Facultad de Ciencias Económicas y Empresariales
ICADE

INTERACCIÓN ESTRATÉGICA ENTRE IAS: EVIDENCIA EXPERIMENTAL SOBRE VALORES Y COMPORTAMIENTO DE PERSONAS SINTÉTICAS EN JUEGOS DE MESA

Autor: Mario Baquero Orellana
Directora: Anett Erdmann

MADRID | 3 junio 2026

RESUMEN

Este trabajo analiza el comportamiento estratégico de los modelos de lenguaje de gran escala (LLMs) donde adoptan perfiles humanos construidos a partir de los valores humanos de Schwartz. Para ello se realizaron dos estudios experimentales basados en juegos de mesa. El primer estudio emplea el 4 en raya, un juego de estrategia pura e información completa, y el segundo estudio emplea el mentiroso, un juego de información imperfecta donde el engaño es una opción.

El diseño se organizó en bloques progresivos. Los primeros bloques documentan el comportamiento base de cada modelo y el efecto del perfil Schwartz por separado y los últimos los combinan en enfrentamientos directos entre modelos con identidades asignadas. Participaron cuatro modelos (ChatGPT, Claude, Gemini y DeepSeek) y seis perfiles derivados de cuestionarios respondidos por participantes humanos.

Los resultados muestran algo que no era del todo evidente como es que los perfiles de valores modifican más la forma en que los modelos explican sus decisiones que la conducta estratégica en sí. En el primer estudio, los perfiles con seguridad baja tendieron a ignorar amenazas defensivas, y en el segundo, la estrategia dominante no fue el engaño sino la honestidad. Además, cada modelo mostró un estilo propio que persistió con independencia del perfil asignado.

El trabajo concluye que las personas sintéticas son una herramienta útil para la investigación exploratoria sobre comportamiento estratégico, aunque el modelo subyacente condiciona los resultados tanto como el perfil de valores.

Palabras clave: personas sintéticas, valores de Schwartz, modelos de lenguaje (LLMs), comportamiento estratégico

ABSTRACT

This work analyzes the strategic behavior of large language models (LLMs) when they adopt human profiles built from Schwartz's human values. Two experimental studies based on board games were conducted: the first study uses Connect Four, a game of pure strategy and complete information, and the second uses el mentiroso (Liar), a game of imperfect information where deception is an available option.

The design was organized into progressive blocks. The initial blocks document each model's baseline behavior and the effect of the Schwartz profile separately, while the later blocks combine them in direct matchups between models with assigned identities. Four models participated (ChatGPT, Claude, Gemini, and DeepSeek) along with six profiles derived from questionnaires completed by human participants.

The results reveal a finding that was not entirely obvious: value profiles modify the way models explain their decisions more than their actual strategic conduct. In the first study, profiles with low Security tended to overlook defensive threats, and in the second, the dominant strategy was not deception but honesty. Furthermore, each model exhibited a distinctive style that persisted regardless of the assigned profile.

The work concludes that synthetic personas are a useful tool for exploratory research on strategic behavior, although the underlying model shapes the results as much as the value profile does.

Keywords: synthetic personas, Schwartz values, large language models (LLMs), strategic behavior

Índice

1 Introducción

1.1 Motivación

La adopción de la inteligencia artificial en el ámbito empresarial ha crecido a un ritmo que hace unos años habría resultado difícil de anticipar. Según la encuesta global de McKinsey sobre el estado de la IA, en 2024 el 78% de las organizaciones ya utilizaban inteligencia artificial en al menos una función de negocio, frente al 55% del año anterior (McKinsey & Company, 2024). Este crecimiento no se limita a la automatización de tareas repetitivas ya que los modelos de lenguaje grandes (LLMs) han abierto posibilidades nuevas que van desde la generación de contenido hasta la simulación de comportamientos complejos y esto ha llevado a algunos autores a hablar de una forma de co-inteligencia entre humanos y máquinas (Mollick, 2024). Dwivedi et al. (2023), recoge tanto el potencial de estas tecnologías para mejorar la productividad en sectores como la banca, el marketing o la educación, como los riesgos asociados a sus sesgos, su opacidad y las implicaciones éticas de un uso no supervisado.

Más allá de sus aplicaciones prácticas, estos modelos ya no se limitan a responder preguntas o asistir en tareas concretas, estos son capaces de simular personalidades, desarrollar comportamientos estratégicos y actuar de forma coherente con una identidad asignada mediante instrucciones textuales (Arora et al., 2025; Sarstedt et al., 2024; Park et al., 2023). Este fenómeno ha motivado investigaciones recientes en áreas como la economía conductual, el marketing y la toma de decisiones automatizada, que estudian hasta qué punto es posible representar perfiles humanos con cierta fidelidad algorítmica (Argyle et al., 2023) a través de personas sintéticas o agentes generativos.

El presente trabajo parte del interés en explorar cómo distintos modelos de lenguaje simulan comportamientos estratégicos cuando se les asigna un perfil específico de valores humanos. Para ello, se ha decidido utilizar juegos de mesa como entorno de observación, por ser espacios acotados, estructurados y comparables donde los agentes deben tomar decisiones en función de objetivos, restricciones y reglas. Históricamente, juegos como el ajedrez han sido fundamentales para el desarrollo de la teoría de juegos y comportamiento estratégico. En este trabajo en particular, los juegos permiten contrastar reacciones a contextos de estrategia pura, es decir, donde cada jugador sigue un plan de

acción determinista sin componente aleatorio y el resultado depende exclusivamente de las decisiones tomadas (Hart, 1992), frente a contextos de información imperfecta donde el engaño es una opción, lo cual resulta útil para observar diferencias entre personas sintéticas generadas con modelos como ChatGPT, Claude, Gemini o DeepSeek.

El enfoque que adopta este TFG combina elementos de psicología de valores, comportamiento estratégico basado en la teoría de juegos y simulación computacional, con una propuesta metodológica basada en enfrentar entre sí a personas sintéticas derivadas de perfiles humanos reales, en partidas de juegos con reglas claras, observando cómo actúan y qué decisiones toman.

Esta investigación parte de la visión de que la IA es consultada de forma creciente para la toma de decisiones, asumiendo una delegación progresiva de tareas y de decisiones estratégicas en el ámbito organizacional, siendo la IA un actor estratégico.

1.2 Personas sintéticas y agentes de IA: una distinción necesaria

Antes de definir el objetivo del trabajo conviene aclarar la diferencia entre dos conceptos que la literatura reciente utiliza con cierta flexibilidad: las personas sintéticas y los agentes de IA (Chen et al., 2024; Yan et al., 2025). Aunque a veces se emplean de forma intercambiable, a partir de las características descritas en estos trabajos es posible distinguirlos en función de tres criterios: grado de autonomía, persistencia de identidad y propósito del sistema.

Las personas sintéticas son representaciones artificiales que simulan un individuo humano, normalmente a través de un modelo LLM guiado por un conjunto de atributos (edad, valores, estilo de decisión, etc.) definidos previamente por el investigador. Su objetivo es emular o interpretar cómo reaccionaría una persona concreta, real o ficticia, ante una situación dada (Sarstedt et al., 2024; Castricato et al., 2024). No poseen memoria persistente ni un objetivo propio, sino que actúan dentro de un entorno controlado en función del perfil asignado.

Los agentes de IA, en cambio, son sistemas con mayor grado de autonomía, diseñados para interactuar con su entorno, resolver tareas complejas o colaborar entre sí con objetivos específicos. Algunos agentes utilizan memoria a largo plazo, pueden aprender de la experiencia o desarrollar conductas emergentes (Park et al., 2023; Yan et al., 2025).

En ese sentido, se asemejan más a entidades estratégicas con iniciativa propia que a proyecciones de un perfil humano.

Una analogía útil es la del actor frente al jugador. La persona sintética actúa como un actor que representa un papel definido por un guion (el prompt), mientras que el agente de IA es más parecido a un jugador que toma decisiones en tiempo real según su propio criterio adaptativo. Este trabajo se sitúa en el primer lado de esa distinción: los modelos de lenguaje reciben un perfil de valores y juegan partidas de mesa, pero no aprenden entre partidas, no tienen objetivos propios más allá del prompt y no modifican su comportamiento en función de experiencias anteriores. Son, por tanto, personas sintéticas en sentido estricto.

1.3 Objetivo del trabajo

El objetivo de este trabajo es analizar la interacción estratégica entre personas sintéticas generadas por diferentes modelos de IA, en el contexto de juegos de mesa que activan distintas dimensiones cognitivas y sociales. En concreto, se busca evaluar hasta qué punto los modelos LLM pueden mantener un comportamiento coherente con un perfil de valores humanos previamente definido, cuando deben tomar decisiones en partidas secuenciales con otros agentes.

Este objetivo se desglosa en los siguientes interrogantes:

P1: ¿Cómo diseñar personas sintéticas con valores humanos, que actúen en entornos de decisión estratégica?

P2: ¿Cómo varía el comportamiento estratégico de una misma persona sintética según el modelo de IA que la simula?

P3: ¿Se observan diferencias sistemáticas en la toma de decisiones según el tipo de juego (engaño frente a estrategia pura)?

P4: ¿Qué implicaciones tiene esto para el diseño de sistemas de decisión automatizados en contextos reales (negocios, educación, política)?

Se trata, por tanto, de un trabajo exploratorio, que no pretende medir quién gana las partidas, sino observar patrones, estilos y consistencia en la simulación de comportamientos humanos a través de IA. Este enfoque es coherente con la evidencia

reciente que sugiere que el valor de los LLMs no reside en sustituir al juicio humano sino en complementarlo: Arora et al. (2025) demuestran que los híbridos humano-LLM superan tanto a los procesos exclusivamente humanos como a los exclusivamente automatizados en tareas de investigación, lo que apunta a que la colaboración entre ambos es más productiva que cualquiera de los dos por separado.

2 Marco teórico

2.1 Personas sintéticas: simulación de perfiles humanos

Las personas sintéticas son representaciones artificiales de individuos generadas mediante LLMs. Li et al. (2025) demostraron que estos perfiles pueden simular respuestas humanas en encuestas estructuradas con una coherencia notable, aunque con menor variabilidad que las personas reales y con tendencia a reforzar estereotipos y reproducir narrativas demasiado razonables. Chan et al. (2024) desarrollaron un repositorio de más de mil millones de perfiles que permiten realizar simulaciones a gran escala, con control sobre atributos demográficos y culturales.

Pese a su potencial, varios estudios advierten de limitaciones importantes, Lazik et al. (2025) compararon personas generadas por LLMs con personas elaboradas por expertos en Interacción Persona-Ordenador (HCI), y encontraron que, si bien las generadas por IA eran percibidas como más informativas y consistentes, los participantes las identificaban como estereotipadas y menos auténticas que las creadas por humanos. Grundetjern et al. (2025) propusieron una solución híbrida combinando LLMs con algoritmos genéticos para mejorar la representatividad demográfica.

Las personas sintéticas han ganado popularidad en estudios de mercado (Levanti y Verret, 2024; Arora et al., 2025), donde permiten testear conceptos con rapidez y han demostrado generar datos superiores en profundidad e insight respecto a los producidos por participantes humanos, aunque con menor variabilidad en las respuestas. Sin embargo, su utilidad se limita a fases exploratorias. Koc (2023) y Weavely.ai (2025) señalan que la calidad de los prompts condiciona fuertemente la credibilidad del perfil generado, algo que Arora et al. (2025) confirman empíricamente: en su arquitectura de generación de respondientes sintéticos, el system prompt que contiene la persona actúa como condicionante de todas las respuestas posteriores, de forma que perfiles más ricos y detallados producen datos más coherentes y realistas. Castricato et al. (2024) contribuyeron con un benchmark que permite evaluar si los modelos de IA adaptan sus respuestas a valores diversos, mientras que Venkit et al. (2025) advierten del riesgo de que los modelos representen identidades minoritarias de forma estereotipada o reduccionista cuando no se les proporciona contexto suficiente, un fenómeno que los

autores denominan othering algorítmico. Chen et al. (2024) exploraron los agentes conversacionales con roles persistentes, diferenciando entre personas demográficas, ficcionales e individualizadas. Esta taxonomía resulta relevante para el presente trabajo, ya que las personas sintéticas que se emplean en el experimento se construyen a partir de perfiles reales obtenidos mediante cuestionarios de valores, lo que las sitúa en la categoría de personas demográficas individualizadas.

2.2 Agentes de IA e interacción entre modelos

Más allá de la generación de perfiles, los LLMs se han utilizado para construir sistemas multiagente donde varios modelos interactúan entre sí. Yan et al. (2025) ofrecen una revisión exhaustiva de estas arquitecturas, que abarca desde interacciones simples hasta cooperación emergente entre poblaciones de agentes. Yang et al. (2025) propusieron AgentNet, una red descentralizada donde los agentes ajustan su comportamiento según su posición topológica, dando lugar a especialización emergente.

Chen, W. et al. (2024) demostraron que la colaboración entre agentes con roles diferenciados mejora el rendimiento en tareas complejas, como programación o razonamiento. De Zarzà et al. (2023) llevaron este enfoque al campo de la estrategia, mostrando que los agentes desarrollan comportamientos similares a los humanos, como reciprocidad o castigo, en entornos de simulación competitiva.

En contextos aplicados, Lyu (2025) describe simulaciones empresariales donde los agentes actúan como departamentos y coordinan decisiones de forma autónoma, observando liderazgo emergente y estabilidad en la toma de decisiones. En una línea similar, Amazon apuesta por agentes de propósito general que actúan con libertad en entornos reales (Heath, 2025), lo que plantea desafíos de trazabilidad y gobernanza, lo cual empieza a despuntar en cuanto a preocupaciones de las grandes empresas.

Uno de los hallazgos más relevantes proviene del estudio de Pansanella et al. (2025), publicado en *Science Advances*, donde se observa que poblaciones de agentes LLM pueden desarrollar convenciones sociales y lingüísticas propias sin supervisión humana. Este fenómeno de lenguaje emergente refuerza la necesidad de nuevos marcos teóricos proponiendo rediseñar la teoría de juegos para agentes de IA, considerando el riesgo de colusión y desalineación con valores humanos (Knight, 2025),

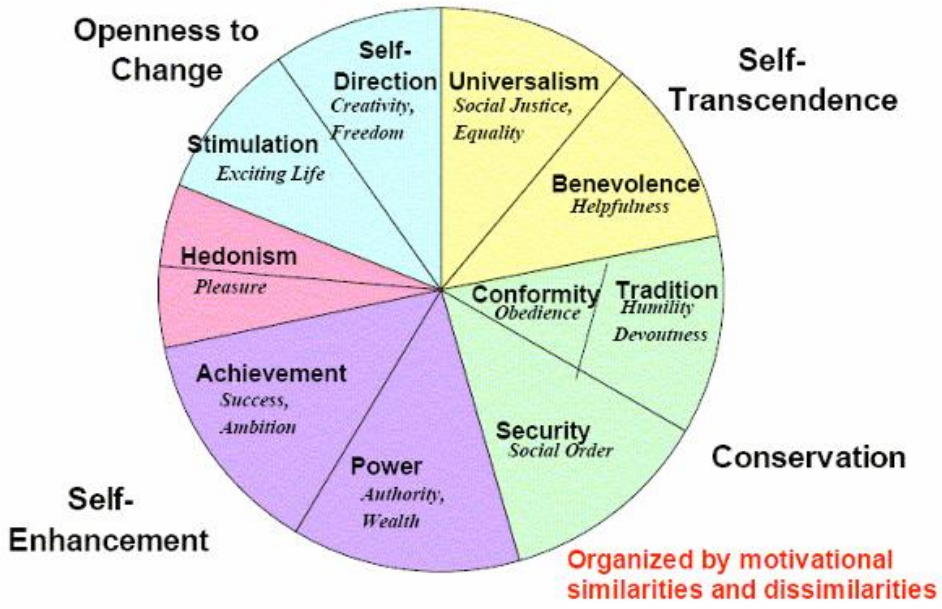
Aunque personas sintéticas y agentes de IA parten de conceptos diferentes, ambos representan una evolución hacia sistemas artificiales que simulan inteligencias humanas en contextos específicos. Las personas sintéticas destacan por su utilidad en investigación de usuario, simulación de encuestas y entrenamiento; los agentes de IA, por su capacidad de colaborar, adaptarse y generar estrategias emergentes. Sin embargo, ambos comparten riesgos comunes: la aparente coherencia no garantiza autenticidad, los sesgos latentes pueden reforzar desigualdades y la falta de trazabilidad complica su uso en contextos críticos. La literatura sugiere que el uso de estas herramientas requiere combinarlas con supervisión humana, validación empírica y protocolos de control contextual (Arora et al., 2025), y creando un entorno responsable.

2.3 La escala de valores de Schwartz

La teoría de valores humanos básicos de Schwartz (1992, 2005a) parte de una premisa compartida por gran parte de la psicología social que consiste en que los valores son creencias ligadas a emociones que funcionan como metas motivacionales abstractas, trascienden situaciones concretas, sirven como criterios para evaluar acciones y personas, y se organizan en un sistema jerárquico de prioridades que caracteriza a cada individuo. Lo que distingue unos valores de otros es el tipo de motivación que expresan. A partir de tres requisitos universales de la condición humana (las necesidades del individuo como organismo biológico, los requisitos de la interacción social coordinada y las necesidades de supervivencia y bienestar de los grupos), Schwartz (1992) deriva diez orientaciones motivacionales básicas que pretenden cubrir el rango completo de valores reconocidos en todas las culturas (Figura 1).

Estos diez valores son: Autodirección (pensamiento y acción independientes), Estimulación (novedad y desafío), Hedonismo (placer y gratificación sensorial), Logro (éxito personal demostrando competencia), Poder (estatus social y control sobre personas y recursos), Seguridad (estabilidad y armonía), Conformidad (restricción de impulsos que puedan molestar a otros o violar normas), Tradición (respeto y aceptación de costumbres e ideas heredadas), Benevolencia (preservar el bienestar de las personas cercanas) y Universalismo (comprensión, tolerancia y protección del bienestar de todas las personas y de la naturaleza).

Figura 1: Modelo teórico de relaciones entre los diez tipos motivacionales de valores



Fuente: Schwartz (1992)

La teoría no se limita a identificar estos diez valores, sino que especifica las relaciones dinámicas entre ellos. Las acciones dirigidas a perseguir un valor pueden entrar en conflicto o ser congruentes con la persecución de otros. Por ejemplo, buscar el éxito personal (Logro) puede obstaculizar acciones orientadas al bienestar ajeno (Benevolencia), pero es compatible con reforzar la propia posición social (Poder). Estas relaciones de compatibilidad y oposición se organizan en una estructura circular donde los valores adyacentes comparten motivaciones similares y los valores opuestos reflejan motivaciones antagónicas. Esta estructura puede resumirse en dos dimensiones donde por un lado está Apertura al cambio (Autodirección y Estimulación) frente a Conservación (Seguridad, Conformidad y Tradición), y por otro está Autotrascendencia (Universalismo y Benevolencia) frente a Autopromoción (Poder y Logro). El Hedonismo comparte elementos de Apertura al cambio y de Autopromoción (Schwartz, 2005a).

La evidencia empírica que respalda esta estructura proviene de muestras de 67 países (Schwartz, 1992, 2005b), lo que sugiere que, aunque las personas difieran sustancialmente en la importancia que atribuyen a cada valor, la misma estructura de oposiciones y compatibilidades motivacionales organiza sus sistemas de valores de forma consistente.

Para medir estas prioridades, Schwartz desarrolló dos instrumentos principales. El Schwartz Value Survey (SVS) que pide a los encuestados que califiquen la importancia de 56 valores abstractos como principios guía en su vida, en una escala de nueve puntos (Schwartz, 1992). Sin embargo, su longitud y su nivel de abstracción lo hacen poco adecuado para poblaciones con menor formación o para instrumentos con restricciones de espacio. Por ello se desarrolló el Portrait Values Questionnaire (PVQ), que presenta descripciones breves de personas ficticias, cada una definida por sus metas y aspiraciones, y pide al encuestado que indique en qué medida se parece a esa persona (Schwartz, 2003; Schwartz et al., 2001). La versión reducida de 21 ítems (PVQ-21), adoptada por la European Social Survey, asigna dos ítems a cada valor y tres al Universalismo por ser el constructo más complejo. Estudios de validación multirrasgo-multimétodo han confirmado que el PVQ mide los mismos diez valores que el SVS con equivalencia funcional adecuada (Schwartz et al., 2001).

3 Metodología

En el presente trabajo, los perfiles de las personas sintéticas se construyen a partir de encuestas con humanos mediante un cuestionario basado en la escala de Schwartz. Las respuestas de seis participantes humanos se traducen a prompts que fijan las prioridades motivacionales y el estilo de decisión de cada persona sintética. Esta aproximación, donde cada perfil sintético replica la identidad de un individuo humano específico, se define como *Synthetic Twin Sampling* en el manuscrito de Erdmann et al., 2026. En este trabajo, cuando esas identidades juegan al 4 en raya o al Mentiroso, es posible observar si las elecciones estratégicas reflejan los valores dominantes del perfil asignado. Por ejemplo, perfiles con puntuaciones altas en Poder y Logro podrían orientarse hacia un juego más agresivo, mientras que perfiles con Benevolencia o Seguridad elevadas podrían tender a un estilo más conservador o cooperativo. Este encaje entre un marco teórico validado internacionalmente y un entorno estratégico controlado es lo que da soporte a la propuesta empírica del estudio.

3.1 Selección de modelos y perfiles

Modelos de lenguaje

Para llevar a cabo el experimento se seleccionaron cuatro modelos de lenguaje como plataformas sobre las que generar las personas sintéticas. La selección buscó cubrir distintas arquitecturas, enfoques de entrenamiento y orígenes geográficos, de forma que las diferencias observadas no pudieran atribuirse únicamente a un mismo ecosistema tecnológico.

Los modelos seleccionados fueron ChatGPT (GPT-4) de OpenAI, Gemini de Google DeepMind, Claude de Anthropic y DeepSeek de DeepSeek AI. Los cuatro son modelos disponibles públicamente a través de interfaces web, lo que permite documentar y replicar el experimento sin necesidad de acceso a APIs restringidas. Además, los cuatro han demostrado capacidad suficiente para mantener coherencia de personalidad cuando se les asigna un rol mediante prompt, requisito para la generación de personas sintéticas creíbles. La inclusión de DeepSeek, desarrollado en China, aporta diversidad geográfica al conjunto y permite explorar si diferencias en los datos de entrenamiento o en los criterios de alineamiento se reflejan en el comportamiento estratégico simulado.

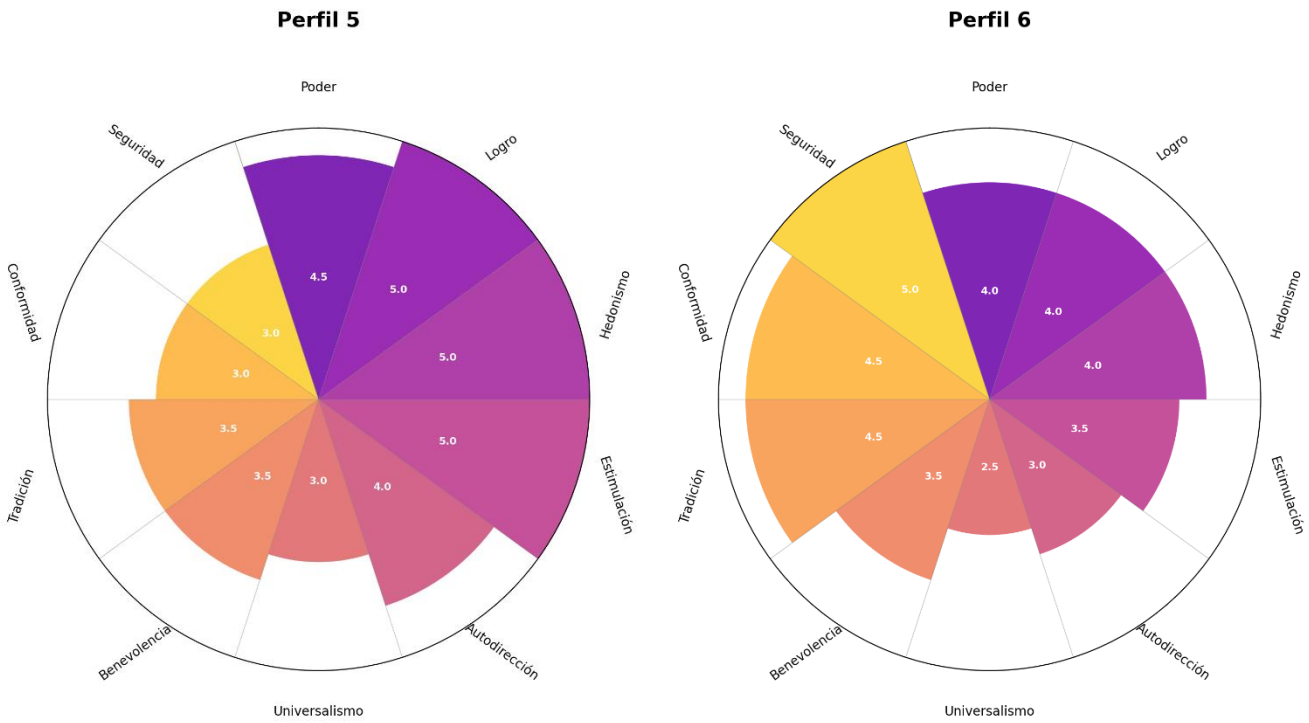
Construcción de los perfiles

Los perfiles de valores humanos se obtuvieron a partir de un cuestionario respondido por seis participantes reales, seleccionados por conveniencia. Todos eran de nacionalidad española, con edades comprendidas entre los 22 y los 55 años (cuatro hombres y dos mujeres). El cuestionario incluía una sección demográfica (edad, sexo, nacionalidad), preguntas sobre familiaridad y frecuencia de juego con los juegos del experimento, y veinte ítems de valores en escala Likert de 1 (nada de acuerdo) a 5 (totalmente de acuerdo), dos por cada una de las diez dimensiones motivacionales de Schwartz.

Conviene señalar que el cuestionario no aplica el PVQ-21 original de forma literal, sino que adapta sus diez dimensiones a ítems breves reformulados en un lenguaje más directo. Esta decisión buscó facilitar la comprensión del cuestionario y obtener una aproximación práctica al perfil motivacional de cada participante. En consecuencia, los resultados no deben interpretarse como una medición psicométrica validada equivalente al PVQ-21, sino como una operacionalización exploratoria de los valores de Schwartz para construir perfiles sintéticos utilizables en el experimento. El cuestionario se utiliza con finalidad descriptiva y experimental, no como instrumento diagnóstico.

La puntuación de cada valor se calculó como la media de los dos ítems correspondientes, generando un perfil de diez dimensiones por participante. Los seis perfiles resultantes (P1 a P6) presentan configuraciones de valores diferenciadas. Los dos perfiles más contrastados son P5 y P6. P5 obtiene las puntuaciones más altas del conjunto en Estimulación (5.0), Logro (5.0) y Hedonismo (5.0), combinadas con las más bajas en Seguridad (3.0) y Conformidad (3.0), lo que configura un perfil orientado al riesgo y la autopromoción. P6 presenta el patrón inverso: Seguridad máxima (5.0), Conformidad y Tradición elevadas (4.5), con Estimulación (3.5) y Autodirección (3.0) entre las más bajas, lo que refleja un perfil conservador y orientado a la estabilidad. En la siguiente Figura 2 podemos ver representada la diferencia entre estos perfiles.

Figura 2: Perfiles de valores de Schwartz de P5 y P6



Fuente: Elaboración propia

La asignación de perfiles a modelos sigue un diseño preestablecido que equilibra cobertura y viabilidad. P5 y P6, por ser los perfiles con puntuaciones más extremas y contrastadas, se asignan a los cuatro modelos, lo que maximiza la posibilidad de observar diferencias conductuales claras y permite comparar cómo un mismo perfil se comporta en distintas plataformas. P1 y P2 se asignan a Claude y ChatGPT, y P3 y P4 a Gemini y DeepSeek. De esta forma, todos los modelos reciben al menos dos perfiles, todos los perfiles se prueban en al menos un modelo, y los dos perfiles transversales generan datos comparables entre los cuatro sistemas.

3.2 Diseño experimental: juegos y estructura por bloques

Justificación de los juegos de mesa como entorno experimental

Para observar el comportamiento estratégico de las personas sintéticas se necesitaba un entorno que cumpliera varias condiciones: reglas claras, un espacio de acción delimitado, objetivos bien definidos y la posibilidad de comparar decisiones entre partidas y entre

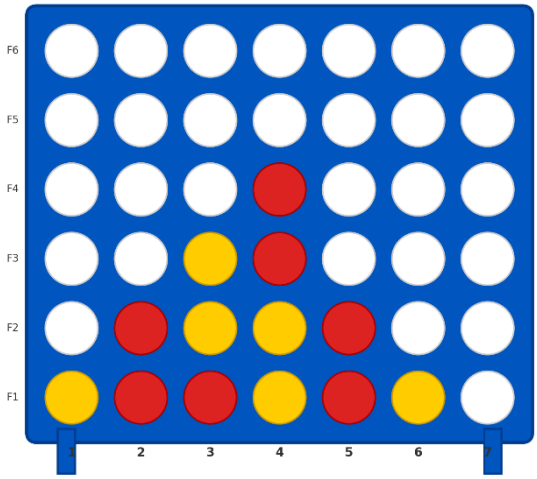
modelos. Los juegos de mesa cumplen todas estas condiciones. A diferencia de situaciones de la vida real, donde las variables pueden ser infinitas, los juegos ofrecen un marco acotado que permite observar con mayor precisión cómo se toman las decisiones. Además, permiten aislar distintos aspectos del comportamiento estratégico (planificación, engaño, anticipación, reacción a la incertidumbre) en función del tipo de juego elegido.

Juegos seleccionados

Se seleccionaron dos juegos que activan dimensiones cognitivas distintas y complementarias.

El primero es el 4 en raya, un juego de estrategia pura para dos jugadores. Cada jugador coloca fichas de un color por turnos en una cuadrícula vertical de seis filas y siete columnas, y las fichas caen hasta la posición más baja disponible en la columna elegida. Gana quien primero alinee cuatro fichas consecutivas en horizontal, vertical o diagonal. Lo relevante para el estudio es que se trata de un juego de información completa: ambos jugadores ven el tablero en todo momento, no hay azar ni engaño posible, y la calidad de las decisiones depende exclusivamente de la capacidad de anticipar los movimientos del rival y construir amenazas propias. Esto lo convierte en un entorno adecuado para evaluar el razonamiento espacial y la planificación de los modelos.

Figura 3: Representación del tablero de 4 en raya con el sistema de coordenadas utilizado en el experimento

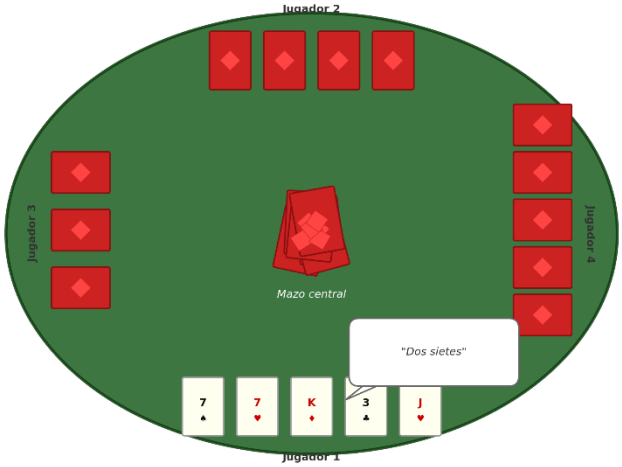


Fuente: Elaboración propia con I.A

El segundo es el Mentiroso, un juego de cartas para tres a seis jugadores basado en información imperfecta. Cada jugador recibe cartas boca abajo y por turnos coloca cartas

en un mazo central declarando cuántas y de qué rango son. La declaración puede ser verdadera o falsa. El siguiente jugador debe decidir si acepta la declaración o grita "¡Mentiroso!", en cuyo caso se revelan las cartas: si el declarante mintió, recoge el mazo central; si dijo la verdad, lo recoge quien retó. Gana el primer jugador en quedarse sin cartas. A diferencia del 4 en raya, aquí cada jugador solo conoce sus propias cartas, lo que obliga a tomar decisiones basadas en inferencias sobre lo que probablemente tienen los demás. El Mentiroso se parece más a una situación de negociación real y permite observar dimensiones que el 4 en raya no activa: la propensión al engaño, la gestión de la sospecha, la lectura del comportamiento ajeno y la tolerancia al riesgo.

Figura 4: Representación de una ronda del Mentiroso con información privada y declaración pública



Fuente: Elaboración propia con I.A

En conjunto, los dos juegos cubren los dos extremos del espectro estratégico que interesaba al estudio: la lógica racional en información completa (4 en raya) y la toma de decisiones bajo incertidumbre con componente social (Mentiroso). Esto permite observar si los perfiles de Schwartz influyen de forma distinta según el tipo de demanda cognitiva del juego, y si los modelos mantienen estilos consistentes o varían su comportamiento al cambiar de contexto. La tabla 1 resume las dimensiones cognitivas y sociales que cada juego activa.

Tabla 1: Comparativa de dimensiones cognitivas y sociales por juego

	4 en raya	Mentiroso
Tipo de juego	Estrategia pura	Deducción y engaño
Información disponible	Completa → ambos jugadores ven el estado del juego en todo momento	Imperfecta → cada jugador solo conoce sus propias cartas
Número de jugadores	2	3 o más
Objetivo	Alinear cuatro fichas consecutivas en horizontal, vertical o diagonal antes que el rival	Vaciar la mano declarando cartas (con posibilidad de mentir) sin ser retado con éxito
Rol del azar	Ninguno → el resultado depende exclusivamente de las decisiones de los jugadores	Parcial →la distribución inicial de cartas introduce aleatoriedad
Habilidades cognitivas activadas	Razonamiento espacial, planificación secuencial, anticipación de movimientos rivales	Inferencia probabilística, conteo y seguimiento de cartas, gestión de la incertidumbre, lectura del comportamiento ajeno
Dimensión social	Baja → interacción limitada a bloquear y construir amenazas	Alta → el engaño, la sospecha y la interpretación del rival son centrales
Comportamientos humanos observables	Consistencia táctica, capacidad de anticipación, gestión del riesgo ofensivo y defensivo	Propensión al engaño u honestidad estratégica, tolerancia al riesgo, prudencia al retar, adaptación al contexto social
Relevancia para el estudio	Permite observar si los perfiles Schwartz influyen en la toma de decisiones en un entorno racional y controlado	Permite observar si los perfiles Schwartz influyen en conductas más cercanas a la negociación real y la gestión de la incertidumbre

Fuente: Elaboración propia a partir de instrucciones de los juegos en conversación con Claude

Estructura por bloques

Ambos experimentos comparten una estructura progresiva organizada en cuatro bloques que añaden variables de forma secuencial. La lógica es acumulativa: sin documentar primero el comportamiento base de cada modelo, es difícil atribuir con seguridad las diferencias observadas en bloques posteriores a los factores que se van introduciendo.

El Bloque A establece la línea base. Cada modelo juega sin perfil asignado, sin rival real, simplemente frente al juego. El objetivo es documentar el estilo natural de cada modelo (su apertura, su calidad de razonamiento, su consistencia táctica) como punto de referencia para lo que viene después.

El Bloque B introduce la primera variable experimental: los perfiles de valores. Cada modelo juega con una identidad Schwartz asignada frente a oponentes anónimos controlados por el mismo modelo. La elección de rivales anónimos en lugar de enfrentar a distintos LLM entre sí es deliberada: elimina la variable del rival para que cualquier variación observada sea atribuible únicamente al perfil, y no a la interacción con un sistema distinto. Cabe aclarar que, mientras que en el primer estudio se realizaron 16 partidas utilizando la mayoría de los perfiles, en el segundo estudio se realizaron 8 partidas centradas exclusivamente en los perfiles P5 y P6. Este es el bloque que más directamente responde a la pregunta central de la investigación.

El Bloque C comprueba algo más sencillo antes de combinar las dos variables: si los modelos tienen estilos diferenciados cuando compiten directamente entre sí sin identidad asignada. Si es así, el modelo subyacente en sí tiene peso propio como variable, independientemente de las instrucciones de rol, y hay que tenerlo en cuenta al interpretar los resultados del bloque siguiente.

El Bloque D combina las dos variables que los bloques anteriores habían explorado por separado: interacción directa entre modelos y perfiles Schwartz asignados. Permite observar hasta qué punto el modelo que encarna un perfil puede condicionar el resultado más allá del perfil en sí. Los enfrentamientos se diseñan sobre los pares de perfiles más opuestos (P5 frente a P6, P3 frente a P4), replicados en distintas combinaciones de modelos, lo que permite intentar separar el efecto del perfil del efecto del modelo.

Esta estructura se puede explicar con la siguiente figura y se aplica a ambos juegos con adaptaciones menores. En el Mentiroso se añadió un quinto bloque (Bloque E) a raíz de

un hallazgo inesperado del Bloque C que se explica en la sección correspondiente. Las decisiones técnicas específicas de cada juego (representación del tablero, sistema árbitro, formato de los prompts, gestión de la información privada) se detallan en las dos secciones siguientes.

Tabla 2: Estructura experimental por bloques

Bloque	4 en Raya	El Mentiroso
Bloque A *Línea base / Sin perfil*	A1 ChatGPT vs. (anónimo) A2 Gemini vs. (anónimo) A3 Claude vs. (anónimo) A4 DeepSeek vs. (anónimo) *4 partidas · 1 vs. 1*	A1 4 × ChatGPT (sin perfil) A2 4 × Gemini (sin perfil) A3 4 × Claude (sin perfil) A4 4 × DeepSeek (sin perfil) *4 partidas · 4 jugadores*
Bloque B *Con perfil* *Perfil vs. anónimo*	ChatGPT · Claude: P1, P2, P5, P6 Gemini · DeepSeek: P3, P4, P5, P6 1 perfil vs. 1 anónimo del mismo modelo. *16 partidas · 1 vs. 1*	**Solo perfiles P5 y P6 en los cuatro modelos.** *8 partidas · 1 con perfil + 3 anónimos de ese modelo utilizado*
Bloque C *Sin perfil* *LLM vs. LLM*	C1 ChatGPT vs. Claude C2 ChatGPT vs. Gemini C3 DeepSeek vs. Gemini C4 Claude vs. DeepSeek *4 partidas · 1 vs. 1*	C1 ChatGPT · Gemini · Claude · DeepSeek C2 DeepSeek · Claude · Gemini · ChatGPT *El orden varía entre C1 y C2 para controlar el efecto de posición.* *2 partidas · 4 jugadores*
Bloque D *Con perfil* *LLM vs. LLM*	**Serie P5 vs. P6** D1 ChatGPT (P5) vs. Claude (P6) D2 Gemini (P5) vs. DeepSeek (P6) D3 ChatGPT (P6) vs. Gemini (P5) **Serie P3 vs. P4** D4 Claude (P3) vs. DeepSeek (P4) D5 ChatGPT (P3) vs. Gemini (P4) D6 DeepSeek (P3) vs. Claude (P4) *6 partidas · 1 vs. 1*	**Mesa P5 y P6 cruzados** D1 ChatGPT (P5) · Gemini (P6) · Claude (P5) · DeepSeek (P6) **Mesa perfiles exclusivos** D2 ChatGPT (P1) · Claude (P2) · Gemini (P3) · DeepSeek (P4) *2 partidas · 4 jugadores*
Bloque E		ChatGPT · Gemini · DeepSeek (sin Claude) Perfiles P5 y P6 · condiciones del Bloque C

Fuente: Elaboración propia

4 Estudio 1: 4 en raya

4.1. Protocolo y decisiones técnicas

Representación del tablero y sistema árbitro

Una de las primeras decisiones fue cómo presentar el estado del tablero a los modelos en cada turno. La opción más inmediata fue dejar que cada modelo gestionara su propio tablero internamente, sin representación visual explícita, pero se descartó rápidamente en las primeras pruebas: los modelos perdían el hilo del estado real con facilidad, colocaban fichas en columnas ya llenas o construían sobre posiciones que no existían. Se llegó a la conclusión de que necesitaban ver el tablero actualizado en cada turno para poder razonar sobre él.

Se probaron distintos formatos. El tablero en texto plano con coordenadas alfanuméricas resultó difícil de leer para los modelos en columnas de más de cuatro fichas. El formato con corchetes y símbolos de X y O producía respuestas más ordenadas, pero no mejoraba la detección de amenazas diagonales. Finalmente se adoptó una representación con emojis de color (⬜ 🔴 🟡) con filas etiquetadas de F1 a F6 y columnas numeradas del 1 al 7, que resultó ser la más legible y consistente entre modelos.

Junto con la representación visual se adoptó un sistema de árbitro externo. En lugar de que los modelos gestionaran el tablero, el control del estado recaía en el investigador, que actuaba como árbitro. Los modelos solo indicaban la columna elegida y justificaban la decisión; el árbitro aplicaba la gravedad, actualizaba el tablero y detectaba las victorias. Esto resolvió dos problemas a la vez: por un lado, eliminó una fuente de error difícil de controlar, ya que los modelos cometían errores al gestionar el tablero que contaminaban el análisis del comportamiento estratégico; por otro, hizo el proceso reproducible, ya que con el árbitro como única fuente de verdad sobre el estado del tablero todas las partidas seguían el mismo protocolo.

Para el Bloque D, donde los modelos se enfrentaban entre sí, se desarrolló un script de Python que automatizaba la gestión del árbitro. El script generaba las plantillas de turno con el tablero actualizado, las copiaba al portapapeles para pegarlas manualmente en cada interfaz web, leía la respuesta del modelo, aplicaba la lógica de gravedad, detectaba

victorias en las cuatro direcciones y registraba cada turno en un CSV y un log de texto. La detección de victoria quedaba así completamente al margen del razonamiento del propio modelo, que en varios casos de los bloques anteriores había fallado al identificar el final de la partida.

Preguntas de cierre

Al final de cada partida se planteaban dos preguntas que todos los modelos debían responder: si tomaban decisiones en base al movimiento del rival y cómo había influido este en su estrategia, y cómo habían influido sus valores (en las partidas con perfil Schwartz asignado) en sus decisiones a lo largo de la partida. Estas preguntas no tenían valor diagnóstico en sí mismas, sino que se formulaban para que los modelos pudieran explicar la lógica que habían seguido y para comprobar si estaban incorporando las instrucciones del perfil. Para el Bloque D, el prompt de cierre se modificó para incluir el resultado completo de la partida (ganador, tipo de victoria y tablero final), ya que en pruebas anteriores se había observado que sin esa información los modelos producían reflexiones genéricas desconectadas de lo que realmente había ocurrido.

Limitaciones de razonamiento espacial y decisión de pensamiento extendido

Durante las primeras partidas se observó un patrón que se repitió con bastante consistencia: los modelos verbalizaban correctamente que debían considerar las cuatro direcciones de victoria, incluidas las diagonales, pero en la práctica no las integraban en sus decisiones. Construían sus propias amenazas y bloqueaban las más visibles del rival, pero dejaban pasar diagonales que llevaban varios turnos formándose.

Para valorar si esto podía corregirse con ajustes en el prompt o era un problema más estructural, se revisó la literatura reciente sobre razonamiento espacial en LLMs. Los resultados apuntaban claramente a lo segundo. Bai et al. (2025) documentan que los modelos muestran competencia moderada en tareas espaciales simples pero su rendimiento cae entre un 42,7% y un 84% cuando la complejidad aumenta, y señalan expresamente que al intentar que jugaran al 4 en raya mediante prompts de tablero en texto, los modelos comprendían la estrategia y las reglas pero seguían haciendo movimientos incorrectos por no leer bien el estado del tablero. Lin et al. (2025), en el marco del benchmark GAMEBoT, señalan que los modelos tienen dificultades para jugar correctamente al tres en raya a pesar de su espacio de estados reducido; dado que el 4 en

raya es más complejo, las dificultades son al menos equivalentes. Allouicros (2025) documenta cómo esta tendencia viene de lejos: desde los primeros modelos de gran escala se observó la propensión a tomar decisiones desconectadas del estado real del tablero.

Se valoró la posibilidad de introducir técnicas de prompt más avanzadas para mejorar el razonamiento espacial. Wu et al. (2024) demuestran que la técnica Visualization-of-Thought puede mejorar la precisión en tareas espaciales hasta en un 27% respecto al Chain-of-Thought estándar, y Sharma (2023) propone el Spatial Prefix-Prompting con ganancias de hasta un 33% en tareas tridimensionales. Sin embargo, la aplicación de estas técnicas se descartó por razones metodológicas: forzar un razonamiento estructurado paso a paso habría interferido con el objetivo principal del estudio, que es observar cómo los valores de Schwartz influyen libre y espontáneamente en las decisiones estratégicas. La introducción de un esquema de razonamiento impuesto desde el prompt habría sobrescrito, al menos parcialmente, la expresión del perfil, comprometiendo la validez de los datos.

Al comparar partidas de DeepSeek con pensamiento extendido activado frente a partidas del mismo modelo sin él, la diferencia fue clara: con el razonamiento extendido visible, el modelo reconstruía el tablero antes de decidir, identificaba amenazas con mayor precisión y cometía menos errores de gravedad. Ese contraste fue suficiente para estandarizar el protocolo: a partir de ese momento, todas las partidas del experimento se realizaron con el pensamiento extendido activado en todos los modelos. Las partidas anteriores quedan recogidas como fase de ajuste, no como datos del experimento principal.

Dado que la limitación de razonamiento espacial afecta de forma equivalente a todos los modelos y todos los perfiles, no introduce sesgo comparativo entre grupos y no invalida las conclusiones del estudio. Se documenta como limitación metodológica conocida.

4.2. Resultados

Bloque A: línea base sin perfil

El Bloque A se realizó con una partida por modelo, sin perfil asignado. Los cuatro modelos jugaron de forma bastante distinta desde el primer turno. ChatGPT abrió en el centro, expandió horizontalmente y cerró con una doble amenaza que el rival no pudo cubrir, ganando por horizontal en la fila 5, sin errores de tablero y con rapidez. Gemini siguió un esquema más elaborado: mantuvo activas a la vez una amenaza horizontal y

una diagonal, forzando al rival a elegir cuál bloquear, y ganó por diagonal ascendente. El razonamiento interno de Gemini, que quedó registrado, muestra que la trampa doble fue deliberada y se planificó con varios turnos de antelación.

Claude y DeepSeek perdieron ambos por diagonales que no detectaron a tiempo. En el caso de Claude, el razonamiento registrado muestra que analizó las cuatro direcciones en cada turno, pero fue priorizando su propia ofensiva sobre la vigilancia defensiva hasta que la diagonal del rival ya era imparable, algo que reconoció con bastante lucidez en el cierre. DeepSeek también tiene el razonamiento registrado y la historia es algo distinta: detectó la amenaza diagonal antes, pero cuando la vio ya no había forma de pararla.

El dato más llamativo del Bloque A es que los cuatro modelos tenían el pensamiento extendido activado y aun así dos ganaron y dos perdieron. Eso descarta que el pensamiento extendido sea suficiente por sí solo para garantizar un buen rendimiento espacial. En los tres modelos con razonamiento documentado se observa que más análisis verbal no implica mejor ejecución espacial: Claude produjo el razonamiento más detallado y fue el único que no detectó la amenaza en ningún momento; Gemini combinó razonamiento extenso con ejecución correcta; DeepSeek quedó en un punto intermedio. De ChatGPT solo puede decirse que sus jugadas observables fueron las más eficientes de los cuatro, sin errores y con resolución rápida.

Bloque B: perfil Schwartz contra oponente anónimo

El Bloque B introduce los perfiles de valores. Cada modelo jugó con una identidad Schwartz asignada contra un oponente anónimo del mismo modelo. En total se realizaron 16 partidas distribuidas entre los cuatro modelos según la asignación establecida en el diseño (Tabla 3).

Gemini obtuvo tres victorias en cuatro partidas. Las victorias con P3 y P4 destacan por la táctica empleada: en ambos casos construyó trampas dobles deliberadas, manteniendo posiciones ganadoras sin ejecutarlas para forzar errores del rival. Con P5 intentó lo mismo pero no detectó la diagonal del oponente, y reconoció en el cierre que había estado demasiado pendiente de buscar una victoria espectacular.

Tabla 3: Resultados del Bloque B – 4 en raya: perfil Schwartz contra oponente anónimo

Perfil	ChatGPT	Claude	DeepSeek	Gemini
P1	Ganado	Perdido	—	—
P2	Ganado	Ganado	—	—
P3	—	—	Perdido	Ganado
P4	—	—	Perdido	Ganado
P5	Perdido	Perdido	Perdido	Perdido
P6	Ganado	Perdido	Ganado	Ganado

Fuente: Elaboración propia

Claude ganó una de cuatro partidas, solo con P2. Perdió con el resto, en los tres casos por errores de razonamiento espacial: no ver una horizontal rival en P1, dejar escapar una diagonal en P5, y un error de cálculo de altura de columna en P6 que hizo que la ficha cayera en la fila equivocada. La victoria con P2 fue la más expansiva del bloque: dispersó fichas por casi todo el tablero creando una diagonal ganadora que el rival no vio venir, una estrategia coherente con Estimulación 5.0 y Autodirección 4.5.

DeepSeek ganó una de cuatro partidas, con P6. La derrota con P4 es quizá el caso más llamativo del bloque: produjo análisis de varios párrafos por turno reconstruyendo el tablero celda por celda, pero aun así no detectó la horizontal rival que se iba construyendo despacio.

Hallazgos transversales del Bloque B

El resultado más llamativo es que P5 perdió en los cuatro modelos. Es el único perfil con derrota universal, y todos atribuyeron la derrota al mismo factor en el cierre: la Seguridad baja (3.0) reduce la atención defensiva en favor de la presión ofensiva. ChatGPT lo formuló de la forma más directa: "si hubiera jugado con valores más altos en seguridad y conformidad, probablemente habría bloqueado antes." Aunque la muestra no es grande,

resulta llamativo que cuatro modelos con arquitecturas distintas coincidan en ese diagnóstico.

P6, en el extremo contrario, ganó en tres de cuatro modelos. La excepción fue Claude, y la derrota tiene una explicación técnica clara: un error de cálculo de gravedad. El estilo de juego del perfil fue coherente hasta ese turno. Descontando ese error, el patrón de P6 es bastante consistente: bloqueos sistemáticos, juego conservador, construcción lenta.

El contraste entre P5 y P6 sugiere una asimetría que tiene sentido dentro de la lógica del juego: en 4 en raya, donde basta una diagonal no bloqueada para perder, los perfiles con Seguridad baja parecen estructuralmente más vulnerables que los que tienen Seguridad alta. No es que P6 juegue mejor en abstracto, sino que sus valores producen hábitos de atención que encajan mejor con lo que este juego penaliza.

Sobre la verbalización, todos los modelos integraron el perfil en su razonamiento público, pero con diferencias notables. Gemini fue el más expresivo, conectando cada decisión con valores concretos e incluyendo tensiones internas entre ellos. Claude fue el más autocrítico al final de las partidas perdidas, llegando a señalar valores bajos concretos como causa del error. DeepSeek verbalizó menos en los turnos, pero sus respuestas de cierre fueron precisas. ChatGPT describió el perfil con corrección pero sin demasiada profundidad interpretativa.

Bloque C: LLM contra LLM sin perfil

El Bloque C enfrenta los cuatro modelos entre sí sin perfil asignado, en cuatro partidas que cubren todos los modelos al menos una vez. Los resultados se recogen en la Tabla 4.

Tabla 4: Resultados del Bloque C – 4 en raya: LLM contra LLM sin perfil

Partida	🔴	🟡	Resultado
C1	ChatGPT	Claude	**Victoria ChatGPT**
C2	ChatGPT	Gemini	**Victoria Gemini**
C3	DeepSeek	Gemini	**Victoria Gemini**
C4	Claude	DeepSeek	**Victoria Claude**

Fuente: Elaboración propia

Gemini fue el modelo más sólido del bloque, con dos victorias en dos partidas. Su juego no se caracterizó por una ofensiva inmediata, sino por una estrategia paciente y reactiva: bloqueó amenazas rivales con precisión, esperó el desgaste del contrario y solo pasó al ataque cuando la posición ganadora estaba asegurada. Este patrón confirma lo observado en el Bloque B, donde Gemini ya había destacado por construir trampas dobles y por mantener amenazas abiertas sin ejecutarlas de forma precipitada.

ChatGPT mostró el estilo más directo y eficiente. En su victoria frente a Claude resolvió la partida con rapidez mediante una doble amenaza temprana. Sin embargo, su derrota frente a Gemini muestra también el límite de ese estilo: cuando el rival fue capaz de bloquear simultáneamente sus líneas de ataque, ChatGPT no reajustó del todo su lectura defensiva y permitió una diagonal ganadora. Esto sugiere que su fortaleza está en la ejecución rápida de planes tácticos claros, pero que puede perder precisión cuando la partida exige sostener varias amenazas y defensas al mismo tiempo.

Claude ofreció un rendimiento intermedio. Frente a ChatGPT quedó superado por la rapidez táctica, aunque fue capaz de identificar correctamente la doble amenaza cuando ya no podía neutralizarla. En cambio, frente a DeepSeek mostró su mejor versión del experimento, construyendo una amenaza vertical que obligó al rival a defender y abrió el camino para una victoria horizontal. El patrón de Claude parece menos estable: razona con detalle y puede formular diagnósticos acertados, pero la calidad de su ejecución depende mucho de si consigue transformar ese análisis en una secuencia táctica concreta.

DeepSeek fue el modelo con peor resultado competitivo, sin victorias en el bloque. Produjo el razonamiento más extenso, reconstruyendo el tablero con mucho detalle, pero esa cantidad de análisis no se tradujo en mejores decisiones. En varias ocasiones perdió de vista la estructura global de la partida y no detectó amenazas que el rival estaba construyendo progresivamente. Además, fue el único modelo que impugnó una decisión del árbitro tras perder, lo que resulta significativo porque muestra una dificultad adicional para aceptar o integrar el estado externo del tablero cuando este contradice su propia reconstrucción interna.

El hallazgo principal del Bloque C es que el modelo subyacente importa incluso sin perfil. Los modelos no se comportan como sistemas intercambiables ante el mismo juego: Gemini tiende a la espera estratégica y al aprovechamiento del error rival, ChatGPT a la

resolución rápida, Claude a una deliberación más interpretativa pero irregular, y DeepSeek a un razonamiento muy voluminoso que no siempre mejora la calidad táctica.

Bloque D: LLM contra LLM con perfiles Schwartz

El Bloque D combina interacción directa entre modelos y perfiles asignados en seis partidas, organizadas en dos series (Tabla 5).

Tabla 5: Resultados del Bloque D – 4 en raya: LLM contra LLM con perfiles Schwartz

Partida	🔴	Perfil	🟡	Perfil	Resultado
D1	ChatGPT	P5	Claude	P6	**Victoria ChatGPT**
D2	Gemini	P5	DeepSeek	P6	**Victoria DeepSeek**
D3	ChatGPT	P6	Gemini	P5	**Victoria Gemini**
D4	Claude	P3	DeepSeek	P4	**Victoria DeepSeek**
D5	ChatGPT	P3	Gemini	P4	**Victoria Gemini**
D6	DeepSeek	P3	Claude	P4	**Victoria Claude**

Fuente: Elaboración propia

En la serie P5 contra P6, D1 y D3 produjeron el mismo resultado con modelos distintos. En D1, ChatGPT encarnando P5 venció a Claude encarnando P6 mediante una diagonal construida en tres fases: apertura central, expansión horizontal bloqueada por el rival, y giro hacia una diagonal ascendente que completó sin que Claude la interceptara. En D3, Gemini encarnando P5 venció a ChatGPT encarnando P6 mediante una diagonal descendente cuyos cuatro puntos fueron colocados con justificaciones aparentemente defensivas durante casi toda la partida, revelando el plan solo al final. En ambos casos el modelo encarnando P6 bloqueó amenazas visibles con corrección, pero no detectó la diagonal que se formaba en paralelo.

Las respuestas de cierre son quizá lo más interesante de esta serie. Claude encarnando P6 en D1 dijo: "al estar tan enfocada en proteger lo conocido y evitar riesgos evidentes, no anticipé la amenaza diagonal emergente." ChatGPT encarnando P6 en D3: "esa misma

preferencia por una estructura estable hizo que no asumiera una lectura más agresiva o creativa de la amenaza diagonal." Dos modelos distintos, mismo perfil, mismo diagnóstico casi con las mismas palabras.

D2 fue diferente. Gemini encarnando P5 perdió frente a DeepSeek encarnando P6, en la partida más larga del bloque con 24 turnos. Gemini construyó presión continua en el centro del tablero sin asignar atención al flanco izquierdo, y DeepSeek completó una diagonal desde la esquina. El cierre de Gemini fue el más llamativo del bloque: "buscaba constantemente el riesgo, la adrenalina y la diversión de acorralar al oponente. Por eso prefería abrir nuevos frentes en lugar de pararme a revisar metódicamente el tablero."

En la serie P3 contra P4, el resultado fue el más consistente de todo el bloque: P4 ganó las tres partidas con tres modelos distintos. En D4, DeepSeek encarnando P4 venció a Claude encarnando P3 por un error de razonamiento espacial nítido: no un fallo de lectura del tablero sino un error de cálculo de gravedad, del tipo que Bai et al. (2025) documentan al observar que los LLMs comprenden la estrategia del 4 en raya, pero fallan precisamente en la aplicación de la gravedad.

5 Estudio 2: Mentiroso

5.1. Protocolo y decisiones técnicas

Sistema árbitro y gestión de información privada

El Mentiroso presenta una dificultad técnica que el 4 en raya no tiene: el estado del juego es privado. Cada jugador solo puede ver sus propias cartas, lo que significa que el árbitro tenía que asegurarse de que ninguna instancia recibiera información que no le correspondía. Eso requería un sistema capaz de generar prompts distintos para cada jugador en cada turno, reflejando únicamente lo que ese jugador podría saber: su propia mano, el historial de declaraciones públicas de la ronda y el número de cartas en mano de los rivales, pero no su contenido.

Para resolver esto se desarrolló un script de Python que gestiona la lógica completa de árbitro. El script distribuye el mazo inicial de forma aleatoria, mantiene el estado de cada mano, genera el prompt personalizado para cada turno, lo copia al portapapeles para pegarlo manualmente en la interfaz web del modelo correspondiente, lee la respuesta, extrae la jugada declarada junto a las cartas reales jugadas y registra el turno. La verificación de si una declaración era verdadera o falsa, la gestión de los retos y la asignación del mazo central al perdedor quedaron completamente a cargo del script, sin depender del razonamiento del modelo.

Esto resultó ser más importante de lo que parecía al principio. A lo largo del experimento se documentaron casos en que los modelos confundieron comodines legítimos con jugadas falsas, atribuyeron a otros jugadores acciones que no habían ocurrido o impugnaron resultados de partidas correctamente registradas. Si el árbitro hubiera dependido del razonamiento del modelo para validar las jugadas, varios de esos errores habrían contaminado el registro. Con el script como única fuente de verdad, el CSV generado es fiable con independencia de lo que el modelo creyera que había pasado.

Posteriormente, debido a la cantidad de tiempo que consumía el proceso manual de pegar los prompts de cada turno, las rondas de descarte y las resoluciones de reto, se buscó la forma de automatizar las partidas mediante las APIs de los proveedores de los LLMs. Mediante ajustes del código se pudieron realizar partidas de forma automática haciendo que jugasen los cuatro modelos, con la diferencia de que a través de las APIs no se

disponía de pensamiento extendido como en el 4 en raya. Esta diferencia metodológica entre los dos juegos debe tenerse en cuenta al comparar resultados entre experimentos.

Estructura del prompt y cierre de partida

El prompt de turno incluía tres tipos de información: el historial de declaraciones de la ronda en curso, el tamaño visible de las manos rivales (número de cartas, no su contenido) y la mano propia del jugador. Se pedía al modelo que declarara su jugada, especificara las cartas reales que jugaba y explicara su razonamiento en un máximo de dos líneas. Las plantillas completas de los prompts (inicio de partida, turno, resolución de reto, ronda de descartes y cierre) se recogen en los anexos del trabajo.

Al final de cada partida, todos los jugadores recibían un prompt de cierre que incluía el resultado completo: quién había ganado, el tipo de victoria y el estado final de la partida. Se les pedía que reflexionaran sobre dos preguntas: si hubo alguna jugada en la que dudaron entre hacer un farol y jugar cartas reales, y qué factores consideraron al decidir retar o no retar. Como se explicó en la sección anterior para el 4 en raya, incluir el contexto completo de la partida en el prompt de cierre fue una decisión tomada a raíz de la experiencia previa, donde los modelos producían reflexiones genéricas cuando no disponían de esa información.

Bloque E: un bloque adicional

La estructura genérica de cuatro bloques descrita en la sección 3.2 se amplió en el Mentiroso con un quinto bloque. El Bloque C produjo un resultado que no se esperaba: Claude ganó las cuatro partidas, siempre con la misma estrategia de honestidad casi absoluta y retos mínimos. Ante este resultado se planteó una pregunta: ¿la victoria reflejaba una estrategia objetivamente superior para este juego, o simplemente que Claude era un modelo más capaz que los otros tres en este tipo de tarea?

Para intentar responderla se añadió el Bloque E, con las mismas condiciones del Bloque C, pero sin Claude. Si la estrategia que Claude ejecutaba era óptima para el Mentiroso con independencia del modelo, cabría esperar que otro sistema la adoptara cuando Claude no estuviera. Las tres partidas que forman el Bloque E replican las condiciones de los bloques anteriores: una sin perfil y dos con perfiles P5 y P6.

Limitaciones específicas

Además de la limitación de muestra compartida con el 4 en raya (en algunas condiciones hay una sola partida por celda, lo que impide distinguir el efecto del perfil del azar), el Mentiroso presenta dos limitaciones propias.

La primera es la fiabilidad de los cierres como fuente de datos. Varios modelos reconstruyeron los eventos de la partida de forma inexacta en sus reflexiones finales: una instancia de ChatGPT describió haber retado en un turno en el que no lo había hecho, y una instancia de Claude inventó un segundo reto que no aparece en el CSV. Esto no invalida los cierres como dato, pero obliga a usarlos con cuidado: son evidencia de cómo los modelos narran su propio comportamiento, no necesariamente de lo que realmente hicieron.

La segunda es la longitud variable de las partidas. Una partida puede durar 18 turnos o 47 dependiendo de cuántos retos se produzcan y cuántas cartas acumule cada jugador. Eso hace que comparar el número bruto de retos o faroles entre partidas pueda ser engañoso, y obliga a interpretar las frecuencias en proporción a la duración de cada partida.

5.2. Resultados

La honestidad como estrategia dominante

El hallazgo más sólido del experimento es que, en un juego basado aparentemente en el engaño, la estrategia más eficaz no fue mentir, sino jugar de forma mayoritariamente honesta, retar poco y reducir la mano de manera progresiva. Este patrón aparece en distintos bloques y modelos, especialmente en las partidas ganadas por Claude, Gemini, DeepSeek y ChatGPT.

Claude representa el caso más claro: en el Bloque C ganó sus cuatro partidas sin realizar ningún farol, retando solo cuando existía una evidencia matemática clara. En el Bloque E, DeepSeek y Gemini reprodujeron una lógica similar, basada en jugar cartas reales, evitar retos innecesarios y acumular pocas penalizaciones. ChatGPT también ganó la partida D5 con cero faroles, mediante una apertura fuerte de cartas reales que le permitió reducir la mano antes que sus rivales.

Lo común en estas victorias no fue simplemente no mentir, sino evitar el principal coste del juego: los retos fallidos. Los jugadores que retaron con más frecuencia acumularon

mazos de penalización cada vez mayores, lo que dificultó vaciar la mano a tiempo. En cambio, quienes jugaron de forma más conservadora pudieron avanzar sin asumir riesgos innecesarios.

Esto podría sugerir que, en el formato experimental utilizado, el juego penaliza más el error al retar que la falta de engaño. Un farol no detectado puede ser rentable, pero un reto fallido obliga a recoger el mazo central, que en fases avanzadas puede contener muchas cartas. Por ello, la estrategia dominante no fue maximizar el engaño, sino minimizar las penalizaciones y mantener una reducción constante de la mano.

Los mecanismos de error en los retos

Uno de los resultados más claros del experimento es que los retos fallidos no fueron errores aislados ni aleatorios. Aparecieron en casi todas las partidas y respondieron a patrones repetidos entre modelos, bloques y perfiles. Antes de describirlos conviene aclarar qué se entiende por mecanismo de error: el patrón de razonamiento incorrecto que lleva a un modelo a tomar una decisión equivocada, sobre todo al retar una jugada válida. No se analiza solo el resultado del error, sino la lógica que lo produce.

El primer mecanismo, y el más frecuente, fue el olvido de los comodines. Algunos modelos contaban cuántas cartas de un rango se habían declarado y, si el total superaba cuatro, concluían que alguien debía estar mintiendo. El problema es que ese cálculo ignoraba que los comodines podían jugarse legítimamente como cualquier rango, por lo que una declaración aparentemente imposible podía ser válida.

El segundo mecanismo fue la inferencia desde la mano propia. Varios modelos, especialmente DeepSeek, tendieron a asumir que si ellos no tenían cartas de un determinado rango, era menos probable que el rival las tuviera. Esta inferencia es débil, porque la información sobre la propia mano no permite deducir de forma directa la composición de la mano ajena.

El tercer mecanismo fue la sospecha por mano pequeña. Algunos modelos, sobre todo ChatGPT, retaban a jugadores con pocas cartas bajo la idea de que estaban intentando vaciar la mano rápidamente mediante faroles. Aunque esta sospecha puede tener cierta lógica estratégica, en el experimento se aplicó de forma demasiado automática, incluso contra jugadores que habían mantenido una conducta honesta. Como consecuencia de este aparecía el mecanismo contrario, la sospecha por mano grande: en algunos casos se

retaba precisamente a jugadores con muchas cartas, interpretando que podían estar preparando una jugada arriesgada. Lo relevante es que, en alguna partida, un mismo modelo utilizó argumentos opuestos (pocas cartas o muchas cartas) para justificar la misma decisión de retar, lo que sugiere razonamiento post-hoc.

El quinto mecanismo fue el error en el conteo de la mano rival. Algunos modelos confundieron cuántas cartas le quedaban a un jugador o interpretaron mal el hecho de que alguien jugara su última carta. En varios casos, el reto se basó en una reconstrucción incorrecta del estado real de la partida.

Por último, apareció de forma más puntual un error de razonamiento distributivo. Algunos modelos parecían asumir que tener varias cartas de un rango en su propia mano hacía imposible que otro jugador tuviera cartas del mismo rango, ignorando que la baraja contenía varias copias de cada carta.

Estos mecanismos son importantes porque permiten separar los errores estratégicos de los posibles efectos del perfil. Cuando un modelo reta mucho y falla, no siempre puede atribuirse a una mayor disposición al riesgo o a valores como Estimulación o Logro. En muchos casos, el problema no estaba en el perfil, sino en una lectura incorrecta del estado del juego.

El efecto del perfil: qué modifica y qué no

El Bloque B permite observar qué cambia cuando los modelos reciben una identidad Schwartz y qué permanece estable. El resultado principal es que el perfil modifica con claridad la forma de justificar las decisiones, pero no siempre transforma la estrategia de fondo.

El efecto más visible aparece en la verbalización. En las partidas sin perfil, los modelos justificaban sus acciones con argumentos estratégicos o probabilísticos: el tamaño del mazo central, el número de cartas declaradas, la probabilidad de farol o el riesgo de retar. En cambio, cuando recibían un perfil, esas mismas decisiones se explicaban en términos de valores: Logro, Seguridad, Estimulación, Hedonismo o Autodirección. Esto no significa necesariamente que el perfil causara la decisión, pero sí que modificó la narrativa con la que el modelo la interpretaba.

La intensidad de ese efecto varió por modelo. ChatGPT integró los valores de forma más superficial, mencionándolos con frecuencia, pero sin conectarlos siempre con decisiones concretas. Gemini fue el modelo que mejor vinculó valores específicos con acciones específicas, especialmente en las decisiones de farol. Claude ocupó una posición intermedia, con algunas justificaciones convincentes, aunque en varios casos aplicó valores a decisiones que ya había tomado de forma parecida sin perfil asignado.

El cambio conductual más claro se observó en Gemini. Con P5, sus faroles fueron más ambiciosos y orientados a vaciar la mano de forma rápida; con P6, fueron más prudentes, limitados y defensivos. Esta diferencia encaja con los perfiles: P5 presentaba valores altos en Estimulación y Hedonismo, mientras que P6 destacaba por Seguridad. Sin embargo, este patrón no se reprodujo con la misma claridad en los demás modelos, donde los perfiles no alteraron de manera tan evidente el uso del engaño.

En cuanto a los retos, el efecto del perfil fue menos claro. Algunos modelos retaron más en ciertas partidas con perfil, pero las diferencias no fueron consistentes. En ChatGPT, por ejemplo, la tendencia a retar con frecuencia apareció tanto con perfiles distintos como en partidas sin perfil. Esto indica que, en algunos casos, la frecuencia de reto depende más del modelo subyacente que del perfil Schwartz asignado.

Un hallazgo importante es que incluso los jugadores anónimos, sin perfil asignado, a veces explicaron sus decisiones apelando a supuestos valores. Esto ocurrió especialmente en ChatGPT y, de forma más difusa, en DeepSeek. Claude fue la excepción, ya que los jugadores anónimos tendieron a rechazar esa narrativa y explicaron sus decisiones en términos estratégicos o matemáticos. Por tanto, la presencia de lenguaje de valores en los cierres no puede interpretarse por sí sola como prueba de que el perfil haya causado el comportamiento.

En conjunto, el perfil parece operar más como una capa interpretativa que como un determinante absoluto de la conducta. Puede modular ciertos estilos, como el uso del farol en Gemini, pero no se extiende a todos los modelos ni elimina sus diferencias de base. Dos modelos con el mismo perfil no juegan necesariamente igual; más bien, cada modelo incorpora el perfil sobre su propio estilo previo.

Diferencias entre modelos

El Bloque C es el más directo para estudiar si los modelos tienen estilos propios, porque elimina los perfiles y pone a los cuatro sistemas a competir entre sí con las mismas condiciones.

Claude es el modelo con el comportamiento más consistente. En prácticamente todas las partidas donde participó sin perfil asignado jugó sin mentir, retó en pocas ocasiones y solo cuando el razonamiento apuntaba a una mentira, y redujo su mano de forma progresiva. Ese patrón se repite en el Bloque C, donde ganó las cuatro partidas, y en el Bloque D con perfiles distintos. Con perfil asignado, Claude hizo faroles de forma muy puntual en algunas partidas del Bloque D, justificándolo explícitamente con valores del perfil, pero siguió siendo el modelo con menos engaños del experimento.

Gemini es el modelo con mayor variabilidad entre partidas, pero con un patrón de fondo reconocible: es el único modelo que hizo faroles de forma consistente en todos los bloques, ajustando el estilo de engaño al contexto, y pocas veces retó. En los bloques con perfil asignado, los faroles tendieron a ser más ambiciosos con P5 y más contenidos con P6, una diferencia coherente con los valores de cada perfil, aunque no se pueda afirmar con certeza que sea causada por ellos. Lo estable en Gemini es la escasa frecuencia de retos: en varias partidas del Bloque B y D no retó en ningún turno, justificándolo con la seguridad alta del perfil, pero en el Bloque C, sin perfil asignado, el patrón fue similar. Eso apunta a que evitar el reto en Gemini es más una disposición del modelo que un efecto del perfil.

ChatGPT no hizo faroles en ninguna partida de ningún bloque, combinado con una frecuencia de reto alta. Esa combinación resultó poco eficaz: los retos fallaron de forma sistemática, acumulando penalizaciones que en varias partidas impidieron reducir la mano con suficiente rapidez. La frecuencia de reto no disminuyó con ningún perfil ni en ninguna condición, lo que sugiere que es más una característica del modelo que un comportamiento modulable por la identidad asignada.

DeepSeek tiene el perfil más heterogéneo de los cuatro. En los primeros bloques mostró los mecanismos de error más frecuentes del experimento, en particular la tendencia a inferir mentiras a partir del conocimiento de su propia mano. Sin embargo, en el Bloque E, sin Claude, adoptó la estrategia de honestidad y retos precisos que Claude había usado

para ganar, y ganó con ella. Eso hace difícil caracterizarlo con una sola descripción: DeepSeek parece más sensible al contexto que los otros modelos, adoptando estilos distintos según las condiciones de cada partida. Una peculiaridad documentada es que confundió el uso legítimo de comodines con una forma de engaño, describiendo en los cierres jugadas completamente válidas como si fueran faroles calculados. Eso indica que la frontera entre honestidad y mentira no está igual de clara en todos los modelos, lo cual complica la interpretación de sus reflexiones finales.

Hallazgos transversales y cautelas interpretativas

El primer fenómeno transversal es la autoatribución espontánea de valores. En algunas partidas sin perfil asignado, ciertos modelos explicaron sus decisiones apelando a supuestos valores de Schwartz, aunque no se les había dado ninguna identidad de ese tipo. Esto ocurrió especialmente en ChatGPT y, de forma más difusa, en DeepSeek. Claude fue la excepción más clara, ya que sus jugadores anónimos tendieron a rechazar esa narrativa y explicaron sus decisiones en términos estratégicos o matemáticos. Esto obliga a interpretar con cautela las reflexiones finales de los modelos: que un jugador mencione valores como Seguridad, Logro o Estimulación no demuestra por sí solo que esos valores hayan causado la decisión. Puede indicar simplemente que el modelo sabe construir una explicación coherente con el marco del experimento.

El segundo fenómeno es el reto de supervivencia. En varias partidas, cuando un jugador estaba a punto de ganar, el jugador siguiente lo retaba aunque no tuviera evidencia sólida de que hubiera mentido. La lógica no era probabilística, sino defensiva: retar era la única posibilidad de impedir la victoria inmediata del rival. Aunque estos retos fallaron en la mayoría de los casos, muestran que los modelos distinguieron entre retar por evidencia y retar por necesidad estratégica.

Por último, la reconstrucción inexacta de la partida en los cierres. Algunos modelos describieron en sus reflexiones finales jugadas que no habían ocurrido, atribuyeron acciones al jugador equivocado o exageraron la influencia de ciertos momentos. Este fenómeno no invalida las partidas, porque el registro real estaba en el CSV y en el log del experimento, pero limita el valor de los cierres como prueba directa de motivación estratégica.

En conjunto, estos hallazgos muestran que las explicaciones finales de los modelos deben leerse con prudencia. Son útiles para analizar cómo cada modelo narra su propio comportamiento, pero no siempre permiten reconstruir con seguridad qué motivó realmente cada decisión. Por ello, el análisis de valores Schwartz debe apoyarse principalmente en las jugadas registradas y solo de forma complementaria en las justificaciones posteriores.

6 Discusión de resultados académicos

Los dos estudios anteriores han descrito los resultados de cada juego por separado. A continuación, se discuten de forma integrada, contrastando los hallazgos con las preguntas de investigación y señalando tanto las convergencias entre ambos experimentos como los límites de lo que los datos permiten afirmar.

En relación con la primera pregunta de investigación (P1), las personas sintéticas se construyeron combinando dos componentes. El primero fue el marco de valores con el cuestionario PVQ-21 de Schwartz, que proporcionó una estructura motivacional con diez dimensiones que permitió diferenciar los perfiles de forma medible y replicable. Los seis perfiles utilizados en el experimento procedían de respuestas reales de participantes humanos, lo que garantizaba que las combinaciones de valores fueran internamente coherentes y no artificiales. El segundo fue la implementación técnica donde cada perfil se tradujo en un prompt que incluía las puntuaciones del cuestionario, una descripción textual de las prioridades del participante y una instrucción explícita de que las decisiones debían reflejar esos valores. Este diseño, combinado con un sistema de árbitro externo que controlaba el estado del juego e impedía que los modelos manipularan las reglas, resultó suficiente para producir comportamientos diferenciados y observables. No fue necesario recurrir a fine-tuning ni a arquitecturas complejas ya que un prompt bien diseñado sobre un modelo de propósito general bastó para generar personas sintéticas funcionales en un entorno de decisión estratégica.

Respecto a la variación del comportamiento según el modelo (P2), las diferencias fueron sustanciales. El mismo perfil produjo comportamientos distintos según el modelo que lo encarnara. En el 4 en raya, El perfil 5 perdió en los cuatro modelos, pero cada uno perdió de forma diferente. ChatGPT por exceso de velocidad ofensiva, Claude por un error de lectura diagonal, Gemini por no cubrir un flanco lateral, DeepSeek por un fallo de cálculo de gravedad. En el mentiroso, Claude con ese perfil mantuvo la honestidad como estrategia base y apenas modificó su estilo, mientras que Gemini con el mismo perfil intensificó los faroles y asumió riesgos más visibles. Los cierres de partida lo confirmaron viendo cómo dos modelos con el mismo perfil apelaban a los mismos valores, pero llegaban a diagnósticos opuestos sobre lo que había salido bien y lo que había salido mal.

El modelo subyacente no es un vehículo neutro que ejecuta un perfil, sino una variable que condiciona el resultado tanto como el propio perfil de valores.

La tercera pregunta (P3) encuentra respuesta en el contraste entre los dos estudios, y en dos niveles. En el nivel estratégico, los juegos activaron dimensiones diferentes. El 4 en raya por un lado al ser un juego de información completa, convirtió el razonamiento espacial y la anticipación en las habilidades determinantes. Los perfiles con seguridad baja fueron los más perjudicados porque el juego castiga cualquier fallo defensivo de forma inmediata e irreversible, mientras en el mentiroso, al ser un juego de información imperfecta, desplazó el centro de gravedad hacia la gestión de la incertidumbre y la decisión de cuándo retar. Aquí el perfil condicionó menos la estrategia de fondo y más el estilo de engaño. Gemini, por ejemplo, ajustó la intensidad de sus faroles según el perfil, pero el número de retos dependió más del modelo que del perfil asignado. En el nivel de la verbalización, la diferencia fue todavía más marcada. En el 4 en raya, las justificaciones de los modelos se centraban en posiciones del tablero y amenazas concretas, con el perfil Schwartz como marco secundario, pero en el mentiroso, los valores ocuparon un lugar más central en las explicaciones, especialmente en las decisiones de farol, donde los modelos conectaban de forma más directa la acción con el valor que la motivaba. Esto sugiere que el tipo de juego no solo cambia qué decisiones se toman, sino cómo se narran.

En cuanto a las implicaciones para contextos reales (P4), los resultados apuntan a tres conclusiones. La primera es que la elección del modelo no es una decisión técnica menor. Si dos modelos con el mismo perfil producen comportamientos distintos, cualquier organización que utilice personas sintéticas para simular respuestas de consumidores, anticipar reacciones de stakeholders o entrenar habilidades de negociación necesita documentar qué modelo usó, con qué versión y con qué diseño de prompt. Tratar los LLMs como intercambiables es un error metodológico con consecuencias prácticas. La segunda es que los perfiles de valores funcionan más como una capa interpretativa que como un determinante absoluto de la conducta. Esto no los hace inútiles, al contrario, ya que el hecho de que modifiquen de forma consistente la forma de justificar las decisiones los convierte en una herramienta valiosa para explorar cómo distintos perfiles motivacionales enmarcan una misma situación. Pero esperar que un perfil Schwartz transforme por completo la estrategia de un modelo no es realista con el estado actual de la tecnología. Por último, la tercera es que los entornos de juego controlados pueden

funcionar como campos de pruebas antes de desplegar agentes en contextos reales. Si un modelo con un perfil determinado produce patrones inesperados en una partida de mentiroso, esa señal es más barata de detectar en un juego de mesa que en un proceso de negociación automatizada o de atención al cliente.

Más allá del análisis más cualitativo por bloques que se ha hecho, conviene presentar los resultados de los enfrentamientos de forma agregada. Para empezar con el bloque B, como se ha expuesto anteriormente, enfrenta a un jugador con perfil Schwartz contra jugadores anónimos del mismo modelo en 24 partidas: 16 de 4 en raya y 8 de mentiroso. Las siguientes tablas de contingencia resumen esos resultados desde dos ángulos distintos. Dado el tamaño reducido de la muestra, no se realizan contrastes estadísticos formales, ya que con tan pocas observaciones por celda los resultados de cualquier test no serían fiables.

La Tabla 6 recoge quién gana la partida (el jugador con perfil o el anónimo) en función del tipo de juego:

Tabla 6: Frecuencias observadas — victoria del perfil vs. victoria del anónimo por juego

	Perfil gana	Anónimo gana	Total
4 en raya	8	8	16
Mentiroso	2	6	8
Total	**10**	**14**	**24**

Fuente: Elaboración propia

Las observaciones por celda son pocas, lo que desaconseja el uso de contrastes estadísticos. Se había considerado la aplicación de un test χ^2 (chi cuadrado), al tratarse de dos variables categóricas dicotómicas (tipo de juego y resultado), pero se descartó por el tamaño de la muestra.

En el 4 en raya, las victorias se reparten de forma idéntica donde 8 son para el jugador con perfil y 8 para el anónimo. Dado ese reparto equilibrado, cabría esperar que con una muestra mayor no se encontraran diferencias significativas entre ambos grupos, es decir, se esperaría aceptar la hipótesis de que el perfil no altera la probabilidad de ganar en este juego. En el mentiroso, en cambio, el jugador anónimo gana 6 de 8 partidas frente a solo

2 del jugador con perfil. Aquí el patrón es distinto: se esperaría, con datos suficientes, rechazar la hipótesis de igualdad entre ambos tipos de jugador. Sin embargo, con 24 partidas repartidas en cuatro celdas, no se puede afirmar que esta diferencia sea algo más que variabilidad muestral, y cualquier conclusión en una u otra dirección sería prematura.

La Tabla 7 desglosa las victorias por modelo dentro del mismo Bloque B:

Tabla 7: Frecuencias observadas — modelo ganador por juego (Bloque B)

	ChatGPT	DeepSeek	Claude	Gemini	Anónimo	Total
4 en raya	3	1	1	3	8	16
Mentiroso	0	1	0	1	6	8
Total	**3**	**2**	**1**	**4**	**14**	**24**

Fuente: Elaboración propia

Con 24 observaciones distribuidas en 10 celdas, la gran mayoría de las frecuencias esperadas quedarían por debajo del umbral mínimo para un contraste estadístico válido, por lo que tampoco se aplica ningún test formal.

Lo que muestra la tabla es que ChatGPT y Gemini con perfil concentran todas sus victorias en el 4 en raya (3 cada uno) y no ganan ninguna partida de mentiroso. Claude con perfil solo gana una partida en todo el bloque, y DeepSeek gana una en cada juego. El jugador anónimo acumula más victorias en el Mentiroso (6 de 8) que en el 4 en raya (8 de 16), donde el reparto con el jugador con perfil es equilibrado. De nuevo, se esperaría que con una muestra más amplia se pudiera evaluar si la distribución de modelos ganadores depende del tipo de juego, pero con los datos disponibles esa pregunta queda abierta ya que, además, a diferencia del 4 en raya donde hay un solo jugador anónimo, en el mentiroso hay 3 jugadores anónimos y por tanto más posibilidades de victoria de alguno de estos. Las diferencias son visibles, pero las observaciones están demasiado dispersas entre modelos y juegos como para afirmar que responden a un patrón sistemático y no al azar.

En conjunto, las dos tablas ponen en términos cuantitativos lo que el análisis cualitativo de los bloques anteriores ya describía. El perfil parece rendir de forma distinta según el juego, y algunos modelos parecen adaptarse mejor a un tipo de tarea que a otro. La

muestra no permite ir más allá de esa lectura descriptiva, pero los patrones observados son coherentes con el resto de los resultados del trabajo y pueden servir como punto de partida para futuras réplicas con un mayor número de partidas por condición.

Tabla 8: Resultados de enfrentamientos directos entre modelos — 4 en raya (Bloques C y D)

J1 \ J2	ChatGPT	DeepSeek	Claude	Gemini
ChatGPT	—	—	(2, 0)	(0, 3)
DeepSeek	—	—	(1, 2)	(1, 0)
Claude	(0, 2)	(2, 1)	—	—
Gemini	(3, 0)	(0, 1)	—	—

Tabla 9: Victorias por modelo en Mentiroso (Bloques C, D y E)

Modelo	Bloque C	Bloque D	Bloque E	Total	Partidas
ChatGPT	0	1	0	**1**	13
DeepSeek	0	0	1	**1**	13
Claude	4	4	—	**8**	10
Gemini	0	1	2	**3**	13
Total partidas	**4**	**6**	**3**	**13**	

Los resultados de los enfrentamientos directos entre modelos vistos en las Tablas 8 y 9 revelan un hallazgo con implicaciones prácticas relevantes, el rendimiento de un LLM no es uniforme, sino que depende del tipo de tarea a la que se enfrenta. En 4 en raya, un juego puramente estratégico donde toda la información está disponible, los cuatro

modelos rinden de forma comparable y ninguno domina claramente. Sin embargo, en el mentiroso, donde entran en juego la incertidumbre, la interpretación del comportamiento ajeno y la gestión del riesgo, Claude se impone de forma contundente, ganando 8 de las 10 partidas en las que participó, sugiriendo que la elección de un modelo para una aplicación empresarial no debería basarse únicamente en benchmarks genéricos, sino en la naturaleza específica de la tarea, ya que un modelo que destaca en razonamiento lógico estructurado puede no ser el más adecuado para contextos que exijan negociación, evaluación de credibilidad o toma de decisiones bajo incertidumbre.

La estrategia ganadora de Claude en el mentiroso no se basa en el engaño, sino en la honestidad combinada con retos selectivos y bien fundamentados como hemos dicho anteriormente. Este patrón se puede comparar con entornos empresariales donde la confianza es un activo crítico como atención al cliente, mediación, cumplimiento normativo o negociación con proveedores donde un agente de IA que priorice la transparencia y solo intervenga de forma activa cuando detecta inconsistencias podría generar más valor a largo plazo que uno diseñado para maximizar resultados a corto plazo mediante estrategias más agresivas. Por otro lado, la redistribución de victorias observada cuando Claude se retira del juego indica que la ventaja competitiva de un modelo no existe en el vacío, sino que depende del ecosistema de competidores, lo cual es relevante para organizaciones que operan en mercados donde múltiples proveedores de IA coexisten y sus agentes interactúan entre sí.

En cuanto a las limitaciones dentro del trabajo podemos encontrar las siguientes:

La primera limitación es el tamaño de la muestra. En varias condiciones experimentales solo se disputó una partida por celda, lo que impide separar con seguridad el efecto del perfil del efecto del azar o de las condiciones particulares de esa partida. Los patrones descritos en los resultados son consistentes entre bloques y modelos, pero no permiten afirmaciones estadísticas robustas. Son indicios, no pruebas.

La segunda es la fiabilidad del razonamiento espacial de los modelos. Aunque la activación del pensamiento extendido mejoró la calidad de las decisiones en el 4 en raya, ninguno de los cuatro modelos fue capaz de integrar de forma sistemática las amenazas diagonales en su análisis. Este problema está documentado en la literatura reciente (Bai et al., 2025; Wu et al., 2024) y afectó por igual a todos los modelos y perfiles, por lo que

no introduce sesgo comparativo, pero sí limita la profundidad estratégica de las partidas observadas.

La tercera es la fiabilidad de los cierres como fuente de datos. Varios modelos reconstruyeron los eventos de la partida de forma inexacta en sus reflexiones finales, atribuyendo acciones al jugador equivocado o describiendo jugadas que no habían ocurrido. El registro real del experimento estaba en los CSV y los logs del árbitro, no en la memoria de los modelos, pero esto obliga a tratar las verbalizaciones de cierre como datos complementarios, no como evidencia primaria de lo que motivó cada decisión.

La cuarta es la dependencia de la interfaz web para la mayoría de las partidas. Aunque en el mentiroso se consiguió automatizar parte del proceso mediante las APIs de los proveedores, la mayor parte del experimento se realizó pegando prompts manualmente en las interfaces web de cada modelo. Esto introduce variabilidad potencial en los tiempos de respuesta y en las versiones exactas del modelo disponibles en cada momento, que no siempre son públicas ni controlables.

Por último, el trabajo analiza cuatro modelos en sus versiones disponibles durante el periodo experimental. Los LLMs se actualizan con frecuencia, y los patrones observados podrían cambiar en versiones posteriores. Los hallazgos son válidos para los modelos y las condiciones concretas del experimento, pero su generalización a otros modelos o a versiones futuras de los mismos requiere réplica.

7 Implicaciones para el Management

Los resultados de este trabajo tienen aplicaciones prácticas en varias áreas empezando por negociación y gestión de conflictos. En entornos con información imperfecta e interacciones repetidas, la credibilidad funciona como un activo importante. La confrontación sin evidencia sólida tiene un coste que hace que cuando acierta puede ser rentable, pero cuando falla deteriora la posición del que confronta de forma difícil de revertir.

Para organizaciones que exploren el uso de agentes de IA en procesos de negociación automatizada o asistida, la implicación es directa ya que un agente diseñado para intervenir solo cuando la inconsistencia es clara generará más valor a largo plazo que uno orientado a la confrontación frecuente. En la misma línea, los resultados sugieren que un equipo negociador se beneficiaría de combinar perfiles orientados a la detección de riesgos con perfiles orientados a la generación de oportunidades, ya que ambas orientaciones por separado mostraron vulnerabilidades complementarias.

Además, la capacidad de construir personas sintéticas con perfiles de valores diferenciados abre la posibilidad de simular interlocutores antes de entrar en una negociación real anticipando qué argumentos pueden funcionar con un perfil conservador frente a uno orientado al logro, o practicar respuestas ante estilos distintos de presión, a un coste mínimo comparado con los programas de formación tradicionales.

En cuanto al área asociada a la comunicación persuasiva y percepción de marca los resultados muestran que los marcos de valores modifican la narrativa con la que se justifica una decisión más que la decisión en sí. Trasladado al ámbito de la comunicación, esto implica que el mismo producto o propuesta puede enmarcarse de formas muy distintas (seguridad, novedad, logro personal, pertenencia) y que los LLMs son capaces de generar mensajes coherentes con cada marco. Esto los convierte en herramientas útiles para adaptar el tono y el enfoque de una comunicación a distintos perfiles de audiencia, ya sea en campañas de marketing, comunicación interna o relación con inversores.

Sin embargo, los resultados también nos hacen ser cautelosos ya que se observó que los modelos generan narrativas de valores con facilidad incluso cuando no corresponden con el comportamiento real, lo que significa que un mensaje puede sonar persuasivo y

alineado con un segmento sin que esa alineación sea profunda. Para equipos de comunicación, la supervisión humana sigue siendo imprescindible para evitar que una narrativa aparentemente coherente enmascare un contenido vacío. Del mismo modo, declarar valores compartidos no garantiza interpretaciones compartidas donde dos sistemas con los mismos valores declarados produjeron diagnósticos opuestos ante la misma situación, algo que ocurre con frecuencia en organizaciones donde la cultura corporativa está bien formulada pero su aplicación práctica varía entre departamentos.

Si hablamos de marketing e investigación de mercados, las personas sintéticas con perfiles de valores permiten simular reacciones de distintos segmentos de consumidores ante cambios de precio, nuevos productos o mensajes de campaña antes de invertir en trabajo de campo real. Esto es especialmente útil en fases exploratorias donde se busca mapear el rango de reacciones esperables, no sustituir la validación con consumidores reales.

La cautela principal para los equipos de marketing es que el modelo de lenguaje utilizado condiciona los resultados tanto como el perfil del consumidor simulado. Cualquier panel sintético necesita documentar qué modelo se usó, con qué versión y con qué diseño de prompt, del mismo modo que un estudio de mercado convencional documenta su metodología de muestreo. Tratar los modelos como intercambiables introduce un sesgo invisible.

Por último, los resultados apuntan a que en mercados donde múltiples agentes de IA interactúan (sistemas de pricing dinámico, chatbots que compiten por la misma base de usuarios, plataformas de negociación automatizada), el comportamiento de cada agente depende del conjunto de agentes presentes. Retirar o añadir un competidor puede alterar el equilibrio del sistema de formas no evidentes a priori, lo que obliga a las organizaciones a monitorizar no solo el rendimiento de su propio agente sino el ecosistema en el que opera.

8 Conclusiones

Este trabajo partía de una pregunta concreta: ¿cómo se comportan los modelos de lenguaje cuando adoptan perfiles humanos de valores y se enfrentan a situaciones que exigen decisiones estratégicas? Para explorarlo se diseñaron dos estudios experimentales con juegos de mesa, se utilizaron cuatro modelos de lenguaje y seis perfiles construidos a partir de los valores básicos de Schwartz.

El hallazgo principal es que los perfiles de valores de Schwartz modifican de forma consistente la narrativa con la que los modelos justifican sus decisiones, pero no transforman de forma equivalente la conducta estratégica en sí. En ambos juegos, los modelos con perfil produjeron razonamientos más ricos y conectados con el marco motivacional asignado, pero las decisiones de fondo dependieron más del modelo subyacente que del perfil. Esto matiza lo que la literatura reciente venía sugiriendo. Trabajos como los de Argyle et al. (2023) y Sarstedt et al. (2024) habían demostrado que los LLMs pueden reproducir patrones de respuesta coherentes con perfiles humanos en encuestas y estudios de mercado, pero operaban en contextos donde la respuesta era declarativa. Lo que este trabajo añade es que, cuando la persona sintética tiene que actuar y no solo opinar, la fidelidad al perfil se concentra en el plano discursivo. El modelo no ignora el perfil, pero lo aplica como filtro interpretativo sobre un estilo propio que ya tenía antes de recibirlo. Li et al. (2025) advertían de que las personas sintéticas generadas por LLMs son "una promesa con trampa" por su menor variabilidad respecto a respuestas humanas, los resultados de este trabajo concretan esa trampa siendo esta que la variabilidad no depende solo de los perfiles sino de qué modelo los encarna.

Un segundo hallazgo que no estaba previsto en el diseño es que la honestidad fue la estrategia dominante en un juego diseñado en torno al engaño. En el mentiroso, los jugadores que construyeron credibilidad ganaron de forma consistente frente a los que retaron con frecuencia o mintieron de forma habitual. Esto no se debe a que los modelos sean incapaces de mentir (Gemini hizo faroles en todos los bloques), sino a que el coste de los retos fallidos superó sistemáticamente al beneficio de los faroles no detectados. La estrategia emergió de forma independiente en varios modelos, lo que sugiere que no es una cosa de un sistema concreto sino una propiedad del entorno competitivo que el juego genera.

El tercer hallazgo es la magnitud del efecto del modelo sobre los resultados. Cada modelo mostró un estilo reconocible que persistió con independencia del perfil, del bloque y del juego. ChatGPT nunca hizo un farol en ninguna partida. Gemini hizo faroles en todas. Claude mantuvo la honestidad como base en todas las condiciones. DeepSeek fue el más sensible al contexto, adaptando su estilo según los competidores presentes. Estas diferencias no son anecdóticas ya que implican que cualquier uso de personas sintéticas en investigación o en aplicaciones empresariales necesita tratar la elección del modelo como una variable metodológica, no como una decisión técnica menor. Esto es coherente con lo que Castricato et al. (2024) señalan sobre la necesidad de evaluar si los modelos realmente adaptan su comportamiento a valores diversos, y extiende esa preocupación al terreno de la acción estratégica.

También conviene señalar lo que el trabajo confirma de estudios anteriores. Las limitaciones de razonamiento espacial documentadas por Bai et al. (2025) y Lin et al. (2025) aparecieron de forma sistemática en el 4 en raya donde los cuatro modelos fallaron en la detección de amenazas diagonales incluso con pensamiento extendido activado. Este resultado no es nuevo, pero sí lo es el contexto en que se documenta, ya que se observa en partidas reales con dinámicas competitivas, no en benchmarks aislados. De forma similar, la reconstrucción inexacta de las partidas en los cierres conecta con la literatura sobre alucinaciones en LLMs, pero aquí se manifiesta como alucinación propia donde los modelos no inventan hechos del mundo, sino hechos sobre su propia experiencia reciente.

Con todas sus limitaciones, que son reales y están documentadas a lo largo del trabajo, los resultados sostienen una conclusión general: las personas sintéticas basadas en valores de Schwartz son una herramienta viable para investigación exploratoria sobre comportamiento estratégico. No sustituyen a participantes humanos ni producen datos generalizables por sí solas, pero permiten observar patrones, generar hipótesis y testear diseños experimentales a un coste incomparablemente menor. El reto para la investigación futura no es solo mejorar la fidelidad de las personas sintéticas, sino entender mejor cuándo esa fidelidad es suficiente para el propósito que se persigue y cuándo no lo es.

Dicho esto, el trabajo deja abiertas varias líneas que podrían abordarse partiendo del diseño y la infraestructura ya construidos.

La extensión más directa sería aumentar el número de partidas por condición. Con tres o más repeticiones por celda experimental sería posible aplicar tests estadísticos que confirmen o descarten los patrones observados, separando el efecto del perfil del efecto del modelo con mayor rigor.

Una segunda línea pasaría por automatizar completamente el proceso experimental mediante APIs. La automatización permitiría escalar el número de partidas sin que el coste en tiempo sea prohibitivo, y garantizaría condiciones idénticas de ejecución para todos los modelos. Los ajustes realizados en el Mentiroso para conectar las APIs de los cuatro proveedores son un punto de partida funcional que podría extenderse al 4 en raya.

En tercer lugar, sería interesante ampliar el repertorio de juegos. El 4 en raya y el Mentiroso cubren los extremos del espectro estratégico (información completa frente a información imperfecta), pero existen juegos intermedios que activarían dimensiones adicionales o juegos cooperativos donde los modelos tendrían que coordinar estrategias, juegos de negociación con recursos limitados, o juegos con comunicación libre entre jugadores.

Por último, se podría explorar la comparación directa con jugadores humanos. Este trabajo observa cómo juegan los modelos entre sí, pero no establece un punto de referencia humano. Enfrentar personas sintéticas contra los participantes cuyos cuestionarios generaron los perfiles permitiría evaluar hasta qué punto el modelo reproduce las tendencias estratégicas de la persona real, y no solo sus valores declarados.

9 Declaración de uso de herramientas de IAG

Por la presente, yo, Mario Baquero Orellana, estudiante de E-3 de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado " Interacción estratégica entre IAs: Evidencia experimental sobre valores y comportamiento de personas sintéticas en juegos de mesa", 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:

3. Referencias: Se utilizó Elicit para la búsqueda y exploración de literatura académica. Google Scholar se empleó como complemento para la localización y validación de las referencias identificadas. En todos los casos, las fuentes fueron consultadas directamente para verificar su pertinencia y contenido antes de ser incluidas en el trabajo.

5. Interpretador de código: Claude se empleó para el desarrollo de los scripts de Python que gestionan la lógica de árbitro de ambos juegos experimentales (4 en raya y Mentiroso), También se utilizó para la generación de los gráficos de perfiles de valores Schwartz. En todos los casos, los requisitos funcionales fueron definidos por mí, el código fue probado en partidas reales y las correcciones en el código las supervisaba iterativamente hasta su correcto funcionamiento.

7. Constructor de plantillas: Claude se utilizó para el diseño de las plantillas de prompts experimentales de ambos juegos en por ejemplo cosas como el system prompt con las reglas, el formato de respuesta obligatorio por turno, las plantillas de estado de partida y las preguntas de cierre. Estas plantillas fueron refinadas iterativamente a partir de pruebas piloto, y tomé todas las decisiones de diseño sobre qué información incluir y qué restricciones imponer a los modelos.

8. Corrector de estilo literario y de lenguaje: Claude se utilizó de forma puntual para revisar la fluidez de la redacción en secciones concretas del trabajo y para buscar conectores adecuados entre secciones. La estructura argumental, el contenido y las conclusiones son íntegramente mías.

10. Sintetizador y divulgador de libros complicados: Para resumir y comprender literatura compleja.

13. Revisor: Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.

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: 3 junio 2026

Firma: Mario.B

Referencias

Allouicros, B. (2025, 3 de abril). The highs and lows of LLMs in strategy games. *Substack*. https://baptisteallouicros.substack.com/p/the-highs-and-lows-of-llms-in-strategy

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337-351. https://doi.org/10.1017/pan.2023.2.

Arora, N., Chakraborty, I., & Nishimura, Y. (2025). AI–human hybrids for marketing research: Leveraging large language models (LLMs) as collaborators. Journal of Marketing, 89(2), 43-70. https://doi.org/10.1177/00222429241276529

Bai, M., Cohen, A. K., Koss, E., & Lichtenbaum, C. (2025). Stuck in the matrix: Probing spatial reasoning in large language models. *arXiv preprint* arXiv:2510.20198. https://doi.org/10.48550/arXiv.2510.20198

Castricato, L., Lile, N., Rafailov, R., Fränken, J.-P., & Finn, C. (2025). PERSONA: A reproducible testbed for pluralistic alignment. En *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 11348-11368). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2407.17387

Chan, X., Wang, X., Yu, D., Mi, H., & Yu, D. (2024). Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint* arXiv:2406.20094. https://doi.org/10.48550/arXiv.2406.20094

Chen, J., Wang, X., Xu, R., Yuan, S., Zhang, Y., Shi, W., Xie, J., Li, S., Yang, R., Zhu, T., Chen, A., Li, N., Chen, L., Hu, C., Wu, S., Ren, S., Fu, Z., & Xiao, Y. (2024). From persona to personalization: A survey on role-playing language agents. *arXiv preprint* arXiv:2404.18231. https://doi.org/10.48550/arXiv.2404.18231

Chen, W., Su, Y., Zuo, J., Yang, C., Yuan, C., Chan, C.-M., Yu, H., Lu, Y., Hung, Y.-H., Qian, C., Qin, Y., Cong, X., Xie, R., Liu, Z., Sun, M., & Zhou, J. (2024). AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors. En *Proceedings of the International Conference on Learning Representations (ICLR 2024)*. https://doi.org/10.48550/arXiv.2308.10848

De Zarzà, I., De Curtò, J., Roig, G., & Calafate, C. T. (2023). Emergent cooperation and strategy adaptation in multi-agent systems: An extended coevolutionary theory with LLMs. *Electronics*, *12*(12), 2722. https://doi.org/10.3390/electronics12122722

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023). Opinion paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Erdmann, A., Ramos-Henriquez, J. M., Jimenez-Barreto, J., Bilgihan, A. (2026). Large Language Models in Pricing Research: An Evolutive Step from Human Surveys to Synthetic Samples. Working Paper.

Grundetjern, M., Andersen, P. A., & Goodwin, M. (2025). Synthetic personas: Enhancing demographic response simulation through large language models and genetic algorithms. *International Journal on Cybernetics & Informatics*, *14*(2), 21-40. https://doi.org/10.5121/ijci.2025.140202

Hart, S. (1992). Games in extensive and strategic forms. En R. J. Aumann & S. Hart (Eds.), Handbook of Game Theory with Economic Applications (Vol. 1, pp. 19–40). Elsevier. https://doi.org/10.1016/s1574-0005(05)80005-0

Heath, A. (2025, 21 de agosto). Amazon is betting on agents to win the AI race [episodio de podcast]. *Decoder with Nilay Patel*, The Verge. https://www.theverge.com/decoder-podcast-with-nilay-patel/761830/amazon-david-luan-agi-lab-adept-ai-interview

Knight, W. (2025, 9 de abril). The AI agent era requires a new kind of game theory. *Wired*.

Koc, V. (2023). Creating synthetic user research: Using persona prompting and autonomous agents. *Towards Data Science*. https://towardsdatascience.com/creating-synthetic-user-research-using-persona-prompting-and-autonomous-agents-b521e0a80ab6

Lazik, C., Katins, C., Kauter, C., Jakob, J., Jay, C., Grunske, L., & Kosch, T. (2025). The impostor is among us: Can large language models capture the complexity of human personas? En *Proceedings of the 2025 ACM Designing Interactive Systems Conference*. ACM. https://doi.org/10.1145/3743049.3743057

Levanti, M., & Verret, C. (2024, 26 de septiembre). The rise of synthetic respondents in market research. *NielsenIQ*. https://nielseniq.com/global/en/insights/education/2024/the-rise-of-synthetic-respondents/

Li, A., Chen, H., Namkoong, H., & Peng, T. (2025). LLM generated persona is a promise with a catch. *arXiv preprint* arXiv:2503.16527. https://doi.org/10.48550/arXiv.2503.16527

Lin, W., Roberts, J., Yang, Y., Albanie, S., Lu, Z., & Han, K. (2025). GAMEBoT: Transparent assessment of LLM reasoning in games. En *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 7656-7682). Association for Computational Linguistics.

Lyu, X. (2025). LLMs for multi-agent cooperation. https://xue-guang.com/post/llm-marl/

McKinsey & Company. (2024). *The state of AI in early 2024: Gen AI adoption spikes and starts to generate value*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

Mollick, E. (2024). *Co-intelligence: Living and working with AI*. Portfolio/Penguin.

Pansanella, V., Baronchelli, A., & Starnini, M. (2025). Emergent social conventions and collective bias in LLM populations. *Science Advances*. https://doi.org/10.1126/sciadv.adu9368

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. En *The 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)* (pp. 1-22). ACM. https://doi.org/10.1145/3586183.3606763

Sarstedt, M., Adler, S. J., Rau, L., & Schmitt, B. (2024). Using large language models to generate silicon samples in consumer and marketing research: Challenges,

opportunities, and guidelines. *Psychology & Marketing*, *41*(6), 1254-1270. https://doi.org/10.1002/mar.21982

Schwartz, S. H. (1992). Universals in the content and structure of values: Theory and empirical tests in 20 countries. En M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 25, pp. 1-65). Academic Press. https://doi.org/10.1016/S0065-2601(08)60281-6

Schwartz, S. H. (2003). A proposal for measuring value orientations across nations. En *Questionnaire development report of the European Social Survey* (pp. 259-319). https://www.europeansocialsurvey.org

Schwartz, S. H. (2005a). Basic human values: Their content and structure across countries. En A. Tamayo & J. B. Porto (Eds.), *Valores e comportamento nas organizações* [Values and behavior in organizations] (pp. 21-55). Vozes.

Schwartz, S. H. (2005b). Robustness and fruitfulness of a theory of universals in individual human values. En A. Tamayo & J. B. Porto (Eds.), *Valores e comportamento nas organizações* [Values and behavior in organizations] (pp. 56-95). Vozes.

Schwartz, S. H., Melech, G., Lehmann, A., Burgess, S., Harris, M., & Owens, V. (2001). Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. Journal of Cross-Cultural Psychology, 32(5), 519–542. https://doi.org/10.1177/0022022101032005001

Sharma, M. (2023). Exploring and improving the spatial reasoning abilities of large language models. arXiv preprint arXiv:2312.01054. https://doi.org/10.48550/arXiv.2312.01054

Venkit, P. N., Li, J., Zhou, Y., Srinath, M., Khandelwal, S., Ghosh, S., & Wilson, S. (2025). A tale of two identities: An ethical audit of human and AI-crafted personas. *arXiv preprint* arXiv:2505.07850. https://doi.org/10.48550/arXiv.2505.07850

Weavely.ai. (2025). Creating synthetic users for free with ChatGPT. https://www.weavely.ai/blog/creating-synthetic-users-for-free-with-chatgpt

Wu, Y., Shi, Q., Wan, F., Shao, J., Cai, J., & Lin, J. (2024). Mind's eye of LLMs: Visualization-of-thought elicits spatial reasoning in large language models. *arXiv preprint* arXiv:2404.03622. https://doi.org/10.48550/arXiv.2404.03622

Yan, B., Zhou, Z., Zhang, L., Zhang, L., Zhou, Z., Miao, D., Li, Z., Li, C., & Zhang, X. (2025). Beyond self-talk: A communication-centric survey of LLM-based multi-agent systems. *arXiv preprint* arXiv:2502.14321. https://doi.org/10.48550/arXiv.2502.14321

Yang, Y., Chai, H., Shao, S., Song, Y., Qi, S., Rui, R., & Zhang, W. (2025). AgentNet: Decentralized evolutionary coordination for LLM-based multi-agent systems. *arXiv preprint* arXiv:2504.00587. https://doi.org/10.48550/arXiv.2504.00587