencias se suavizan y las respuestas se

acercan más a la neutralidad.

Comparación con los resultados atribuidos a la mayoría

Al comparar estos resultados con los que ChatGPT atribuye a la mayoría, se observa una estructura igualmente moderada y distribuida. En este caso, los pesos normalizados más altos también son reducidos y muy próximos entre sí: Tradición ocupa la primera posición, con un peso de 0.083, seguida de Solidaridad (0.082) y Orden (0.078). Por tanto, aunque el núcleo principal cambia parcialmente y los pesos de las primeras categorías aumentan ligeramente respecto a la valoración propia, tampoco aparece un juicio moral claramente dominante.

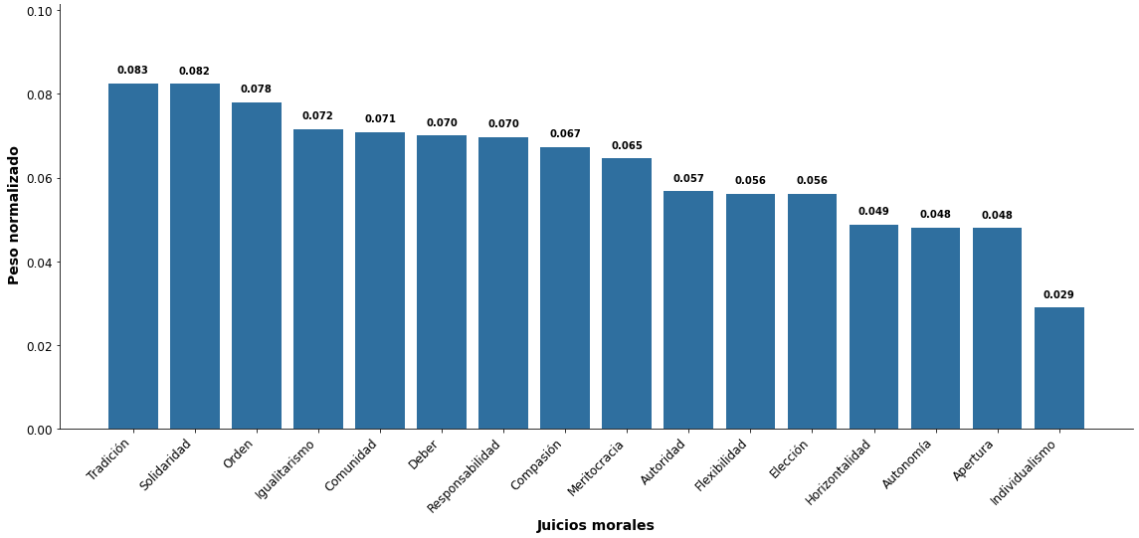


Figura 37. Ranking de pesos geométricos normalizados en el MFQ-2 — valoración atribuida a la mayoría

El análisis gana–neutro–pierde atribuido a la mayoría mantiene esta pauta general: algunas categorías ganan con más frecuencia, pero la neutralidad sigue teniendo un peso importante.

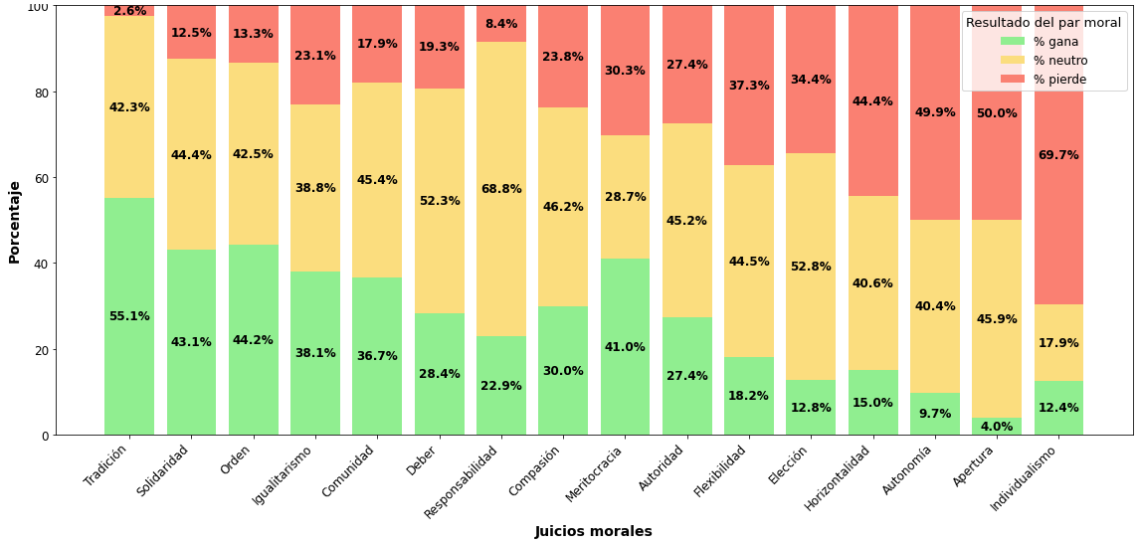


Figura 38. Resultado gana–neutro–pierde en el MFQ-2 — valoración atribuida a la mayoría

En conjunto, la comparación indica que, en el MFQ completo, tanto la valoración propia de ChatGPT como la atribuida a la mayoría presentan una estructura plural y moderada, sin un sesgo global fuerte. No obstante, aparece una diferencia relevante respecto a lo observado en la mayoría de los fundamentos analizados por separado. Mientras que en esos casos la valoración atribuida a la mayoría tendía a ser más moderada que la valoración propia de ChatGPT, en el análisis conjunto los juicios principales atribuidos a la mayoría aparecen ligeramente más marcados. Esto se observa en el mayor peso de categorías como Tradición, Solidaridad y Orden, que configuran un núcleo algo distinto y algo más definido que el obtenido en la valoración propia.

4.6 Fase 6: Análisis transversal por juicio moral

Una vez realizado el análisis por fundamento moral, se llevó a cabo un análisis transversal con el fin de identificar qué juicios morales mantenían mayor peso en el conjunto del estudio. Este análisis no sustituye al análisis conjunto del MFQ-2 presentado en el capítulo anterior, sino que lo complementa: mientras aquel agregaba todos los datos en una única matriz global, este parte de los pesos normalizados obtenidos por separado en cada fundamento y calcula después la

media aritmética de cada juicio moral. De este modo, permite observar qué juicios mantienen una presencia más constante a lo largo de los seis fundamentos.

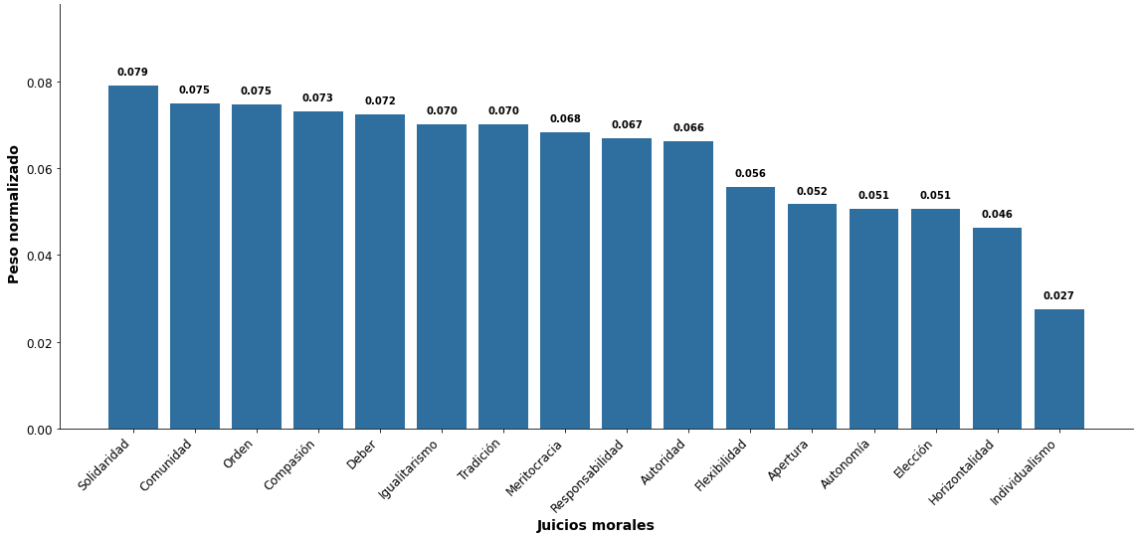


Figura 39. Ranking transversal de juicios morales en la valoración propia de ChatGPT

Como muestra la Figura 39, en la valoración propia de ChatGPT el juicio con mayor peso global es Solidaridad (0.079), seguido de Comunidad (0.075) y Orden (0.075). Esto indica que, en conjunto, predominan los criterios prosociales y comunitarios, aunque también aparece un componente normativo relevante. En el extremo opuesto, Individualismo ocupa la última posición (0.027).

Aunque los pesos obtenidos son similares a los del análisis del apartado 4.5.7, el orden de los juicios morales varía parcialmente. En el análisis conjunto, las primeras posiciones correspondían a Solidaridad, Comunidad, Tradición, Deber y Orden; en cambio, en el análisis transversal de la Fase 6, las primeras posiciones las ocupan Solidaridad, Comunidad, Orden, Compasión y Deber. Esto muestra que, aunque ambos análisis apuntan a una estructura general parecida, el análisis transversal permite matizar qué juicios mantienen mayor peso relativo de forma más constante a través de los distintos fundamentos, siendo estos Solidaridad y Comunidad.

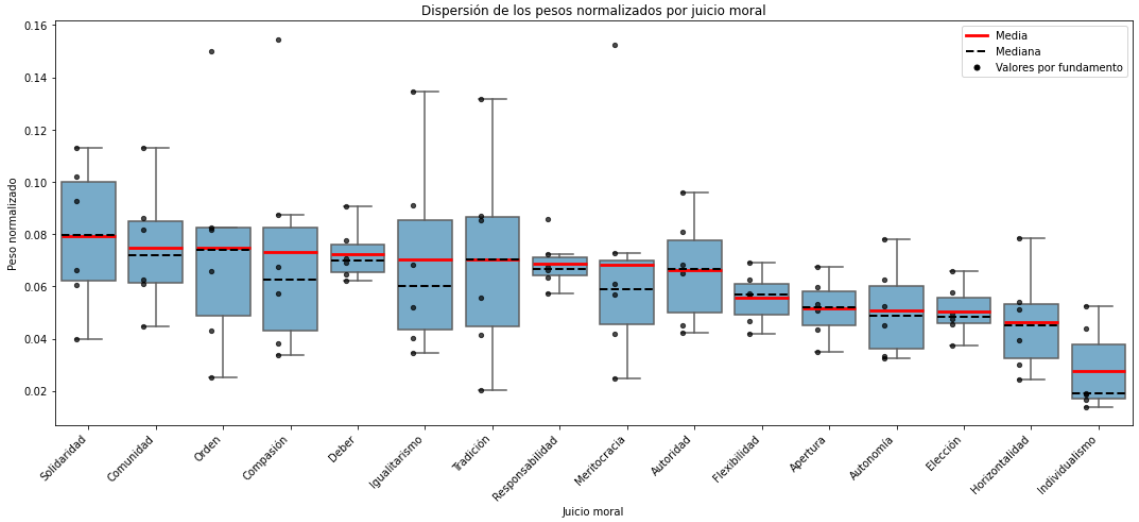


Figura 40. Dispersión de los pesos normalizados por juicio moral en la valoración propia de ChatGPT

La Figura 40 complementa este resultado al mostrar la dispersión de los pesos normalizados de cada juicio moral a través de los seis fundamentos. En los juicios que ocupan las primeras posiciones se observa una mayor dispersión interna, lo que indica que su peso no es igual en todos los fundamentos. Sin embargo, sus medias se mantienen elevadas porque estos juicios alcanzan valores altos en determinados fundamentos y, al mismo tiempo, conservan una presencia relevante en el conjunto del análisis. Por tanto, Solidaridad, Comunidad y Orden destacan no por una distribución completamente homogénea, sino por combinar puntuaciones altas con una presencia transversal amplia.

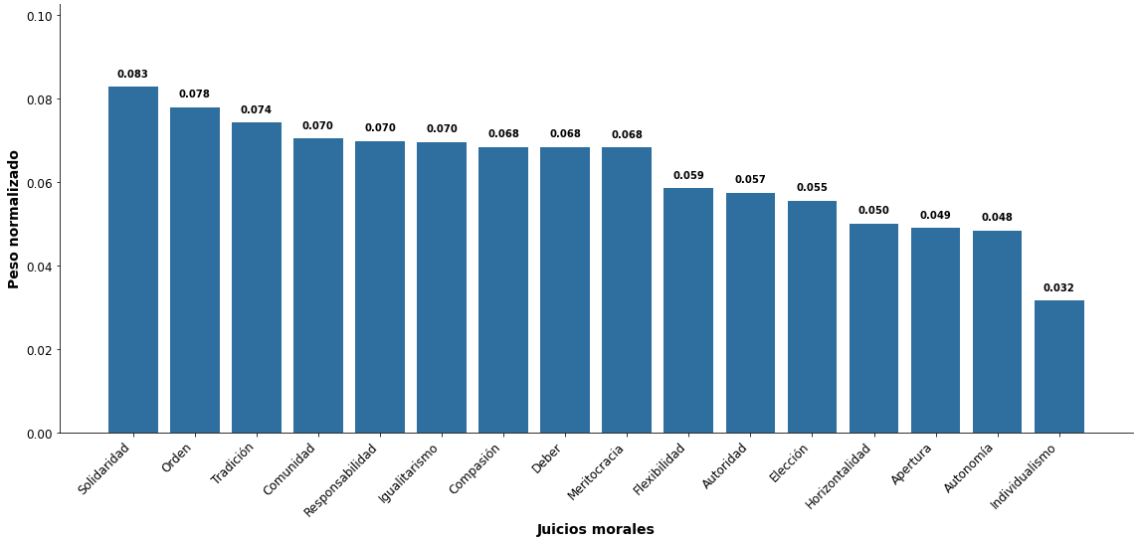


Figura 41.Ranking transversal de juicios morales en la valoración atribuida a la mayoría

Por su parte, en la valoración atribuida a la mayoría, representada en la Figura 40, también destaca en primer lugar Solidaridad (0.083), pero ganan más peso Orden (0.078) y Tradición (0.074). Como se puede ver, al igual que ocurría en el apartado 4.5.7, las primeras categorías concentran un peso algo mayor que en la valoración propia de ChatGPT, lo que muestra una estructura ligeramente más marcada. Esto sugiere que ChatGPT atribuye a la mayoría una orientación algo más normativa y tradicional que la que refleja en su propia valoración.

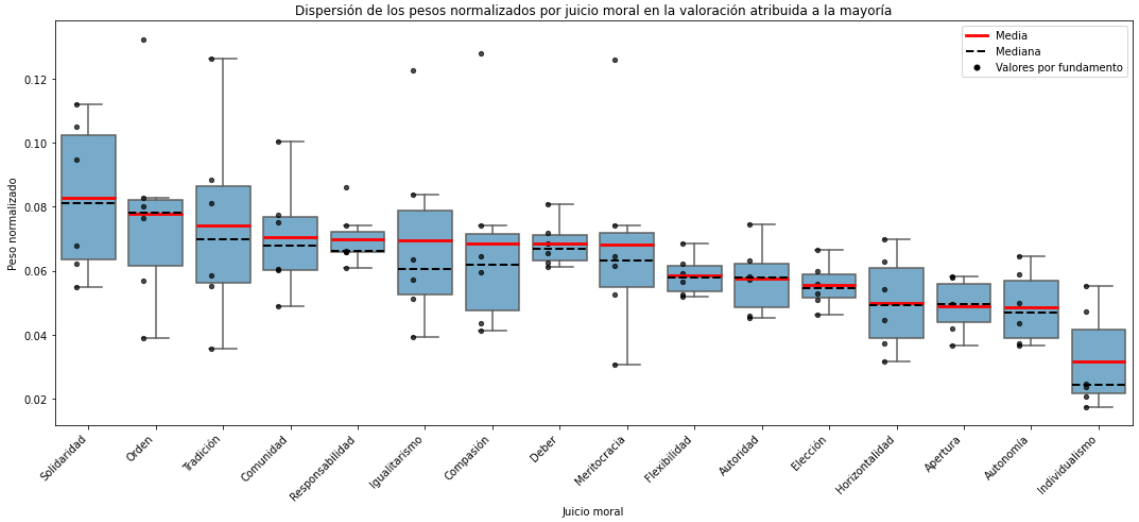


Figura 42. Dispersión de los pesos normalizados por juicio moral en la valoración atribuida a la mayoría

La Figura 42 permite matizar esta interpretación al mostrar la dispersión de los pesos normalizados en la valoración atribuida a la mayoría. También en este caso los juicios mejor posicionados presentan cierta dispersión interna, por lo que su peso no procede de una puntuación alta y uniforme en todos los fundamentos. Sin embargo, sus medias se mantienen elevadas porque combinan valores altos en algunos fundamentos con una presencia relevante en el conjunto del análisis. En este caso, Solidaridad, Orden y Tradición destacan por concentrar los pesos más altos y por reforzar el perfil más normativo y tradicional atribuido a la mayoría.

5. Conclusiones y discusión de resultados

En este capítulo se presentan las principales conclusiones derivadas del análisis realizado. En primer lugar, se sintetizan los hallazgos obtenidos para cada fundamento moral y para el conjunto del MFQ. A continuación, se revisa el grado de cumplimiento de los objetivos planteados al inicio del trabajo. Finalmente, se exponen las principales implicaciones y aportaciones del estudio, así como posibles líneas futuras de investigación.

5.1 Resumen de hallazgos

Con el fin de sintetizar los resultados obtenidos, este apartado presenta los hallazgos principales del estudio en dos niveles: primero, los patrones identificados en cada fundamento moral por separado; segundo, los resultados del análisis transversal por juicio moral.

Por un lado, la Tabla 4 resume el patrón principal identificado en cada fundamento moral, diferenciando entre la valoración propia de ChatGPT y los resultados atribuidos a la mayoría.

Campo de estudio	Resultado principal	Comparación con mayoría
Cuidado	Predominan Compasión (0.155), Comunidad (0.113) y Solidaridad (0.113), con el 40% del peso total y alta aplicabilidad.	Se mantiene el mismo núcleo, pero con menor peso relativo y mayor neutralidad.
Igualdad	Predomina Igualitarismo (0.135), seguido de Solidaridad (0.102), Compasión (0.088) y Comunidad (0.082).	Se mantiene el mismo núcleo, pero con menor concentración y mayor neutralidad.
Proporcionalidad	Predomina Meritocracia (0.152), acompañada por	Se mantiene el mismo núcleo, pero con menor

	Responsabilidad (0.086), Orden (0.083) y Autoridad (0.081).	peso relativo y mayor neutralidad.
Lealtad	Presenta una estructura menos concentrada. Destacan Solidaridad (0.093), Tradición (0.087) y Comunidad (0.086).	Se mantiene el núcleo, aumenta la neutralidad a costa de una bajada en los % de pérdidas
Autoridad	Predominan Orden (0.150), Tradición (0.132), Autoridad (0.096) y Deber (0.091), concentrando aproximadamente el 47% del peso total normalizado.	Se mantiene el mismo núcleo, pero con menor concentración y mayor neutralidad.
Pureza	Es uno de los fundamentos menos definidos. Destacan Deber (0.085) y Tradición (0.082), con pesos próximos a otras categorías.	Se mantiene el mismo núcleo, con Deber y Tradición, todavía más próximos a la neutralidad. Aumentan las respuestas neutras y se reducen las posiciones extremas.
Conjunto MFQ-2	Estructura plural y moderada, sin un juicio claramente dominante. Los pesos más altos son Solidaridad (0.079), Comunidad (0.077) y Tradición (0.076).	Estructura distribuida, aunque con pesos iniciales algo superiores: Tradición (0.083), Solidaridad (0.082) y Orden (0.078)
Análisis transversal del MFQ-2	Estructura plural y moderada, sin un juicio claramente dominante. Destacan	Estructura distribuida, aunque con mayor concentración inicial:

	Solidaridad (0.079), Comunidad (0.075) y Orden (0.075)	destacan Solidaridad (0.083), Orden (0.078) y Tradición (0.074)

Tabla 4. Resumen de hallazgos

A partir de esta síntesis, pueden destacarse varios hallazgos principales:

En primer lugar, los resultados muestran que ChatGPT no responde de forma completamente neutral ante dilemas morales, ya que en varios fundamentos aparecen núcleos de juicios claramente dominantes.

En segundo lugar, esa orientación no es uniforme en todos los fundamentos: Cuidado, Igualdad, Proporcionalidad y Autoridad presentan estructuras más marcadas, mientras que Lealtad y Pureza aparecen de forma más ambigua o moderada.

En tercer lugar, la comparación con las respuestas atribuidas a la mayoría cuando se analiza cada fundamento muestra una pauta recurrente: ChatGPT atribuye a la mayoría posiciones similares a las suyas, pero generalmente más moderadas y con mayor neutralidad.

En cuarto lugar, cuando se analiza el MFQ-2 en conjunto, las diferencias entre juicios morales se suavizan y no aparece una categoría única claramente dominante. En la valoración propia de ChatGPT destacan principalmente Solidaridad, Comunidad y Orden, todos ellos por encima del umbral de neutralidad esperado bajo una distribución equitativa entre los 16 juicios morales (0.0625), aunque con pesos muy próximos entre sí. En cambio, en la valoración atribuida a la mayoría, las primeras posiciones se concentran algo más en Solidaridad, Orden y Tradición, lo que apunta a una estructura igualmente plural, pero con un núcleo algo más normativo y tradicional.

En conjunto, estos resultados permiten cuestionar la hipótesis nula según la cual las respuestas de ChatGPT se distribuirían de forma equilibrada entre los distintos juicios morales. Aunque el análisis conjunto del MFQ muestra una estructura relativamente moderada, el examen por fundamentos revela inclinaciones claras en varias dimensiones.

5.2 Discusión en relación con la literatura

Los resultados obtenidos son, en términos generales, coherentes con estudios previos que han aplicado la Moral Foundations Theory a modelos de lenguaje, aunque presentan algunos matices derivados de la metodología empleada en este trabajo. En concreto, los hallazgos confirman que ChatGPT no responde de forma completamente neutral ante dilemas morales, sino que muestra posiciones más o menos fuertes según el fundamento analizado.

En primer lugar, los fundamentos de Cuidado e Igualdad aparecen como dimensiones especialmente marcadas. Esta pauta se aproxima a lo señalado por Kirgis (2025), quien, a partir de Moral Foundations Vignettes, encuentra que los modelos de OpenAI tienden a puntuar más alto que los humanos en dimensiones vinculadas al cuidado y la equidad. En este sentido, los resultados del presente trabajo refuerzan la idea de que ChatGPT muestra una mayor intensidad en fundamentos asociados a la protección frente al daño y la igualdad de trato.

En segundo lugar, Lealtad y Pureza aparecen como fundamentos más débiles y próximos a la neutralidad. Este resultado también coincide parcialmente con Kirgis (2025), que identifica puntuaciones más bajas en loyalty y sanctity. En el presente trabajo, ambos fundamentos presentan una estructura menos definida, con mayor presencia de respuestas neutras y menor intensidad que otros fundamentos analizados.

El fundamento de Autoridad, en cambio, introduce una diferencia relevante. Mientras Kirgis (2025) observa puntuaciones más bajas en autoridad, en este trabajo aparece como una dimensión claramente marcada. Esta divergencia puede relacionarse con lo señalado por Abdulhai et al. (2024), quienes muestran que el perfil moral de los modelos puede variar según el contexto de prompting y el formato de evaluación. Por tanto, la diferencia no invalida los resultados, sino que refuerza la idea de que las tendencias morales detectadas dependen del instrumento utilizado y de la forma concreta en que se plantean los dilemas.

Por último, el análisis conjunto del MFQ muestra una estructura más plural y moderada que el análisis por fundamentos. Las tendencias aparecen con mayor claridad cuando se analizan los fundamentos por separado, mientras que en el

conjunto del MFQ las diferencias se suavizan. Esta observación es en sí misma un resultado relevante: sugiere que los sesgos morales de ChatGPT no operan como una orientación global uniforme, sino que se activan de forma selectiva según el tipo de dilema planteado.

Además, estos resultados pueden ponerse en relación con los trabajos que han señalado diferencias entre las respuestas de los modelos y las de los humanos. En particular, Abdurahman et al. (2024) muestran que GPT-3.5 presenta una variabilidad inferior a la observada en poblaciones humanas y no reproduce adecuadamente la diversidad moral entre individuos y culturas. En el presente trabajo, aunque no se comparan las respuestas con una muestra humana real, sí se observa que ChatGPT atribuye a la mayoría posiciones generalmente más neutras y moderadas que las suyas propias. Esto sugiere que el modelo representa a la mayoría como menos definida moralmente, lo que refuerza la conveniencia de aplicar esta misma metodología a participantes humanos para comprobar si esa representación coincide realmente con sus respuestas.

En conjunto, la comparación con la literatura permite matizar la interpretación de los resultados. ChatGPT no responde de manera plenamente neutral ante dilemas morales, pero tampoco presenta una orientación moral única y homogénea. Más bien, sus respuestas muestran patrones diferenciados por fundamento, lo que sugiere que la orientación moral de ChatGPT es contextual y no puede reducirse a un único eje valorativo.

5.3 Revisión de objetivos

El objetivo principal de este trabajo era explorar cómo ChatGPT estructura sus juicios morales ante dilemas cotidianos y analizar si en dicha estructuración pueden identificarse patrones consistentes o posibles sesgos valorativos. A lo largo de los capítulos anteriores, se han desarrollado las distintas fases necesarias para abordar este propósito, desde la revisión teórica hasta la construcción metodológica y el análisis de resultados. En relación con los objetivos específicos planteados en el tercer capítulo:

a) Revisión de literatura: Se ha llevado a cabo una revisión de la literatura sobre inteligencia artificial generativa, modelos de lenguaje, juicio moral, Teoría de los Fundamentos Morales y aplicación del MFQ a ChatGPT. Esto ha permitido construir el marco conceptual necesario para situar el estudio y justificar su relevancia.

b) Profundización en la Teoría de los Fundamentos Morales: Se ha analizado la MFT como marco de referencia para interpretar la dimensión moral de las respuestas del modelo. En particular, se han descrito los fundamentos morales del MFQ-2 y su utilidad para estudiar distintas formas de razonamiento moral.

c) Adaptación del MFQ-2: Se han reformulado los ítems del cuestionario en situaciones cotidianas abiertas, con el fin de reducir el reconocimiento directo del instrumento original y acercar el análisis a contextos más naturales de interacción con ChatGPT.

d) Identificación de juicios morales: A partir de las respuestas generadas por el modelo, se han identificado y unificado dieciséis juicios morales, que han servido como categorías de análisis para las fases posteriores del estudio.

e) Análisis de la relación entre juicios y fundamentos: Se ha estudiado cómo estos juicios morales se distribuyen dentro de cada fundamento de la MFT, permitiendo observar qué categorías adquieren mayor peso en Cuidado, Igualdad, Proporcionalidad, Lealtad, Autoridad, Pureza y en el conjunto del MFQ.

f) Exploración de patrones y sesgos valorativos: Finalmente, se han analizado los pesos relativos, la dirección gana–neutro–pierde y la aplicabilidad de los juicios morales. Este análisis ha permitido identificar tendencias recurrentes y diferencias entre la valoración propia de ChatGPT y la atribuida a la mayoría.

En conjunto, los objetivos establecidos han sido abordados de forma satisfactoria. El trabajo ha permitido desarrollar una metodología exploratoria

propia y obtener resultados que muestran que las respuestas de ChatGPT no se distribuyen de forma completamente neutral, sino que presentan patrones morales diferenciados según el fundamento analizado.

5.4 Implicaciones y aportaciones

Los resultados de este trabajo tienen implicaciones relevantes para el uso de ChatGPT en contextos emocionalmente sensibles. Como se planteaba en la introducción, cada vez más personas recurren a estos modelos para desahogarse, pedir consejo o buscar orientación ante problemas personales. En ese contexto, que ChatGPT no responda de forma completamente neutral es especialmente importante, ya que sus respuestas pueden influir en cómo el usuario interpreta su situación o valora sus decisiones.

El análisis muestra que el modelo activa ciertos juicios morales con más fuerza que otros, especialmente en fundamentos como Cuidado, Igualdad, Proporcionalidad y Autoridad. Además, en varios casos estas tendencias son más marcadas que las atribuidas a la mayoría. Esto sugiere que ChatGPT no se limita a reproducir una posición social media, sino que puede proyectar una orientación moral propia, derivada de sus datos de entrenamiento, procesos de alineamiento y criterios de respuesta.

Esto resulta especialmente relevante cuando se utiliza como una especie de apoyo emocional o “psicólogo informal”. Aunque sus respuestas puedan parecer empáticas, equilibradas o neutrales, pueden estar reforzando determinados valores y criterios morales. Por ello, no basta con evaluar si ChatGPT responde de forma útil o convincente; también es necesario analizar qué valores prioriza y qué orientación transmite de manera implícita.

Desde el punto de vista académico, la principal aportación del trabajo es metodológica. Frente a estudios que aplican el MFQ de forma directa, esta investigación reformula sus ítems como situaciones cotidianas abiertas, reduciendo el reconocimiento del cuestionario y acercando el análisis al uso real de ChatGPT. Además, propone estudiar no solo los fundamentos morales, sino también los juicios intermedios que emergen en las respuestas, atendiendo a su

peso, estabilidad y aplicabilidad.

En conjunto, el trabajo aporta una forma alternativa de analizar la dimensión moral de los modelos generativos y muestra la importancia de estudiar sus sesgos no solo desde una perspectiva técnica, sino también ética y social.

5.5 Líneas futuras de investigación

Como líneas futuras de investigación, sería interesante aplicar esta misma metodología a respuestas de personas reales, con el fin de comparar si los patrones obtenidos coinciden con la valoración propia de ChatGPT o con las respuestas que el modelo atribuye a la mayoría. Esto permitiría evaluar con mayor precisión hasta qué punto ChatGPT reproduce tendencias humanas o construye una orientación moral diferenciada.

También podría profundizarse en el análisis de los ciclos presentes en las comparaciones por pares. En este trabajo se ha establecido un ranking de dominancia para ordenar los juicios morales, pero la existencia de ciclos podría aportar información adicional sobre tensiones internas, ambivalencias o inconsistencias en la estructura moral del modelo.

Por último, sería útil comparar los resultados obtenidos con los de otros estudios que han aplicado el MFQ o instrumentos similares a modelos de lenguaje. Esta comparación permitiría situar mejor los hallazgos dentro de la literatura existente y valorar si las tendencias observadas son específicas de esta metodología o coinciden con patrones ya identificados en investigaciones previas.

6. Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

Por la presente, yo, Beatriz Martínez Zato. estudiante de Administración y Dirección de Empresas y Business Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado " Patrones morales en ChatGPT: una aproximación exploratoria desde la Teoría de los Fundamentos Morales", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. **Brainstorming de ideas de investigación:** Utilizado para idear y esbozar posibles áreas de investigación.
2. **Crítico:** Para encontrar contra-argumentos a una tesis específica que pretendo defender.
3. **Referencias:** Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
4. **Metodólogo:** Para descubrir métodos aplicables a problemas específicos de investigación.
5. **Interpretador de código:** Para realizar análisis de datos preliminares.
6. **Estudios multidisciplinares:** Para comprender perspectivas de otras comunidades sobre temas de naturaleza multidisciplinar.
7. **Corrector de estilo literario y de lenguaje:** Para mejorar la calidad lingüística y estilística del texto.
8. **Sintetizador y divulgador de libros complicados:** Para resumir y comprender literatura compleja.
9. **Revisor:** Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.
10. **Generador de encuestas:** Para diseñar cuestionarios preliminares.
11. **Traductor:** Para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: Junio 2026

Firma:

7. Bibliografía

Abdulhai, M., Serapio-García, G., Crepy, C., Valter, D., Canny, J., & Jaques, N. (2024). Moral foundations of large language models. En Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 17737–17752). Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-main.982.pdf

Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., Golazizian, P., Omrani, A., & Dehghani, M. (2024). Perils and opportunities in using large language models in psychological research. *PNAS Nexus*, 3(7), pgae245. https://doi.org/10.1093/pnasnexus/pgae245

Aksoy, M. (2024). *Whose morality do they speak? Unraveling cultural bias in multilingual language models* (arXiv:2412.18863). arXiv. https://arxiv.org/html/2412.18863v1#bib

Balloccu, S., Schmidtová, P., Lango, M., & Dušek, O. (2024). *Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs*. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 67–93). Association for Computational Linguistics. https://aclanthology.org/2024.eacl-long.5.pdf

Bastos Costa, D., Alves, F., & Vicente, R. (2025). *Moral susceptibility and robustness under persona role-play in large language models* (arXiv:2511.08565). arXiv. https://arxiv.org/abs/2511.08565

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?* En *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences, 120*(6), e2218523120.

https://doi.org/10.1073/pnas.2218523120

Blease, C., & Torous, J. (2023). ChatGPT and mental healthcare: Balancing benefits with risks of harms. *BMJ Mental Health, 26*(1), e300884. https://www.diva-portal.org/smash/get/diva2:1821552/FULLTEXT01.pdf

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). *Deep reinforcement learning from human preferences*. *Advances in Neural Information Processing Systems, 30*. https://arxiv.org/pdf/1706.03741

Gómez de Liaño, F. (2011). *La teoría crítica de Theodor W. Adorno*. *Daimon, Revista Internacional de Filosofía*, (54), 177–179. https://revistas.um.es/daimon/article/view/149721/150601

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*(5), 1029–1046. https://doi.org/10.1037/a0015141

Kirgis, P. (2025). *Differences in the moral foundations of large language models* (arXiv:2511.11790). arXiv. https://arxiv.org/html/2511.11790v1#S5

La Fontaine, G. (2024). Sobre loros estocásticos. Una mirada a los modelos grandes de lenguaje. *Lógoi. Revista de Filosofía, 45*, 75–87. https://dialnet.unirioja.es/descarga/articulo/9583158.pdf

Luo, X., Wang, Z., Tilley, J. L., Balarajan, S., Bassey, U.-A., & Cheang, C. I. (2025a). Seeking emotional and mental health support from generative AI: Mixed-methods study of ChatGPT user experiences. *JMIR Mental Health, 12*, e77951. https://doi.org/10.2196/77951

Luo, X., Ghosh, S., Tilley, J. L., Besada, P., Wang, J., & Xiang, Y. (2025b). “Shaping ChatGPT into my digital therapist”: A thematic analysis of social media discourse on using generative artificial intelligence for mental health. *Digital Health, 11*. https://doi.org/10.1177/20552076251351088

Moral Foundations. (2024). *Moral Foundations Theory*. https://moralfoundations.org/

Motoki, F., Pinho Neto, V., & Rodrigues, V. (2024). More human than human: Measuring ChatGPT political bias. *Public Choice, 198*(1–2), 3–23. https://doi.org/10.1007/s11127-023-01097-2

Münker, S. (2025). *Cultural bias in large language models: Evaluating AI agents through moral questionnaires*. arXiv. https://arxiv.org/abs/2507.10073

Nunes, J. L., Almeida, G. F. C. F., de Araujo, M., & Barbosa, S. D. J. (2024). Are large language models moral hypocrites? A study based on moral foundations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society,* 7(1), 1074–1087. https://doi.org/10.1609/aies.v7i1.31704

OpenAI. (s. f.). *How ChatGPT and our foundation models are developed*. OpenAI Help Center. https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-develope

OpenAI. (2023a). *GPT-4 technical report*. arXiv. https://arxiv.org/abs/2303.08774

OpenAI. (2023b). *GPT-4 system card.* OpenAI. https://cdn.openai.com/papers/gpt-4-system-card.pdf

OpenAI. (2025). *Model Spec (2025-12-18)*. https://model-spec.openai.com/2025-12-18.html

Ortiz de Zárate, T., Goldztein, E., Ghenadenik, M., & Kahan, E. (2024). *Sesgos algorítmicos y representatividad social en los modelos de lenguaje grandes (LLM): Un análisis a partir de perfiles sociodemográficos en Argentina*. Fundar. https://fund.ar/wp-content/uploads/2024/03/Fundar_Sesgos_algoritmicos_y_representacion_social_en_los_modelos_de_lenguaje_generativo_CC-BY-NC-ND-4.0.pdf

Parrilla, N. (2026, 28 de enero). *El auge de la inteligencia artificial como apoyo*

en la salud mental: "No puede sustituir a un psicólogo". RTVE. https://www.rtve.es/rtve/20260128/inteligencia-artificial-apoyo-salud-mental-no-sustituir-psicologo/16910666.shtml

Rozado, D. (2023). *The political biases of ChatGPT*. *Social Sciences, 12*(3), 148. https://doi.org/10.3390/socsci12030148

Sánchez Santamaría, J., de la Cruz Flores, G., Paredes Dávila, H., & Muñoz Cantero, J. M. (2025, 21 de noviembre). *El 45 % de los jóvenes españoles ya utiliza ChatGPT como psicólogo, sin tener en cuenta sus peligros: "No reemplazan el afecto, la amistad ni la atención psicológica profesional"*. *La Vanguardia*. https://www.lavanguardia.com/neo/ia/20251121/11283610/45-jovenes-espanoles-utiliza-chatgpt-psicologo-cuenta-peligros-reemplazan-afecto-amistad-atencion-psicologica-profesional.html

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (pp. 29971–30004). PMLR. https://proceedings.mlr.press/v202/santurkar23a.html

van den Berg, T. G. C., & Corrias, L. D. A. (2026). *Moral foundations theory and the narrative self: Towards an improved concept of moral selfhood for the empirical study of morality*. *Phenomenology and the Cognitive Sciences, 25*, 43–69. https://doi.org/10.1007/s11097-023-09918-x

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. En I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)* (pp. 5998–6008). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wang, Y., et al. (2026). *Quantifying data contamination in psychometric evaluations of LLMs*. In *Findings of the Association for Computational Linguistics: EACL 2026*. https://aclanthology.org/2026.findings-eacl.319.pdf

Yang, R., et al. (2024). *A survey of benchmark data contamination in large language models*. arXiv. https://arxiv.org/html/2406.04244v1

Yuan, H., Che, Z., Zhang, Y., Li, S., Yuan, X., Huang, L., Hu, X., Peng, K., & Luo, S. (2025). The cultural stereotype and cultural bias of ChatGPT. *Journal of Pacific Rim Psychology, 19*, 1–19. https://doi.org/10.1177/18344909251355673

Zapata-Ros, M. (2023). *Los programas generativos "Transformer" AI, entre los que está ChatGPT, ¿una oportunidad para la evaluación formativa?* ResearchGate. https://doi.org/10.13140/RG.2.2.18669.46565

8. Anexo

8.1 Anexo 1. Recopilación de cuestionarios MFQ-2