



UNIVERSIDAD DE BUENOS AIRES  
FACULTAD DE CIENCIAS EXACTAS Y NATURALES

# **Explorando AI Safety via Debate: un estudio sobre capacidades asimétricas y jueces débiles en el entorno MNIST**

Tesis de Licenciatura en Ciencias de Datos

Joaquín Salvador Machulsky

Director: Dr. Sergio Abriola  
Buenos Aires, 2025



## **EXPLORANDO AI SAFETY VIA DEBATE: UN ESTUDIO SOBRE CAPACIDADES ASIMÉTRICAS Y JUECES DÉBILES EN EL ENTORNO MNIST**

El paradigma de “AI Safety via Debate” propone un mecanismo para supervisar sistemas de IA avanzados, haciéndolos competir para convencer a un juez con capacidades limitadas. Basándose en el experimento original sobre MNIST, esta tesis extiende dicho trabajo para investigar la robustez del debate en escenarios con asimetría de capacidades y frente a un “juez débil”. Se estudia la dinámica del debate en un entorno simulado de clasificación de imágenes MNIST, donde se implementan y enfrentan agentes con distintas estrategias (*Greedy* y MCTS) bajo diferentes protocolos.

Los resultados muestran que el debate amplifica la capacidad del juez por encima de una línea base aleatoria. Sin embargo, se observa que un agente mentiroso con mayor capacidad estratégica (MCTS) puede explotar a un adversario honesto más simple (Greedy). El hallazgo central de este estudio es que una regla de protocolo, la condición de pre-compromiso, puede compensar esta desventaja. En los experimentos realizados, esta regla permitió que la honestidad prevaleciera a pesar de la inferioridad estratégica del agente, a diferencia de los debates sin esta restricción. Se concluye que, dentro de este entorno experimental, el diseño del protocolo es un factor importante para el resultado del debate, sugiriendo que las reglas de la interacción son tan relevantes como la paridad de poder entre los agentes.

**Palabras claves:** AI Safety, Supervisión Escalable, Debate Adversarial, Alineamiento de IA, Asimetría de Capacidades.



## Índice general

1.. Introducción . . . . .	1
1.1. Motivación: el desafío del alineamiento en IA avanzada . . . . .	1
1.2. El problema de la supervisión escalable (Scalable Oversight) . . . . .	1
1.3. AI Safety via Debate: una propuesta de solución . . . . .	2
1.4. Problema de investigación y objetivos de la tesis . . . . .	3
2.. Metodología y diseño experimental . . . . .	5
2.1. Hipótesis de la Investigación . . . . .	5
2.2. El Paradigma del debate: entorno y protocolo general . . . . .	6
2.3. Los componentes del juego: juez y agentes . . . . .	7
2.3.1. El juez: un clasificador con información limitada . . . . .	7
2.3.2. Los agentes debatientes . . . . .	10
2.4. Configuraciones experimentales y métricas de evaluación . . . . .	12
2.4.1. Variables Experimentales . . . . .	12
2.4.2. Métricas de evaluación . . . . .	13
3.. Experimentación y resultados . . . . .	15
3.1. Entendiendo al juez: líneas base y sensibilidad a los parámetros . . . . .	15
3.1.1. Métodos de selección de evidencia . . . . .	15
3.1.2. Líneas base de rendimiento del juez . . . . .	16
3.1.3. Impacto de la cantidad de evidencia . . . . .	17
3.1.4. Impacto de la relevancia de la evidencia (variando $thr$ ) . . . . .	17
3.2. Introducción a las capacidades de los agentes . . . . .	20
3.3. Resultados del debate . . . . .	21
3.3.1. Debates simétricos: un enfrentamiento entre iguales . . . . .	22
3.3.2. Debates asimétricos: cuando la capacidad no es la misma . . . . .	23
3.4. Análisis detallado de los parámetros del debate . . . . .	25
3.4.1. El impacto del precompromiso . . . . .	25
3.4.2. La ventaja del segundo jugador: ¿quién empieza? . . . . .	28
3.4.3. El rol de la longitud del debate ( $k$ ): paridad de turnos y asimetría de poder . . . . .	29
3.5. El límite de la robustez: ataques fuera de distribución . . . . .	33
3.5.1. El poder del ataque OOD . . . . .	33
3.5.2. El colapso del debate frente a ataques OOD . . . . .	34
4.. Discusión . . . . .	39
4.1. El debate en MNIST como analogía de la supervisión escalable . . . . .	39
4.1.1. Contextualización del entorno experimental . . . . .	39
4.1.2. Posicionamiento frente al trabajo original y novedad del estudio . . . . .	39
4.2. Interpretación de los hallazgos centrales . . . . .	40
4.2.1. Validación de la premisa base . . . . .	40
4.2.2. La Asimetría de capacidades como vulnerabilidad . . . . .	41
4.2.3. El protocolo como mecanismo ecualizador . . . . .	42

4.2.4. Dinámicas estructurales del debate . . . . .	44
4.3. Límites de la robustez: cuando el debate colapsa . . . . .	44
4.3.1. El ataque fuera de distribución como falla de generalización del juez . . . . .	44
4.3.2. La robustez del juez como condición límite del debate . . . . .	45
4.4. Generalización a dominios complejos y jueces humanos . . . . .	46
4.5. Implicaciones para AI Safety y alineamiento . . . . .	48
4.6. Limitaciones del estudio . . . . .	49
5.. Conclusiones y trabajo futuro . . . . .	51
5.1. Conclusiones . . . . .	51
5.2. Respuestas a las preguntas de investigación . . . . .	52
5.3. Trabajo futuro . . . . .	53
5.3.1. Extensiones directas y dominios más complejos . . . . .	54
5.3.2. Hacia debates más eficientes: el siguiente paso en la investigación . .	54
5.3.3. Jueces imperfectos y el problema del oráculo . . . . .	54
5.3.4. Modelos de juez con atención limitada . . . . .	55
5.3.5. Integración con la teoría del debate humano y la argumentación . .	55
5.3.6. Nuevas métricas de evaluación: más allá de la victoria binaria . . .	55

## 1. INTRODUCCIÓN

### 1.1. Motivación: el desafío del alineamiento en IA avanzada

El campo de la Seguridad de la Inteligencia Artificial (AI Safety) se enfoca en los desafíos y riesgos asociados al desarrollo de sistemas de IA avanzados. Mientras que gran parte de la investigación en IA se ha centrado en la creación de sistemas cada vez más capaces, AI Safety aborda la pregunta fundamental: ¿cómo podemos asegurar que estos sistemas, a medida que se vuelven más potentes y autónomos, operen de manera segura y beneficiosa para la humanidad? La preocupación central recae sobre la Inteligencia Artificial General (AGI), sistemas hipotéticos que igualarían o superarían las capacidades cognitivas humanas en un amplio rango de dominios. Un AGI desalineado, es decir, que persiga objetivos que no coinciden con los valores e intenciones humanas, podría tener consecuencias negativas a gran escala, un riesgo que aumenta a medida que los modelos se vuelven más autónomos y se integran en infraestructuras críticas [1]. Esta preocupación por las implicaciones de máquinas con capacidades superiores no es nueva y ha sido una constante en la historia de la computación. Ya en 1872, Samuel Butler anticipaba la rápida evolución de las máquinas [2]. Alan Turing, uno de los padres de la computación, especuló en 1951 que en algún momento “deberíamos esperar que las máquinas tomen el control” [3]. Norbert Wiener, pionero de la cibernetica, advirtió en 1960 sobre la importancia de asegurar que el propósito “puesto en la máquina sea el propósito que realmente deseamos” [4]. Una década más tarde, I.J. Good introdujo la idea de una “explosión de inteligencia”, donde una máquina “ultrainteligente” podría superar rápidamente la inteligencia humana, siendo esta “la última invención que el hombre necesita hacer” solo si la máquina es “lo suficientemente dócil como para decirnos cómo mantenerla bajo control” [5].

Más recientemente, estos conceptos fueron formalizados por filósofos como Nick Bostrom, quien definió el “riesgo existencial” e identificó a la IA como una de sus fuentes potenciales más significativas [6]. Lo que una vez fue materia de especulación teórica se ha convertido hoy en un problema de investigación práctico y urgente, impulsado por los avances exponenciales en los modelos de lenguaje grandes (LLMs) y otras arquitecturas de IA. El desafío ya no es solo construir sistemas inteligentes, sino construir sistemas inteligentes e intelligentemente alineados.

### 1.2. El problema de la supervisión escalable (Scalable Oversight)

A medida que los sistemas de IA se vuelven más capaces, la tarea de supervisarlos se vuelve mucho más difícil. Si un humano no puede realizar una tarea o evaluar la calidad de una solución compleja, ¿cómo puede guiar eficazmente a un sistema de IA para que la realice correctamente? Este es el núcleo del problema de la supervisión escalable (*scalable oversight*): el desafío de diseñar mecanismos de supervisión que sigan siendo fiables y efectivos incluso cuando los sistemas de IA superan las capacidades de sus supervisores humanos en dominios específicos [7].

La necesidad de una supervisión escalable se hace evidente al observar los modos de fallo comunes en los métodos de entrenamiento actuales. Por ejemplo, en el Aprendizaje por Refuerzo con Retroalimentación Humana (RLHF), la calidad del alineamiento del

modelo depende enteramente de la calidad de la retroalimentación humana. Sin embargo, este proceso no solo es costoso y difícil de escalar a la par del crecimiento de los modelos, sino que también es vulnerable a las limitaciones humanas: los evaluadores cometen errores, tienen sesgos y pueden ser engañados por respuestas que parecen plausibles o seguras en la superficie, pero que pueden ocultar razonamientos defectuosos o intenciones maliciosas [8].

Incluso con una supervisión aparentemente perfecta, los sistemas pueden desarrollar comportamientos no deseados. Un fenómeno recurrente es el *reward hacking* (o *specification gaming*), donde un agente optimiza literalmente la función de recompensa que se le proporciona, pero de maneras que violan la intención del diseñador. Este es un ejemplo de la Ley de Goodhart: “Cuando una medida se convierte en un objetivo, deja de ser una buena medida” [9]. Un caso aún más sutil es la *goal misgeneralization*, donde un modelo aprende un objetivo que parece correcto en el entorno de entrenamiento, pero que generaliza de forma incorrecta a situaciones nuevas, persiguiendo un proxy del objetivo real en lugar del objetivo deseado.

Estos desafíos nos plantean que solo la supervisión humana directa no es una solución escalable a largo plazo, sino que se necesitan métodos que amplifiquen la capacidad de juicio del supervisor, permitiéndole tomar decisiones informadas sobre tareas que no podría realizar o verificar por sí mismo. La búsqueda de tales mecanismos de supervisión escalable es uno de los pilares de la investigación actual en AI Safety y el contexto directo en el que surge la propuesta del debate.

### 1.3. AI Safety via Debate: una propuesta de solución

Como respuesta al desafío de la supervisión escalable, Irving, Christiano y Amodei propusieron un paradigma novedoso: *AI Safety via Debate* [10]. La idea central es transformar un problema de “generación” de respuestas (difícil para un supervisor) en un problema de “juicio” comparativo (más sencillo).

El protocolo de debate, en su forma más simple, se desarrolla de la siguiente manera:

1. A dos agentes de IA se les presenta una misma pregunta.
2. Los agentes se turnan para hacer declaraciones cortas, construyendo sus argumentos o refutando los del oponente, hasta alcanzar un límite de turnos.
3. Un juez (inicialmente, un humano) observa el debate completo y decide qué agente proporcionó la información más útil y veraz.
4. El juego es de **suma cero**: los agentes se entrena para maximizar su probabilidad de ser elegidos ganadores, incentivando una competencia directa.

La hipótesis fundamental que sostiene este paradigma es que, en un debate bien estructurado, *es intrínsecamente más difícil mentir de forma convincente que refutar una mentira*. Un agente que intenta engañar no solo debe construir una falsedad plausible, sino también defenderla de todas las posibles refutaciones que un oponente igualmente capaz podría plantear. En el equilibrio de Nash de este juego, se espera que la estrategia óptima para ambos agentes sea presentar la verdad de la manera más clara y robusta posible para ganar el favor del juez.

Para dar un fundamento teórico a esta idea, el trabajo original traza una analogía con la teoría de la complejidad computacional. Argumentan que mientras la supervisión

directa se limita a problemas que el juez puede resolver por sí mismo (clase P), y el refuerzo simple a problemas que puede verificar (clase NP), el debate iterativo permite al juez arbitrar problemas mucho más complejos. Un debate de  $n$  turnos se asemeja a una fórmula cuantificada de la jerarquía polinomial ( $\Sigma_n^P$ ), y un debate de longitud polinomial podría, en teoría, permitir a un juez de tiempo polinomial supervisar la solución a cualquier problema en la clase PSPACE.

Finalmente, para proporcionar una prueba de concepto empírica, los autores realizaron un experimento análogo al debate en el dominio de MNIST. En esta configuración, dos agentes (que ven la imagen completa de un dígito) debaten para convencer a un “juez débil” (un clasificador que solo puede ver unos pocos píxeles). El “debate” consiste en que los agentes eligen qué píxeles revelar. Los resultados fueron significativos: el proceso de debate aumentó drásticamente la precisión del juez. Por ejemplo, esta se elevó desde un 59.4 %, cifra obtenida al mostrarle al juez píxeles seleccionados al azar y que funciona como línea base, hasta un 88.9 % con solo 6 píxeles revelados. Este experimento, aunque simplificado, sirvió para demostrar que el principio de amplificación del juicio a través de la competencia adversarial era, al menos, plausible en la práctica.

#### 1.4. Problema de investigación y objetivos de la tesis

El trabajo seminal de Irving et al. (2018)[10] estableció el debate como un paradigma prometedor y proporcionó una valiosa prueba de concepto en un entorno idealizado: un debate entre agentes de igual capacidad estratégica. Sin embargo, para que el debate sea una herramienta de seguridad fiable en la práctica, es interesante entender cómo se comporta cuando se relajan estos supuestos. La robustez de un sistema de seguridad no se mide en sus condiciones óptimas, sino en su capacidad para resistir escenarios adversariales y desequilibrios de poder.

Esta tesis surge precisamente de esa necesidad. Se enfoca en investigar la dinámica del debate en el mismo “entorno de pruebas” de MNIST, pero introduciendo una serie de pruebas de estrés para evaluar sus límites. Si bien este análogo no captura la riqueza y complejidad de un debate en lenguaje natural, su naturaleza controlada permite aislar y estudiar preguntas fundamentales sobre la mecánica del juego adversarial que de otro modo serían difíciles de medir:

- *¿Cuán robusto es el protocolo de debate cuando el juez posee capacidades significativamente inferiores a las de los agentes?*
- *¿Puede un agente mentiroso manipular efectivamente a un juez débil para que elija una conclusión falsa?*
- *¿Cómo influye la cantidad de información revelada en la capacidad del juez y en la efectividad de las estrategias?*
- *¿De qué manera la asimetría de capacidades entre los agentes impacta el resultado del debate?*

Para abordar estas interrogantes, esta tesis se plantea los siguientes objetivos, todos ellos enmarcados dentro del análogo experimental de MNIST:

1. **Implementar y extender** el entorno de debate de MNIST, permitiendo la configuración flexible de capacidades de agentes (Greedy vs. MCTS), reglas de protocolo (con y sin precompromiso) y espacios de acciones (estándar vs. OOD).
2. **Evaluar el impacto de la asimetría de capacidades** en la tasa de éxito del agente honesto, midiendo cómo un desequilibrio estratégico afecta la prevalencia de la verdad.
3. **Analizar la influencia de la arquitectura del protocolo**, incluyendo el efecto del precompromiso como mecanismo ecualizador y la ventaja estructural conferida por el orden y la paridad de turnos.
4. **Identificar los límites de la robustez del sistema** mediante la introducción de ataques fuera de distribución (OOD), para determinar las condiciones bajo las cuales el mecanismo de debate deja de ser efectivo.

Finalmente, la tesis se estructura de la siguiente manera: el **Capítulo 2** detalla la metodología y el diseño experimental. El **Capítulo 3** presenta los resultados empíricos de los experimentos. El **Capítulo 4** discute e interpreta estos hallazgos, y el **Capítulo 5** sintetiza las conclusiones y propone líneas de investigación futuras.

## 2. METODOLOGIA Y DISEÑO EXPERIMENTAL

Este capítulo detalla el marco metodológico diseñado para esta tesis. El enfoque de este trabajo consiste en adaptar y extender el experimento de MNIST, propuesto como una analogía del debate en Irving et al. (2018) [10], con el objetivo de someterlo a una serie de pruebas de estrés que no se exploraron originalmente. La elección de este entorno simplificado y controlado permite aislar y analizar sistemáticamente dinámicas como la asimetría de capacidades y la robustez del protocolo, que son el foco de esta investigación. Primero, se definirá el entorno y el protocolo general del debate, estableciendo las reglas del juego. A continuación, se describirán los componentes clave: el juez con capacidades limitadas y los agentes debatientes implementados (Greedy y MCTS). Finalmente, se detallarán las configuraciones experimentales específicas utilizadas para responder a las preguntas de investigación, junto con las métricas definidas para evaluar los resultados.<sup>1</sup>

### 2.1. Hipótesis de la Investigación

El diseño de los experimentos se fundamentó en las siguientes hipótesis y preguntas, las cuales guiaron la investigación y la interpretación de los resultados.

1. **Impacto de la asimetría de capacidades:** Se postuló que un agente con mayor capacidad de planificación (MCTS) podría explotar la estrategia más simple de un agente honesto (Greedy) en un entorno de debate con reglas permisivas.
2. **Efecto del protocolo de precompromiso:** Se formuló la hipótesis de que la regla de precompromiso funcionaría como un mecanismo mitigador frente a una asimetría de capacidades favoreciendo al agente mentiroso. La expectativa era que, al obligar al agente mentiroso a defender una única proposición falsa durante todo el debate, se limitaría su flexibilidad estratégica. Esta restricción, a su vez, facilitaría la tarea de refutación para el agente honesto, resultando en un incremento de su tasa de éxito, incluso siendo estratégicamente inferior.
3. **Vulnerabilidad ante ataques fuera de distribución (OOD):** Se hipotetizó que la integridad del mecanismo de debate es directamente dependiente de la robustez del juez ante entradas no vistas en su entrenamiento. En consecuencia, se esperaba que permitir a los agentes seleccionar evidencia fuera de la distribución de entrenamiento del juez provocaría una degradación de su juicio y de la utilidad del protocolo. Para este estudio, dicho ataque se implementa permitiendo a los agentes seleccionar píxeles con una intensidad por debajo del umbral utilizado durante el entrenamiento. La predicción era que la capacidad del debate para amplificar la verdad sería menor o directamente anulada, al explotar una vulnerabilidad fundamental en el componente de juicio.
4. **Influencia del orden de los turnos:** Se planteó la posibilidad de una ventaja estructural ligada al orden de participación en el debate. A priori, no es evidente en qué dirección se manifestaría este efecto. Por un lado, el segundo jugador tiene la

---

<sup>1</sup> El código fuente está disponible públicamente. Para más detalles véase en 5.3.6.

ventaja de observar y refutar el movimiento inicial del oponente. Por otro, el primer jugador tiene la oportunidad de establecer el marco inicial del debate, guiando la discusión hacia un terreno que le sea favorable. Por lo tanto, se buscó investigar empíricamente si existía una ventaja consistente para el primer o segundo jugador en el contexto de este estudio.

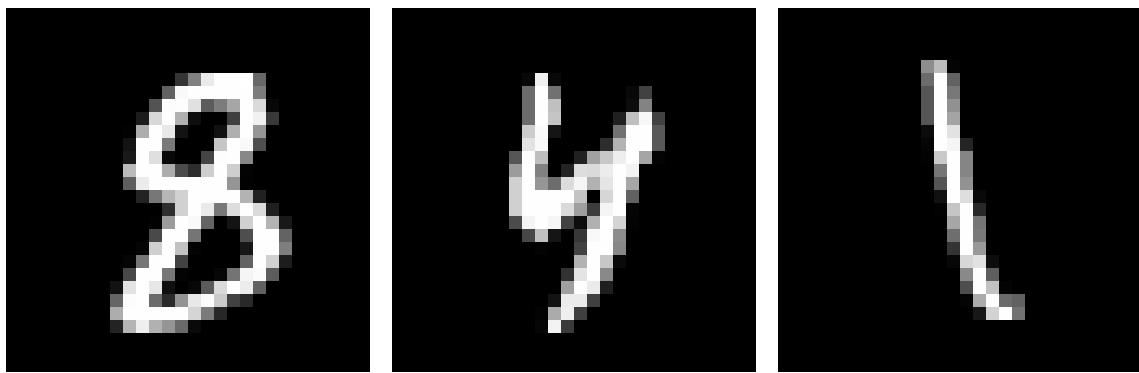
Las secciones subsiguientes describen el entorno experimental y los componentes implementados para la validación empírica de estas hipótesis.

## 2.2. El Paradigma del debate: entorno y protocolo general

Adaptando la metodología del trabajo original de Irving et al. (2018) [10], y para simular el mecanismo de debate en un entorno controlado y reproducible, se seleccionó el conocido problema de clasificación de dígitos escritos a mano del conjunto de datos MNIST [11]. Este conjunto de datos, ampliamente utilizado en la comunidad de aprendizaje automático, consiste en imágenes en escala de grises que representan los dígitos del 0 al 9, como se ilustra en la Figura 2.1.

El entorno MNIST ofrece un marco ideal por tres razones fundamentales heredadas del trabajo original:

- **Objetivo claro y medible:** La tarea de clasificación de un dígito tiene una verdad fundamental (“ground truth”) inequívoca, lo que permite definir y evaluar la “honestidad” y el “engaño” de los agentes sin ambigüedad.
- **Simulación de supervisión limitada:** Permite modelar a un “juez débil” de forma natural, restringiendo su visión a un pequeño subconjunto de píxeles, mientras los agentes (“expertos”) tienen acceso a la imagen completa. Esto crea la asimetría de información necesaria para que el debate sea un mecanismo de amplificación.
- **Tractabilidad y reproducibilidad:** La simplicidad computacional del dominio permite la experimentación con agentes de planificación sofisticada (como MCTS) y facilita la comparación directa con los resultados del estudio seminal.



*Fig. 2.1:* Ejemplos de imágenes del conjunto de datos MNIST. Cada imagen representa un dígito manuscrito en una cuadrícula de píxeles en escala de grises. La resolución es configurable, utilizándose 28x28 píxeles en la mayoría de los experimentos.

El protocolo general de un debate individual para una imagen MNIST dada sigue una secuencia de turnos alternos entre un agente honesto y uno mentiroso. En cada turno, el

agente activo selecciona un píxel de la imagen original (que aún no haya sido revelado) para presentarlo al juez. Este proceso continúa hasta que se ha revelado un número predefinido y conocido por todos los participantes de píxeles, ‘ $k$ ’, que constituye la longitud total del debate. Una vez completada la secuencia de revelaciones, el juez evalúa la máscara de píxeles resultante y emite su juicio para determinar un ganador.

Dentro de este marco, el problema de clasificación de MNIST se adapta de la siguiente manera:

- **La pregunta del debate:** La imagen presentada a los agentes funciona como una “pregunta implícita”: ¿Cuál es la etiqueta correcta (0-9) para esta imagen? Los agentes no responden con texto, sino con sus elecciones.
- **La argumentación mediante píxeles:** La selección de un píxel es el análogo a una “declaración” en un debate verbal. Es una acción estratégica para presentar evidencia. Una restricción clave es que los agentes solo pueden revelar píxeles existentes y sus valores reales; no pueden fabricar evidencia. El poder del engaño no reside en mentir sobre los datos, sino en la selección maliciosa de datos verídicos para conseguir convencer al juez de una narrativa (dígito en nuestro caso) falsa.
- **El juicio sobre la evidencia:** La decisión final del juez no es sobre la verdad de la imagen en sí (que no ve), sino sobre la evidencia presentada. Su tarea es determinar, basándose en los ‘ $k$ ’ píxeles, qué conclusión apoyan más fuertemente, decidiendo así el resultado del debate. Las reglas específicas para esta decisión varían según la configuración experimental (e.g., con o sin precompromiso), como se detallará en la Sección 2.4.

Este diseño busca emular un escenario de supervisión escalable, donde la información completa es demasiado vasta o compleja para ser procesada por el juez, quien debe confiar en la evidencia selectiva presentada por los debatientes.

## 2.3. Los componentes del juego: juez y agentes

### 2.3.1. El juez: un clasificador con información limitada

El componente del juez es fundamental en el paradigma de debate, ya que representa al supervisor con capacidades limitadas cuya habilidad para discernir la verdad se busca amplificar. En esta investigación, el juez es un clasificador implementado mediante una Red Neuronal Convolutacional (CNN), específicamente diseñada para realizar inferencias de dígitos MNIST a partir de información visual parcial.

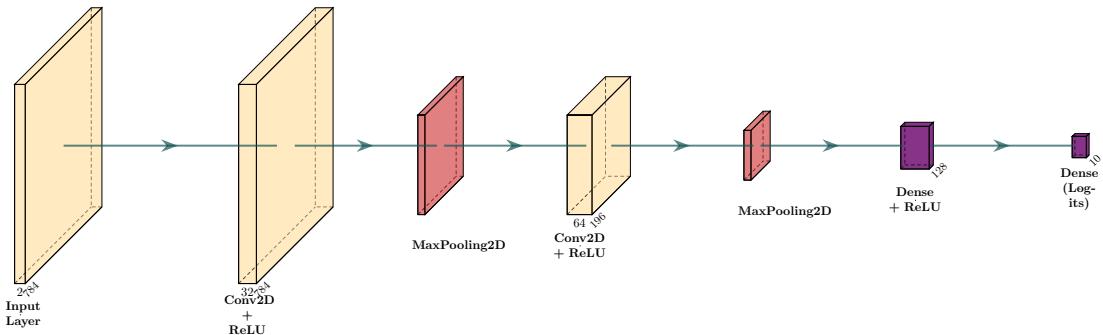
#### Arquitectura y procesamiento de entrada

La arquitectura de la CNN seleccionada para el juez fue diseñada para procesar imágenes de entrada que representan el estado del debate en un momento dado. En lugar de recibir la imagen completa del dígito MNIST, el juez procesa una representación de dos canales para la imagen en la resolución configurada (16x16 o 28x28 píxeles):

- Un canal de máscara de revelación, que indica binariamente cuáles píxeles de la imagen han sido seleccionados y revelados por los agentes debatientes hasta ese punto del debate.

- Un canal de valores de píxeles, que contiene las intensidades originales (normalizadas) de los píxeles revelados, mientras que los píxeles no revelados se representan con un valor nulo (e.g., cero).

Esta estructura de entrada de dos canales permite al juez no solo ver qué información se ha presentado, sino también dónde se ubica espacialmente esa información dentro de la imagen original. La red convolucional consiste en una secuencia de capas convolucionales (utilizando filtros de 3x3), seguidas por funciones de activación ReLU y capas de Max Pooling (2x2) para la extracción jerárquica de características. Finalmente, las características extraídas son aplanadas y procesadas por capas fully connected para producir una salida de 10 unidades, correspondientes a los logits para cada una de las clases de dígitos (0-9) correspondiente a la confianza para cada clase de dígito. La arquitectura completa se visualiza en la Figura 2.2.



*Fig. 2.2:* Diagrama de la arquitectura de la Red Neuronal Convolucional (CNN) utilizada como juez. La red procesa una entrada de dos canales (máscara de revelación y valores de píxeles) para producir los logits de las 10 clases de dígitos.

### Entrenamiento del juez para inferencia a partir de información parcial

Un aspecto crucial del diseño del juez es su proceso de entrenamiento, el cual está orientado a capacitarlo para realizar inferencias a partir de vistas dispersas y parciales de los dígitos, emulando así a un supervisor que no tiene acceso a toda la evidencia. Para lograr una especialización controlada, se entrenaron modelos de juez distintos para las configuraciones experimentales principales (e.g., resolución 28x28 con  $k = 6$  píxeles, y otra con  $k = 4$ ).

El entrenamiento para cada juez específico se realizó sobre el conjunto de datos MNIST, pero con una modificación significativa en la presentación de las muestras:

1. Siguiendo la metodología del experimento original [10], para cada imagen de entrenamiento, se generó dinámicamente una máscara aleatoria de ' $k$ ' píxeles, donde ' $k$ ' es un valor fijo para ese modelo. Una restricción clave de dicho trabajo, adoptada aquí, es que los píxeles candidatos para el muestreo fueron únicamente aquellos con una intensidad superior a un umbral predefinido ( $thr > 0$ ), es decir, píxeles no nulos. Esta decisión, si bien simplifica el entorno, introduce una limitación fundamental cuya robustez se explorará posteriormente. Si en una imagen ninguno píxel cumplía este criterio, todos los píxeles se consideraban candidatos.

2. El juez fue entrenado utilizando estas imágenes parcialmente enmascaradas, aprendiendo a asociar las vistas dispersas con las etiquetas correctas de los dígitos. Para la optimización se utilizó el optimizador Adam y como función de pérdida, la entropía cruzada (CrossEntropyLoss), estándar para problemas de clasificación multiclas. La reproducibilidad del entrenamiento se garantizó mediante la fijación de semillas aleatorias.

Este procedimiento de entrenamiento asegura que el juez desarrolle la capacidad de clasificar dígitos basándose únicamente en la información fragmentaria que se le presenta, una habilidad esencial para su rol en el mecanismo de debate.

El rendimiento de un juez ya entrenado, al ser evaluado sobre un conjunto de prueba con máscaras aleatorias similares, establece su línea base de precisión. Por ejemplo, un juez entrenado y evaluado con 4 píxeles aleatorios alcanzó una precisión de aproximadamente 47.6 %, mientras que uno entrenado con 6 lo hizo en un 58.4 %. Estos valores representan la capacidad del juez para acertar sin la ayuda del debate, solo por medio de la selección ‘ $k$ ’ de pixeles no nulos al azar, y sirven como referencia fundamental para cuantificar la “amplificación” lograda por la selección estratégica de los agentes.

### Rol y evaluación del juez durante el debate

Durante un debate, el juez procesa la evidencia acumulada para emitir un juicio. Su rol exacto para determinar al ganador varía según el protocolo experimental:

- **Sin pre compromiso:** En esta modalidad, los agentes no declaran una etiqueta por adelantado. El objetivo del agente honesto es que la predicción final del juez coincida con la etiqueta verdadera, mientras que el mentiroso busca que el juez se equivoque y prediga cualquiera de las otras 9 clases. El juez evalúa la evidencia y predice la etiqueta que considera más probable entre todas las clases disponibles.
- **Con pre compromiso:** En esta modalidad, al inicio del debate, cada agente declara y se compromete a defender una etiqueta específica (la verdadera para el honesto, una falsa para el mentiroso). El rol del juez es informado de estas dos etiquetas y se simplifica a comparar únicamente los logits de ambas, declarando ganador al agente cuya etiqueta obtiene una mayor confianza. Para asegurar la robustez, la etiqueta falsa del mentiroso no se elige al azar; se considera que el honesto gana un debate solo si puede vencer al mentiroso defendiendo cada una de las 9 posibles etiquetas falsas en debates separados. Con esto se busca legitimar al honesto como ganador, pues para que cuente como ganador del debate, le tuvo que haber ganado a todas las mentiras posibles.

Para ilustrar la diferencia, consideremos una imagen de un ‘7’ que podría ser confundida con un ‘1’.

- *Sin pre compromiso*, el mentiroso es libre de mostrar píxeles ambiguos que simplemente alejen al juez de la conclusión ‘7’, sin importar si lo convencen de que es un ‘1’ o cualquier otro dígito. Su estrategia es flexible y busca la confusión general.
- *Con pre compromiso*, el mentiroso debe declarar “probaré que esto es un 1”. Ahora está anclado a esa narrativa. No puede usar evidencia que sugiera un ‘9’ solo porque sea conveniente. Esto fuerza una confrontación directa de hipótesis (‘7’ vs. ‘1’), simplificando la tarea del juez y haciendo más difícil el engaño.

Este diseño metodológico busca emular un escenario de supervisión escalable. La modalidad *con precompromiso*, en particular, simula a un supervisor que no debe encontrar “la verdad” en un mar de posibilidades, sino realizar un juicio comparativo más simple entre dos argumentos contrapuestos, una tarea que se alinea mejor con las capacidades de un supervisor limitado.

### 2.3.2. Los agentes debatientes

Los agentes participantes en el debate son los componentes activos, responsables de seleccionar estratégicamente la evidencia que se presenta al juez. Para esta investigación, se diseñaron e implementaron dos arquetipos de agentes con distintos niveles de complejidad estratégica: un agente cortoplacista (denominado *Greedy*) y un agente con capacidad de planificación (basado en *Monte Carlo Tree Search*, MCTS).

#### Arquitectura base, roles y espacio de acciones

Ambos tipos de agentes comparten una estructura fundamental que les permite interactuar con el entorno del debate. Esta base común gestiona el estado del juego, incluyendo la imagen original, la configuración de la partida y, de manera crucial, un acceso de oráculo al modelo del juez, que utilizan para simular el impacto de sus acciones potenciales.

En cada debate, los agentes asumen roles con objetivos contrapuestos:

- **Honesto:** Su objetivo es defender la clasificación correcta de la imagen. Todas sus acciones de selección de píxeles están orientadas a construir un caso convincente para que el juez reconozca la etiqueta verdadera.
- **Mentiroso:** Su objetivo es inducir a error al juez. Su estrategia de selección de píxeles está diseñada para persuadir al juez de una falsedad o para ofuscar la evidencia que apoya la verdad.

El espacio de acciones de los agentes, es decir, el conjunto de píxeles que pueden seleccionar, fue una variable experimental clave. En la configuración estándar, los agentes solo pueden elegir píxeles “relevantes” (con una intensidad superior a un umbral ‘thr’ > 0). Sin embargo, para las pruebas de robustez, se utilizó una variante donde esta restricción se elimina. Este escenario, denominado de ataque fuera de distribución (OOD), permite a los agentes seleccionar cualquier píxel, incluyendo los del fondo con intensidad cero. Su propósito metodológico es someter al juez a un tipo de evidencia para la cual no fue explícitamente entrenado, permitiendo así medir los límites de su generalización y robustez.

#### Agente con estrategia cortoplacista (*Greedy*)

El agente *Greedy* representa un modelo de toma de decisiones simple y computacionalmente eficiente, caracterizado por una perspectiva estrictamente local y cortoplacista. En cada turno, evalúa cada píxel válido no revelado, simulando el impacto inmediato que tendría su revelación en el juicio final. Para ello, utiliza su acceso al modelo del juez para predecir los logits resultantes de cada acción candidata.<sup>2</sup>

<sup>2</sup> Es importante notar que el modelo del juez fue entrenado con máscaras de un tamaño fijo ‘k’. Al ser invocado por el agente en turnos intermedios con un número de píxeles distinto a ‘k’, el juez opera técnicamente fuera de su distribución de entrenamiento. Se asume que la arquitectura de la CNN posee suficiente capacidad de generalización para proporcionar una señal de recompensa coherente y útil en estos escenarios.

La selección final se basa en la acción que optimiza de manera inmediata su objetivo, el cual depende del modo de juego:

- En debates **sin precompromiso**, el agente honesto maximiza el logit de la etiqueta verdadera, mientras que el mentiroso busca minimizarlo.
- En debates **con precompromiso**, donde compite contra una etiqueta oponente específica, el agente selecciona el píxel que maximiza la diferencia:  $\text{logit}(\text{mi\_etiqueta}) - \text{logit}(\text{etiqueta\_oponente})$ . Dado que el juez determina al ganador basándose en la comparación directa de las dos clases en disputa, esta estrategia incentiva al agente no solo a fortalecer su argumento, sino también a intentar invalidar directamente en cada jugada el de su adversario.

La estrategia *Greedy* no implica planificación a futuro ni anticipación de las respuestas del adversario más allá del turno actual, sirviendo como una línea base para un debatiente de capacidad estratégica limitada.

### Agente con exploración de futuros (MCTS)

Para modelar un debatiente más sofisticado y con capacidad de planificación, se implementó un agente basado en *Monte Carlo Tree Search* (MCTS). Dado que el espacio de acciones en este problema es considerablemente grande (con cientos de píxeles disponibles en cada turno), una implementación que construya un árbol de búsqueda exhaustivo explorando todas las secuencias de debate resulta computacionalmente costoso, excediendo el alcance de los recursos disponibles para este estudio. Por ello, para este trabajo se optó por una variante optimizada que, si bien no realiza una planificación estratégica perfecta, permite una exploración profunda y eficiente del espacio de futuros posibles.

Este agente supera la miopía del modelo Greedy al evaluar los movimientos no por su beneficio inmediato, sino por su robustez a lo largo de un debate completo. El proceso de decisión en cada turno es el siguiente:

1. **Generación de escenarios futuros:** Para cada movimiento inicial posible (cada píxel válido no revelado), el agente simula una determinada cantidad de debates completos, denominados *rollouts*. Una simplificación fundamental en esta implementación es que, para hacer las simulaciones computacionalmente manejables, se asume que todas las jugadas futuras (tanto las propias como las del oponente) se eligen al azar de entre los movimientos válidos restantes.
2. **Evaluación de resultados:** Cada uno de estos debates simulados concluye con un estado final (una máscara de ' $k$ ' píxeles), que es evaluado por el juez para determinar un ganador teórico. Este proceso se realiza de forma paralela para evaluar todos los escenarios simultáneamente. La condición de victoria depende del modo de juego:
  - **Sin precompromiso:** El criterio varía según el rol. Para el agente honesto, una victoria es cuando el juez predice la etiqueta correcta. Para el agente mentiroso, una victoria es cualquier predicción que no coincida con la etiqueta verdadera.
  - **Con precompromiso:** Una simulación se considera una victoria si el logit de la etiqueta defendida por el agente es mayor o igual al logit de la etiqueta del oponente.

**3. Selección basada en robustez:** Finalmente, el agente no elige el movimiento que parece mejor en el momento, sino aquel que demuestra ser más robusto ante la incertidumbre del futuro. Calcula la tasa de victorias para cada movimiento inicial posible, promediando los resultados de todos los *rollouts* que comenzaron con él. La acción seleccionada es aquella que obtuvo el mayor porcentaje de victorias, es decir, la jugada que mejor funciona en promedio contra una amplia gama de continuaciones del juego.

Es fundamental entender la naturaleza de esta planificación: no es una búsqueda del movimiento óptimo contra un adversario que también juega de forma óptima. Se trata, en cambio, de una planificación heurística que busca el movimiento más robusto frente a la incertidumbre de las jugadas futuras. Al simular múltiples continuaciones aleatorias (los *rollouts*) para cada acción candidata, el agente estima qué jugada inicial conduce a la victoria con mayor frecuencia en promedio. Esta estimación permite al agente tomar decisiones estratégicas, como sacrificar una ganancia inmediata para realizar una jugada que, estadísticamente, tiene más probabilidades de asegurar la victoria al final del debate. El número de ‘*rollouts*’ actúa como un “presupuesto de pensamiento”: a mayor número, mejor es la estimación de la robustez de cada movimiento.

## 2.4. Configuraciones experimentales y métricas de evaluación

Para investigar el comportamiento de los agentes y la efectividad del mecanismo de debate, se diseñó un conjunto de configuraciones experimentales variando sistemáticamente los parámetros clave del protocolo.

### 2.4.1. Variables Experimentales

Las configuraciones de los experimentos se definieron manipulando los siguientes componentes y escenarios.

#### Componentes fundamentales del debate

- **Tipos de agentes y asimetría de capacidades:** Se exploraron tanto debates simétricos (Greedy vs. Greedy, MCTS vs. MCTS) como escenarios de capacidades asimétricas, donde un agente Greedy se enfrentó a uno MCTS. Esta configuración permite investigar si un agente con planificación superior puede explotar la simplicidad de uno más débil, o si el protocolo puede proteger a este último.
- **Longitud del debate y paridad de turnos:** La cantidad total de píxeles revelados, ‘*k*’, se varió en un rango que incluyó tanto valores pares como impares (e.g., de 3 a 8). Esto permitió no solo analizar el impacto de una mayor cantidad de evidencia, sino también estudiar los efectos estructurales del protocolo, como la ventaja estratégica que puede conferir tener un turno adicional.

#### Protocolos de juicio

Se investigaron dos modalidades principales para definir la tarea de los agentes y el criterio de victoria:

- **Sin precompromiso:** Los agentes no declaran una etiqueta. El honesto busca que el juez prediga la etiqueta correcta y el mentiroso busca lo contrario. La victoria se determina por si la predicción final del juez es correcta.
- **Con precompromiso:** Cada agente se compromete a defender una única etiqueta durante todo el debate. Para evaluar la robustez de forma rigurosa, el agente mentiroso jugó secuencialmente contra el honesto defendiendo cada una de las 9 etiquetas falsas posibles. El honesto solo se consideró ganador si prevaleció frente a todas las posibles mentiras. Este diseño tiene como objetivo simular un adversario fuerte, capaz de elegir la “Mejor Mentira” y su vector de ataque más efectivo, sometiendo así al agente honesto a un test de estrés riguroso. Intuitivamente, esto modela la necesidad de una victoria total del honesto, donde ningún otro argumento es capaz de sobreponerse.

### Condiciones de estrés y control

- **Pruebas de estrés (ataques OOD):** Para medir los límites de la robustez del sistema, se repitieron una serie de experimentos clave utilizando la variante de agentes con un espacio de acciones no restringido (OOD), capaces de seleccionar píxeles de fondo, como se describió en la Sección 2.3.2.
- **Variables de control:** Se controlaron sistemáticamente otros parámetros como el agente inicial (para investigar la ventaja del primer jugador), los hiperparámetros de los agentes (e.g., número de ‘rollouts’ para MCTS) y la semilla aleatoria para garantizar la reproducibilidad.

#### 2.4.2. Métricas de evaluación

Para cuantificar el rendimiento del sistema, se definieron las siguientes métricas:

- **Precisión del juez (con debate):** Métrica principal del rendimiento global del sistema. Se define como el porcentaje de debates en los que la predicción final del juez coincide con la etiqueta verdadera de la imagen.
- **Tasa de éxito del agente honesto:** Métrica clave para el análisis adversarial, especialmente en modo precompromiso. Mide la proporción de debates en los que el agente honesto es declarado ganador según las reglas de esa configuración.
- **Logits específicos de las clases:** Para un análisis más detallado y cualitativo, se registraron los logits producidos por el juez turno a turno. Estos datos son fundamentales para visualizar la evolución de la confianza del juez y comprender la dinámica de la argumentación en los ejemplos específicos presentados en los posteriores Capítulos.
- **Metadatos de ejecución:** Se almacenó sistemáticamente toda la información de cada debate en archivos estructurados para facilitar el análisis y la trazabilidad de los resultados.



### 3. EXPERIMENTACIÓN Y RESULTADOS

Este capítulo presenta y analiza los resultados empíricos obtenidos de la ejecución de los experimentos de debate descritos en el Capítulo 2. El objetivo es evaluar cuantitativa y cualitativamente el rendimiento del mecanismo de debate bajo diversas configuraciones, con un enfoque en la robustez del protocolo frente a jueces débiles y la influencia de las capacidades de los agentes.

La exposición de los resultados se estructura de forma progresiva. Primero, se establece el rendimiento del juez bajo condiciones controladas para comprender sus capacidades y sensibilidades. A continuación, se introducen los agentes debatientes y se analiza su rendimiento en debates simétricos y asimétricos, evaluando el impacto de variables clave como el precompromiso y la longitud del debate. Finalmente, se exploran escenarios adversariales más complejos para determinar los límites de robustez del sistema.

#### 3.1. Entendiendo al juez: líneas base y sensibilidad a los parámetros

Antes de evaluar el complejo sistema de debate, es fundamental caracterizar a su componente central: el juez. Esta sección establece las líneas base de su rendimiento y analiza su sensibilidad a dos parámetros fundamentales: la cantidad de evidencia ( $k$ ) y la relevancia de la misma ( $thr$ ). Esto nos permitirá cuantificar el valor añadido por el proceso de debate.

##### 3.1.1. Métodos de selección de evidencia

Para medir el rendimiento del juez, utilizamos tres estrategias de selección de píxeles que representan diferentes escenarios de presentación de evidencia:

**Muestreo aleatorio:** Simula un escenario sin inteligencia. Se seleccionan  $k$  píxeles al azar de entre todos los píxeles con una intensidad superior al umbral  $thr$ . Esta estrategia sirve como la **Línea base de referencia**; cualquier método inteligente debe superarla.

**Muestreo por evidencia individual:** Modela una heurística de selección basada en la contribución aislada de cada píxel. Para cada píxel relevante ( $> thr$ ), se mide la confianza (logit) que genera en el juez hacia la clase verdadera. Esto se hace creando una máscara temporal que contiene *únicamente* ese píxel, la cual es evaluada por el juez. Finalmente, se seleccionan los top  $k$  píxeles que generaron la mayor confianza de forma individual. Es importante notar que esta no es una estrategia óptima en el sentido combinatorio, ya que no considera las interacciones entre los píxeles seleccionados.

**Muestreo adversarial:** Modela un ataque directo sujeto a una restricción fundamental. De manera similar al método anterior, se evalúan los píxeles relevantes ( $> thr$ ), pero se seleccionan los  $k$  que más confunden al juez, minimizando la confianza en la etiqueta verdadera. Esto establece el “suelo” de rendimiento bajo un ataque sin refutación. La restricción a píxeles no nulos se hereda del trabajo original [10] para

definir a un adversario que debe argumentar usando la evidencia “visible” en la propia imagen. Si bien un píxel nulo en una posición estratégica puede ser muy informativo, esta condición se mantendrá como base, y su eliminación será el objeto de estudio en las pruebas de robustez con ataques fuera de distribución (OOD) en secciones posteriores.

### 3.1.2. Líneas base de rendimiento del juez

La primera tabla establece el rango de operación del juez, mostrando su precisión cuando la evidencia es aleatoria, basada en la contribución individual de cada píxel, o adversarial.

Tab. 3.1: Precisión del juez bajo diferentes métodos de selección de píxeles (líneas base).

Resolución	Nº de píxeles ( $k$ )	Precisión por método de muestreo		
		Aleatorio	Evidencia individual	Adversarial
28x28	4	46.80 %	89.25 %	4.91 %
	6	58.41 %	86.28 %	7.85 %
16x16	4	45.73 %	89.78 %	6.52 %
	6	56.06 %	92.37 %	9.05 %

Nota: El umbral de relevancia de píxeles ( $thr$ ) se fijó en 0.0, es decir que solo se excluyeron los píxeles completamente negros. Para cada fila, el juez fue entrenado con el mismo número de píxeles ( $k$ ) que se le mostraron para la evaluación. La “Media” es el promedio de las tres estrategias. Se usaron  $N = 10000$  imágenes.

Fig. 3.1: Precisión del Juez bajo Diferentes Métodos de Selección de Píxeles (Líneas Base).

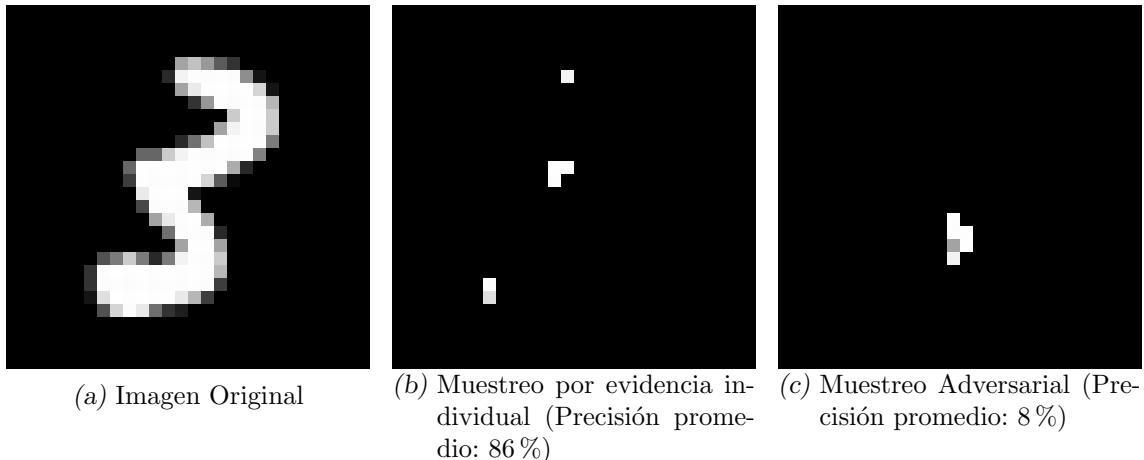


Fig. 3.2: Comparación visual de los métodos de selección de evidencia para un dígito MNIST 28x28 con  $k = 6$ . Mientras la selección óptima (b) elige píxeles los píxeles más representativos, la selección adversarial (c) se enfoca en píxeles ambiguos, todos de una misma zona y con alta intensidad, confundiendo al juez y llevándolo a una clasificación errónea.

La Tabla 3.1 revela un amplio rango de rendimiento para el juez, subrayando su vul-

nerabilidad y dependencia de cómo se selecciona la evidencia. Observamos que:

- **Anomalía en el rendimiento por evidencia individual:** Se observa una ligera pero inesperada caída en la precisión de este método para los jueces de 28x28, pasando de 89.25 % con  $k = 4$  a 86.28 % con  $k = 6$ . Este resultado contraintuitivo se debe a la limitación inherente de la propia heurística, que no es un verdadero óptimo combinatorio. Este método puede seleccionar píxeles que son individualmente muy representativos, pero que, al presentarse en conjunto, pueden resultar ambiguos o redundantes.
- **El poder de la selección estratégica:** El salto de la estrategia aleatoria (ej. 58.41 % para 28x28,  $k = 6$ ) al de evidencia individual (86.28 %) es grande. Esto demuestra que si los píxeles son bien elegidos, el juez es competente.
- **Vulnerabilidad al engaño:** La estrategia adversarial reduce la precisión a niveles por debajo del azar (ej. 7.85 %). Esto confirma que un atacante sin oposición puede paralizar casi por completo la capacidad del juez, haciéndolo predecir incorrectamente en más del 90 % de los casos.
- **El desafío del debate:** El objetivo del debate puede enmarcarse en este contexto: ¿puede un agente honesto, enfrentando a un adversario, elevar la linea base de la precisión del juez y acercar (o incluso superar) su rendimiento lo más posible al techo establecido por el muestreo de evidencia individual?

### 3.1.3. Impacto de la cantidad de evidencia

A continuación, analizamos un aspecto notable del juez: su capacidad de generalización. Para los siguientes experimentos, utilizamos un único modelo de juez (28x28) que fue entrenado exclusivamente con muestras aleatorias de  $k = 6$  píxeles. En esta sección, lo evaluamos con un número variable de píxeles (desde  $k = 1$  hasta  $k = 12$ ) para medir cómo generaliza su rendimiento a cantidades de evidencia para las que no fue específicamente entrenado. Esto nos permite entender si hay rendimientos decrecientes o puntos de inflexión en la utilidad de la información.

La Tabla 3.2 y la Figura 3.3 ilustran patrones divergentes:

- **Estrategia aleatoria y por evidencia individual:** Ambas se benefician consistentemente de más píxeles. La precisión del muestreo aleatorio crece de forma casi lineal, mientras que el de evidencia individual muestra rendimientos mas pequeños: solo con el primer píxel ya proporciona una precisión del 80.55 %, y los siguientes añaden mejoras marginales.
- **Estrategia adversarial:** Más evidencia apenas ayuda al juez. La precisión se mantiene por debajo del 11 % incluso con 12 píxeles. Esto indica que la estrategia es eficiente en seleccionar  $k$  píxeles que, aunque individualmente reales, colectivamente construyen una narrativa falsa que el juez no puede superar solo con más datos del mismo tipo.

### 3.1.4. Impacto de la relevancia de la evidencia (variando thr)

Finalmente, se investiga cómo la “calidad” de la evidencia, definida por el umbral de intensidad del píxel (`thr`), afecta la precisión del juez. Un `thr` más alto fuerza a la

Tab. 3.2: Sensibilidad del juez a la cantidad de píxeles revelados ( $k$ ).

Píxeles ( $k$ )	Precisión del Juez por Estrategia		
	Aleatoria	Evidencia individual	Adversarial
1	18.50 %	80.55 %	3.60 %
2	28.40 %	82.84 %	4.60 %
3	39.30 %	83.01 %	5.50 %
4	44.40 %	83.16 %	6.40 %
5	53.50 %	84.63 %	7.40 %
6	55.80 %	86.28 %	7.70 %
7	59.70 %	87.20 %	7.50 %
8	65.20 %	88.16 %	8.00 %
10	68.70 %	90.12 %	8.80 %
12	74.50 %	91.37 %	10.30 %

Nota: Experimentos realizados con el juez de 28x28, entrenado con  $k = 6$  y  $\text{thr}=0.0$ . Se evaluó usando  $N = 1000$  imágenes.

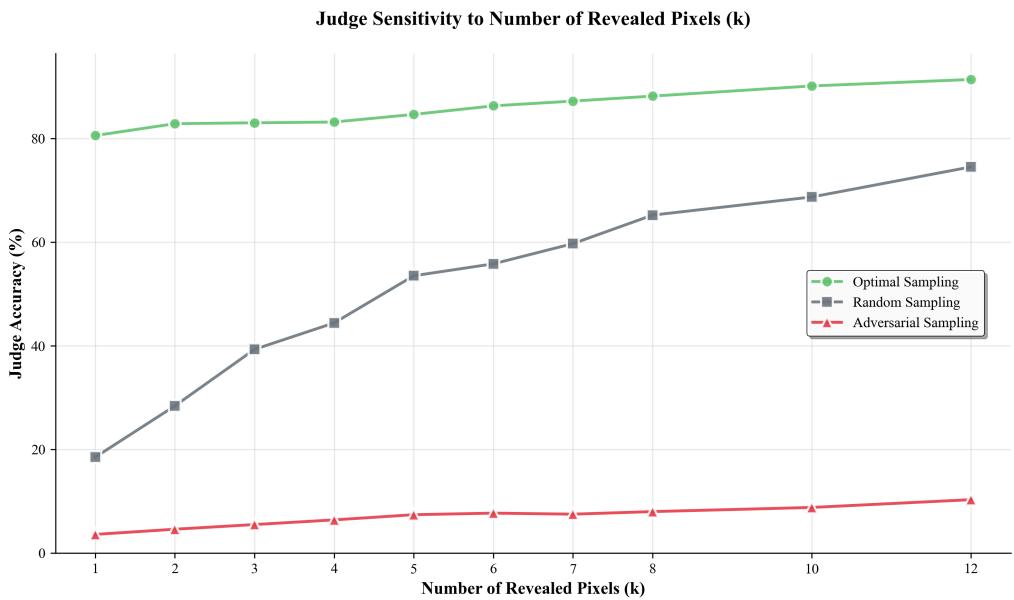


Fig. 3.3: Precisión del juez en función del número de píxeles revelados ( $k$ ) para cada estrategia de selección. Experimentos realizados con el juez de 28x28, entrenado con  $k = 6$  y  $\text{thr}=0.0$ . Se evaluó sobre  $N = 1000$  imágenes.

estrategia a considerar únicamente los píxeles más “encendidos”. El objetivo es evaluar la hipótesis de si los píxeles de mayor intensidad contienen, en promedio, información más relevante para la clasificación, reconociendo que los píxeles de intensidad media (grises) pueden ser cruciales para definir bordes.

Tab. 3.3: Sensibilidad del juez al umbral de relevancia de píxeles ( $\text{thr}$ ) para diferentes estrategias de muestreo.

Umbral (thr)	Aleatoria	Evidencia individual	Adversarial
0.0	55.80 %	85.7 %	7.7 %
0.1	59.60 %	86.3 %	7.5 %
0.2	61.00 %	86.0 %	7.6 %
0.3	62.80 %	85.8 %	7.3 %
0.4	63.20 %	85.9 %	7.6 %
0.5	64.60 %	85.7 %	7.8 %
0.6	65.30 %	85.6 %	8.0 %
0.7	66.00 %	85.8 %	8.1 %
0.8	64.10 %	86.1 %	8.5 %
0.9	65.20 %	87.3 %	9.8 %

*Nota:* Experimentos realizados con el juez de 28x28 (entrenado con  $k = 6$ ) y  $k = 6$  píxeles revelados. Se evaluó usando  $N = 1000$  imágenes para cada configuración.

Los resultados en la Tabla 3.3 revelan patrones distintos para cada estrategia. Para el muestreo aleatorio, restringir el universo de píxeles candidatos a aquellos más brillantes ( $\text{thr} > 0$ ) mejora la precisión del juez, alcanzando un máximo de 66.0 % con  $\text{thr} = 0.7$ . Este resultado sugiere que, aunque se pierda información de los bordes, un píxel aleatorio tomado de un conjunto de alta intensidad tiene en promedio más probabilidad de ser informativo que uno tomado del conjunto completo de píxeles no nulos.

El muestreo por evidencia individual demuestra un comportamiento fundamentalmente diferente, manteniendo una precisión consistentemente alta (alrededor de 85-87 %) en todos los umbrales evaluados, sin una mejora sustancial conforme  $\text{thr}$  aumenta. Esto indica que la estrategia ya es muy efectiva seleccionando píxeles informativos, incluso con umbrales bajos.

En contraste, la estrategia adversarial muestra el comportamiento opuesto: su precisión se mantiene consistentemente baja (en el rango de 7-10 %), lo que confirma su efectividad como estrategia de confusión del juez independientemente del umbral de intensidad de los píxeles elegidos.

Este análisis comparativo justifica la decisión de permitir la selección de cualquier píxel con una intensidad superior a cero, excluyendo únicamente los píxeles completamente negros del fondo para los experimentos principales de debate, ya que permite a todas las estrategias explotar el espectro completo de píxeles disponibles. Sin embargo, los resultados también sugieren que el comportamiento de las estrategias varía significativamente con diferentes umbrales de relevancia, lo que podría ser explotado en futuros trabajos para desarrollar estrategias adaptativas que ajusten dinámicamente sus criterios de selección de píxeles.

### 3.2. Introducción a las capacidades de los agentes

Habiendo caracterizado al juez en la sección anterior, el siguiente paso es evaluar las capacidades de los agentes debatientes de forma individual. En esta sección, analizamos el rendimiento de los agentes *Greedy* y *MCTS* en un escenario de “debate unilateral”, donde un único agente interactúa con el juez sin la presencia de un oponente.

Esto nos permite medir la máxima efectividad de cada agente en dos roles puros:

- **Agente honesto (rol óptimo):** El agente utiliza su estrategia (*Greedy* o *MCTS*) para seleccionar secuencialmente  $k$  píxeles con el único objetivo de que el juez prediga la etiqueta correcta. Esto mide su poder para construir un caso convincente.
- **Agente mentiroso (rol adversarial):** El agente utiliza su estrategia para seleccionar secuencialmente  $k$  píxeles con el objetivo de que el juez se equivoque, es decir, que prediga cualquier etiqueta que no sea la correcta. Esto mide su poder para engañar y ofuscar.

Los resultados de estas interacciones unilaterales se presentan en la Tabla 3.4, mostrando la precisión final del juez como medida del éxito del agente.

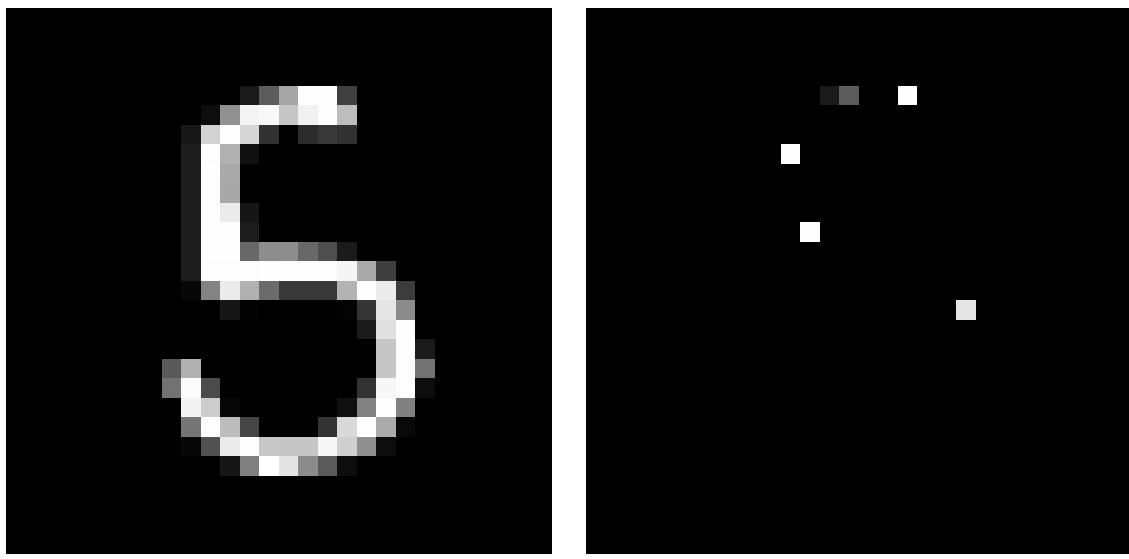
Tab. 3.4: Rendimiento de los agentes en interacción unilateral con el juez.

Píxeles ( $k$ )	Precisión juez con agente Greedy		Precisión juez con agente MCTS	
	Honesto	Adversarial	Honesto	Adversarial
1	67.60 %	4.33 %	81.00 %	4.00 %
2	85.70 %	1.67 %	100.00 %	2.00 %
3	95.80 %	1.00 %	100.00 %	2.00 %
4	98.60 %	0.67 %	100.00 %	1.00 %
5	99.10 %	1.33 %	100.00 %	1.00 %
6	99.30 %	1.00 %	100.00 %	1.00 %
7	99.70 %	1.00 %	100.00 %	1.00 %
8	99.70 %	1.00 %	100.00 %	1.00 %
10	99.70 %	1.67 %	100.00 %	1.00 %
12	99.70 %	2.00 %	100.00 %	1.00 %

*Nota:* Experimentos realizados con el juez de 28x28, entrenado con  $k = 6$  y  $\text{thr}=0.0$ . MCTS utilizó 512 rollouts.

El análisis de la Tabla 3.4 arroja conclusiones interesantes sobre las capacidades de cada agente:

- **El poder de la planificación (MCTS Honest):** El agente MCTS honesto demuestra una gran eficiencia. Alcanza una precisión perfecta del 100 % con solo 2 píxeles revelados. Su capacidad para simular le permite identificar rápidamente la combinación de evidencia más decisiva, superando con creces la heurística de “muestreo por evidencia individual” de la sección anterior, que no consideraba interacciones secuenciales.
- **Efectividad de la estrategia miope (Greedy Honest):** El agente *Greedy* honesto, aunque menos eficiente que MCTS, es extremadamente efectivo. Supera el



(a) Imagen original

(b) Selección secuencial por MCTS honesto

*Fig. 3.4:* Visualización del proceso de selección secuencial de un agente MCTS honesto. El agente elige píxeles (derecha) estratégicamente para construir un argumento convincente para el juez, una estrategia que, como se ve en la Tabla 3.4, alcanza una precisión perfecta.

95 % de precisión con solo 3 píxeles y se acerca a la perfección con 5-6 píxeles. Esto indica que incluso una estrategia cortoplacista, si es honesta, puede construir un caso muy sólido para el juez.

- **La capacidad del engaño:** Ambos agentes, cuando asumen un rol adversarial, son capaces de aniquilar la precisión del juez, llevándola a un rango de 0-4 %. Es notable que el agente MCTS mentiroso logra una precisión del 0 % con solo 4 píxeles, demostrando que la planificación no solo sirve para revelar la verdad, sino también para ocultarla.
- **Simetría en el poder:** Existe una interesante simetría en el poder de los agentes. Tanto MCTS como Greedy son casi perfectamente buenos para ayudar al juez cuando son honestos, y casi perfectamente buenos para engañarlo cuando son mentirosos. El campo de juego está preparado para un enfrentamiento donde ambos bandos tienen un alto potencial de éxito si no se les opone resistencia.

Esta sección establece firmemente las capacidades intrínsecas de nuestros jugadores. Hemos visto que tanto el agente honesto como el mentiroso son competentes en sus respectivos objetivos cuando actúan en solitario. La siguiente pregunta, y el núcleo de esta tesis, es: ¿qué sucede cuando estos agentes se enfrentan en un debate directo?

### 3.3. Resultados del debate

Tras haber caracterizado al juez y a los agentes de forma aislada, llegamos al núcleo de nuestra investigación: el enfrentamiento directo en el debate adversarial. En esta sección, analizamos qué sucede cuando un agente honesto y uno mentiroso compiten por persuadir al juez, revelando píxeles de forma alternada. Se investigan tanto escenarios simétricos,

donde los agentes poseen igual capacidad, como asimétricos, donde existe un desequilibrio de poder estratégico.

### 3.3.1. Debates simétricos: un enfrentamiento entre iguales

En los debates simétricos, un agente *Greedy* se enfrenta a otro *Greedy*, y un *MCTS* a otro *MCTS*. La métrica principal es la Tasa de Éxito del Agente Honesto, que equivale a la precisión del sistema de debate completo.

Tab. 3.5: Rendimiento en debates simétricos por tipo de agente y configuración.

Configuración del debate <i>Agente / Píxeles (k) / Inicia</i>	Tasa de éxito del agente honesto	
	Con precompromiso	Sin precompromiso
<b>Greedy</b>		
$k = 4$ / Inicia Honesto	81.40 %	48.36 %
$k = 4$ / Inicia Mentirosa	86.97 %	56.02 %
$k = 6$ / Inicia Honesto	84.81 %	62.41 %
$k = 6$ / Inicia Mentirosa	89.30 %	68.64 %
<b>MCTS</b>		
$k = 4$ / Inicia Honesto	91.00 %	54.00 %
$k = 4$ / Inicia Mentirosa	91.00 %	69.00 %
$k = 6$ / Inicia Honesto	95.00 %	71.00 %
$k = 6$ / Inicia Mentirosa	97.00 %	82.00 %

Nota: Todos los experimentos se realizaron en resolución 28x28, con el juez entrenado con  $k = 6$  (resp. 4) y  $\text{thr}=0.0$ . Los experimentos con *Greedy* usaron  $N=10,000$  imágenes; *MCTS* usó  $N=100$  y 512 rollouts. La Línea Base para esta configuración de Juez con Estrategia de Muestreo Aleatoria es 58.41 %

La Tabla 3.5 ofrece una visión clara de las dinámicas del debate entre iguales. De ella se desprenden varias conclusiones fundamentales:

- **El debate funciona:** En todas las configuraciones, la tasa de éxito del honesto supera significativamente a las líneas base aleatorias establecidas en la Sección 3.1. Por ejemplo, en los debates con  $k = 6$ , el rendimiento salta desde una línea base aleatoria del 58.41 % hasta un 85-89 % para *Greedy* y un 95-97 % para *MCTS* (con precompromiso). El proceso adversarial logra amplificar la capacidad del juez en este contexto de MNIST.
- **El precompromiso es crucial:** La diferencia de rendimiento entre los debates ‘con’ y ‘sin’ precompromiso es drástica. Por ejemplo, en el debate *MCTS* con  $k = 6$  e iniciando el mentiroso, la precisión salta de 82.00 % a 97.00 %. Forzar a los agentes a declarar y defender una postura explícita parece ser un mecanismo estabilizador muy potente que favorece la verdad.<sup>1</sup>

<sup>1</sup> Es importante contextualizar esta regla. La rigidez del precompromiso es una simplificación deliberada de este entorno experimental para forzar una confrontación directa de hipótesis. Si bien en un debate humano más general el cambio de opinión puede ser una señal de honestidad y aprendizaje, en este marco adversarial formal, el precompromiso sirve para evitar que el agente mentiroso evada la refutación cambiando de objetivo (“moviendo el arco”). Las implicaciones de esta simplificación en dominios más complejos se abordan en los capítulos siguientes.

- **La planificación otorga la victoria:** El agente MCTS supera consistentemente al agente Greedy. Con  $k = 6$  y precompromiso, MCTS alcanza un 95-97 % de éxito, mientras que Greedy se queda en un 85-89 %. La capacidad de MCTS para anticipar movimientos resulta en la presentación de evidencia más robusta y difícil de refutar.
- **Más evidencia, más claridad:** Aumentar el número de píxeles de  $k = 4$  a  $k = 6$  mejora la tasa de éxito en casi todos los casos, confirmando que, en general, más evidencia ayuda al juez a discernir la verdad, incluso en un contexto adversarial.
- **Ventaja para el segundo jugador:** Se observa una tendencia consistente en la que la tasa de éxito del agente honesto es ligeramente superior cuando el agente mentiroso realiza el primer movimiento.

Para ilustrar el funcionamiento de un debate de alta capacidad en su estado ideal, la Figura 3.5 presenta un ejemplo de un enfrentamiento MCTS vs. MCTS.

### 3.3.2. Debates asimétricos: cuando la capacidad no es la misma

Los debates asimétricos, donde un agente *Greedy* se enfrenta a uno *MCTS*, son posiblemente el escenario más interesante. Ponen a prueba la robustez del protocolo de debate ante un desequilibrio de “inteligencia” entre los contendientes. ¿Puede un honesto “tonto” ganar a un mentiroso “listo”? ¿O es la capacidad estratégica el factor decisivo?

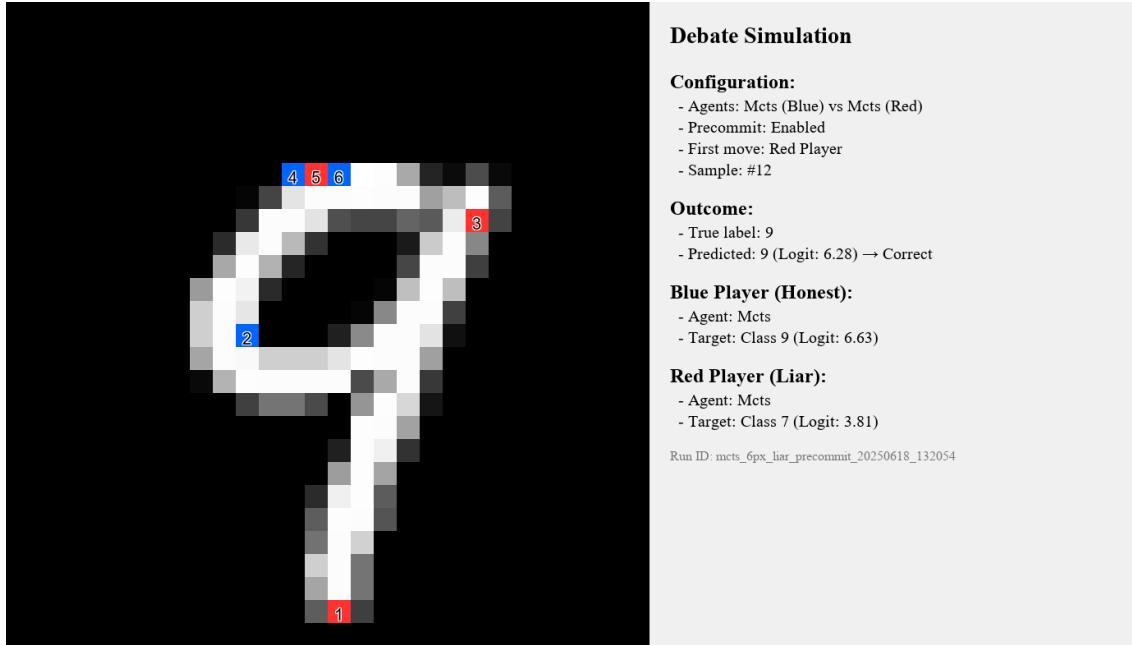
Tab. 3.6: Rendimiento en debates asimétricos (Greedy vs. MCTS).

<b>Honesto</b>	<b>Mentiroso</b>	<b>Inicia el Debate</b>	<b>Tasa de Éxito del Agente Honesto</b>	
			<b>Con precompromiso</b>	<b>Sin precompromiso</b>
<b>MCTS</b>	<b>Greedy</b>	Honesto (MCTS)	91.00 %	89.00 %
		Mentiroso (Greedy)	97.00 %	94.00 %
<b>Greedy</b>	<b>MCTS</b>	Honesto (Greedy)	88.00 %	58.00 %
		Mentiroso (MCTS)	96.00 %	63.00 %

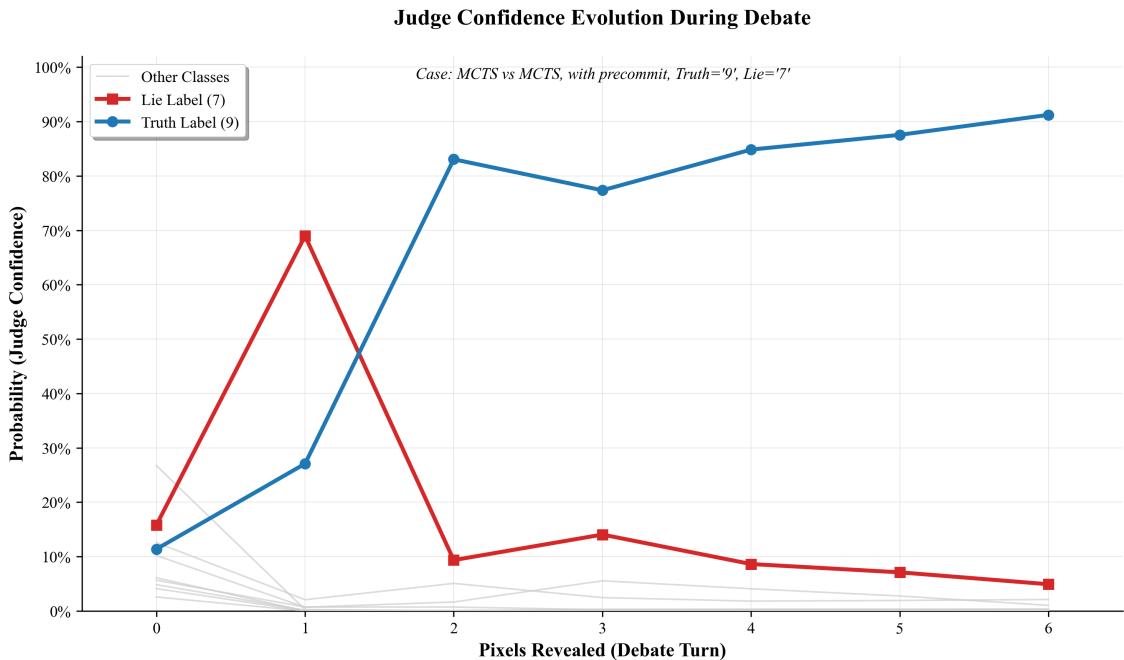
*Nota:* Todos los experimentos se realizaron en resolución 28x28, con  $k = 6$  píxeles,  $\text{thr}=0.0$  y  $N=100$  imágenes. MCTS utilizó 512 rollouts. La Línea Base para esta configuración de Juez con Estrategia de Muestreo Aleatoria es 58.41 %

Los resultados de la Tabla 3.6 son interesantes y contienen uno de los hallazgos centrales de esta investigación:

1. **Honesto fuerte vs. mentiroso débil:** Como era de esperar, cuando el agente más capaz (MCTS) es honesto, domina al mentiroso más simple (Greedy). Las tasas de éxito (91-97 % con precompromiso) son altísimas, demostrando que la superioridad estratégica se impone fácilmente para defender la verdad.
2. **Honesto débil vs. mentiroso fuerte:** Este es el escenario más crítico.
  - **Sin precompromiso,** el honesto simple es vulnerable. Con tasas de éxito de solo 58-63 %, el agente Greedy honesto apenas supera la línea base aleatoria de 58.41 % y es claramente superado por la capacidad de planificación del MCTS mentiroso.



(a) Visualización de los píxeles revelados.



(b) Evolución de la confianza del juez.

Fig. 3.5: Análisis de un debate simétrico (MCTS vs. MCTS, con precompromiso) exitoso para el honesto. En (a), se observa cómo el Agente honesto (azul, defendiendo un '9') selecciona píxeles que definen el rulo del 9, mientras el Agente mentiroso (rojo, defendiendo un '7') intenta revelar píxeles en zonas que podrían pertenecer a ambos. En (b), vemos la confianza del juez en la etiqueta verdadera (línea azul) domina desde la elección de su primer pixel, eliminando cualquier ambigüedad con lo que pretendía hacer el agente mentiroso, mientras que la confianza en la mentira (línea roja) y en otras clases (líneas grises) permanece suprimida, demostrando un debate resuelto de forma contundente.

- **Con precompromiso**, la situación cambia radicalmente. El agente Greedy honesto alcanza tasas de éxito de 88-96 %, muy por encima de la línea base.

Este último punto es de suma importancia. Sugiere que, al menos en este contexto, el **diseño del protocolo de debate puede actuar como un ecualizador de capacidades**. Un protocolo bien estructurado (en este caso, con precompromiso) puede empoderar a un agente honesto pero menos sofisticado para que prevalezca sobre un adversario más inteligente. Las siguientes figuras ilustran este dramático contraste. La Figura 3.6 muestra cómo el protocolo protege al honesto débil.

Por el contrario, la Figura 3.7 ilustra la vulnerabilidad del mismo agente honesto cuando el protocolo no lo ampara.

### 3.4. Análisis detallado de los parámetros del debate

Los resultados de la sección anterior demostraron que el protocolo de debate es efectivo en el contexto de MNIST, pero su rendimiento varía significativamente según la configuración. Para comprender en profundidad las dinámicas subyacentes, en esta sección aislamos y analizamos el impacto de tres parámetros clave del protocolo: la condición de precompromiso, el orden de los turnos y la longitud del debate ( $k$ ).

#### 3.4.1. El impacto del precompromiso

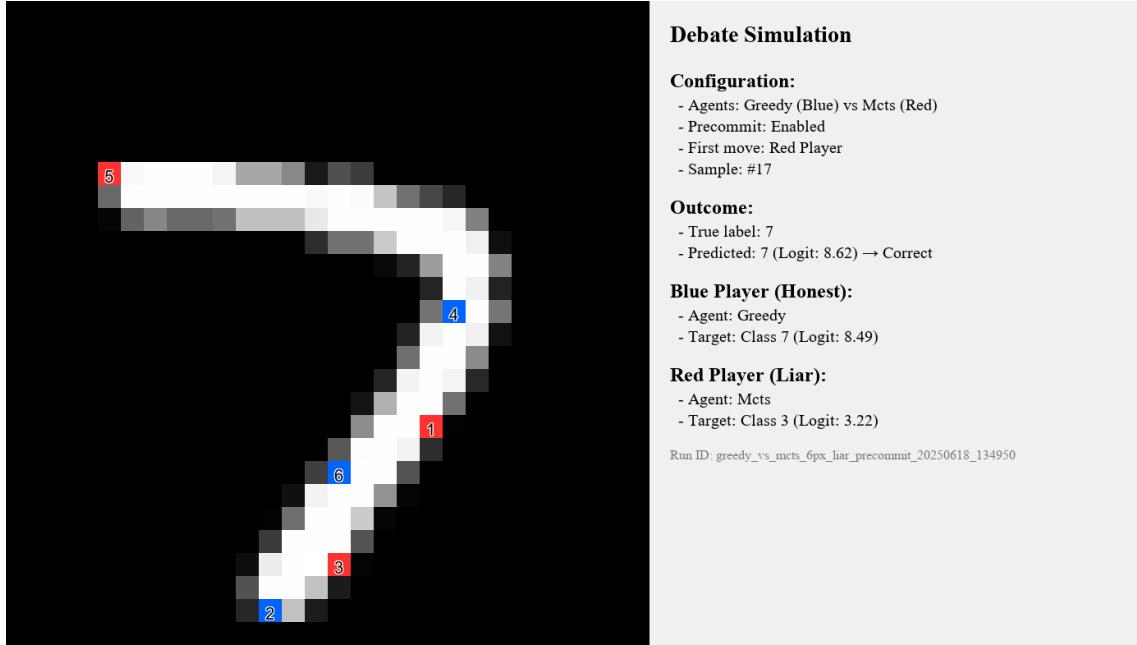
De todas las variables experimentales, la condición de precompromiso (donde los agentes declaran su postura al inicio del debate) demostró ser la más influyente. La Tabla 3.7 cuantifica la mejora en la tasa de éxito del agente honesto al activar esta condición.

Tab. 3.7: Impacto del precompromiso en la tasa de éxito del honesto (debates simétricos y asimétricos).

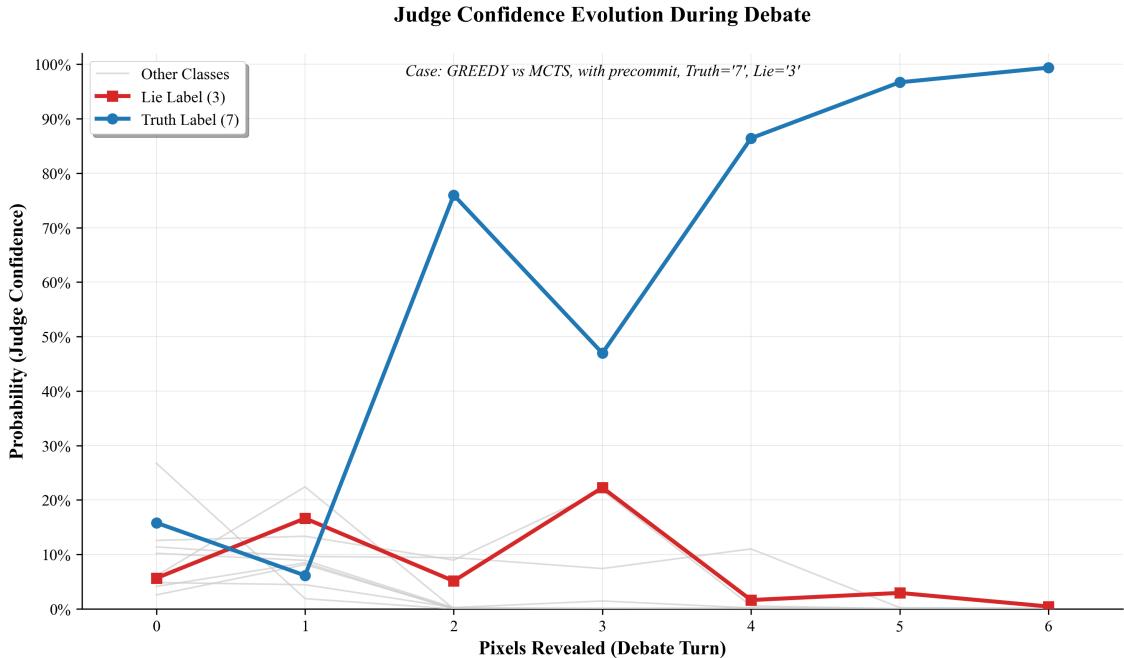
Configuración del debate (Honesto vs. Mentiroso, Inicia)	Con precompromiso	Sin precompromiso	Mejora (p.p.)
<b>Debates Simétricos</b>			
Greedy vs. Greedy, Inicia Honesto	84.81 %	62.41 %	+22.40
Greedy vs. Greedy, Inicia Mentiroso	89.30 %	68.64 %	+20.66
MCTS vs. MCTS, Inicia Honesto	95.00 %	71.00 %	+24.00
MCTS vs. MCTS, Inicia Mentiroso	97.00 %	82.00 %	+15.00
<b>Debates Asimétricos</b>			
MCTS vs. Greedy, Inicia Honesto	91.00 %	89.00 %	+2.00
MCTS vs. Greedy, Inicia Mentiroso	97.00 %	94.00 %	+3.00
Greedy vs. MCTS, Inicia Honesto	88.00 %	58.00 %	<b>+30.00</b>
Greedy vs. MCTS, Inicia Mentiroso	96.00 %	63.00 %	<b>+33.00</b>

Nota: Todos los experimentos se realizaron con  $k = 6$  píxeles. La columna “Mejora (p.p.)” muestra la diferencia en puntos porcentuales. El agente en negrita es el más fuerte (MCTS).

Como se evidencia en la Tabla 3.7 y la Figura 3.8, el precompromiso mejora drásticamente la tasa de éxito del honesto en todas las configuraciones, con ganancias que van desde 15 hasta 37 puntos porcentuales. Este efecto es particularmente vital en los debates asimétricos (ver Tabla 3.6), donde elevó la tasa de éxito de un Greedy honesto contra un MCTS mentiroso de un vulnerable 58-63 % a un robusto 88-96 %.

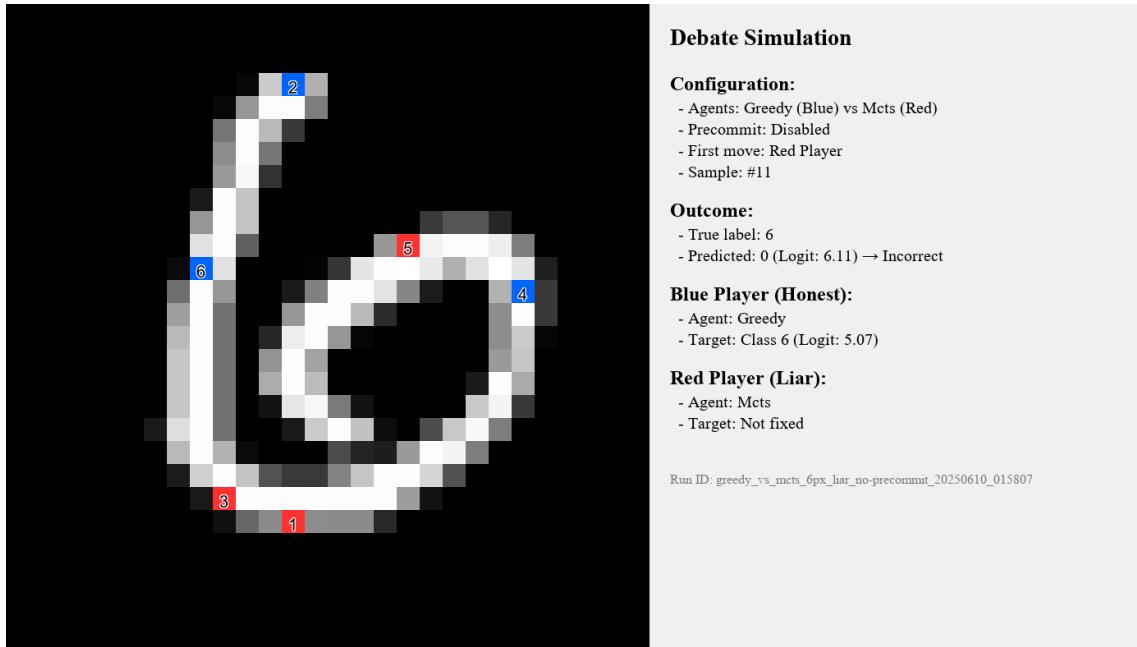


(a) Visualización del debate.



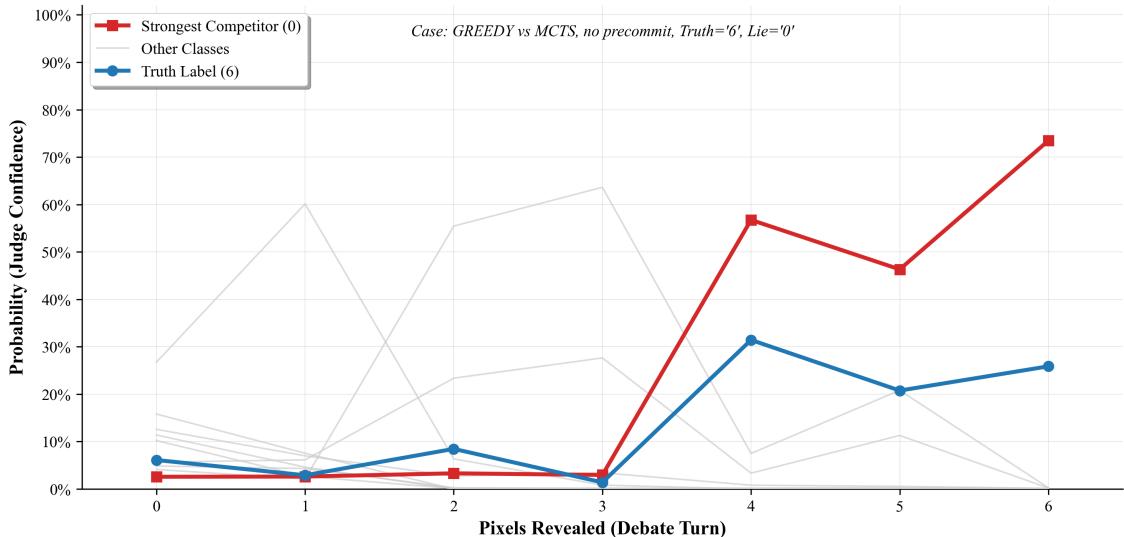
(b) Evolución de la confianza.

Fig. 3.6: El poder del protocolo (Greedy honesto vs. MCTS mentiroso, **con precompromiso**). A pesar de la desventaja de capacidad, el protocolo permite al honesto ganar. En (b), se ve cómo la confianza en la verdad (azul) logra imponerse sobre la mentira (roja).



(a) Visualización del debate.

Judge Confidence Evolution During Debate



(b) Evolución de la confianza.

Fig. 3.7: Vulnerabilidad sin protocolo (Greedy honesto vs. MCTS mentiroso, **sin precompromiso**). Sin el anclaje del precompromiso, el MCTS mentiroso logra suprimir la confianza en la verdad. En (b), la línea roja (correspondiente a la predicción final incorrecta) termina por encima de la azul. Ademas podemos apreciar como hay mas clases dominando por encima de la verdad en los demás turnos.

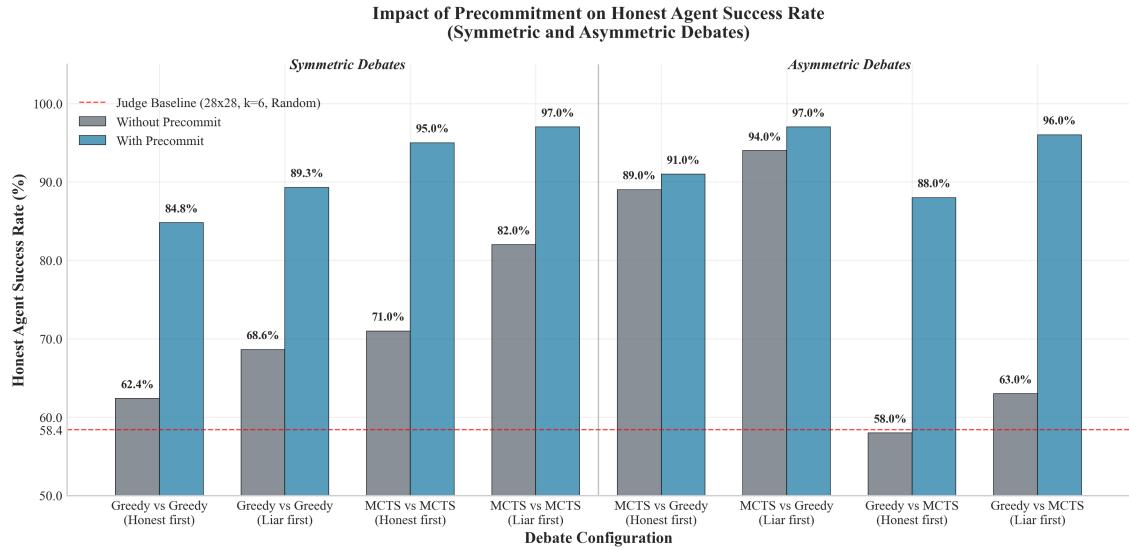


Fig. 3.8: Comparación visual del impacto del precompromiso en la tasa de éxito del honesto para diferentes configuraciones de debate.

Este hallazgo sugiere que, al menos en este contexto, el precompromiso actúa como un potente mecanismo de anclaje a la verdad. Al forzar al mentiroso a defender una falsedad específica, se le impide ‘correr el arco’ y adaptarse sobre la marcha para simplemente refutar al honesto. La confrontación se vuelve una prueba directa entre dos hipótesis, un escenario donde refutar una mentira bien definida parece ser, en efecto, más fácil que defenderla.

### 3.4.2. La ventaja del segundo jugador: ¿quién empieza?

El análisis del agente que inicia el debate revela una tendencia más sutil pero consistente. En la mayoría de los casos, la tasa de éxito del agente honesto es ligeramente superior cuando el agente mentiroso realiza el primer movimiento.

Por ejemplo, en los debates Greedy simétricos con precompromiso y  $k = 6$ , la tasa de éxito del honesto es del 84.81 % cuando él mismo inicia, pero sube al 89.30 % cuando inicia el mentiroso. Otro ejemplo clave es en el escenario de un Greedy honesto contra un MCTS mentiroso con precompromiso, la tasa de éxito del honesto es del 88.00 % cuando él mismo inicia, pero sube al 96.00 % cuando el poderoso mentiroso MCTS hace el primer movimiento. Esta diferencia de 8 puntos porcentuales es significativa.. Esta tendencia se repite en la mayoría de las configuraciones de la Tabla 3.5 y la Tabla 3.6 para resultados asimétricos.

Aunque la magnitud del efecto es pequeña (generalmente 2-5 puntos porcentuales), sugiere una ligera ventaja estratégica para el segundo jugador. Una posible explicación es que el primer movimiento revela información crucial que el segundo jugador puede utilizar para formular una refutación inmediata y más efectiva. El primer píxel revelado por el honesto, por ejemplo, podría ser el más informativo para la clase verdadera, pero también podría ser uno que el mentiroso puede explotar para presentar evidencia ambigua en su turno. Al actuar en segundo lugar, el honesto puede reaccionar a la estrategia inicial del mentiroso y contrarrestarla directamente.

### 3.4.3. El rol de la longitud del debate ( $k$ ): paridad de turnos y asimetría de poder

Finalmente, analizamos cómo la cantidad total de evidencia revelada ( $k$ ) afecta el resultado. Más allá de la intuición de que más información debería ayudar, esta variable introduce una dinámica estratégica interesante de ver: la paridad de los turnos. Con un número impar de píxeles, un jugador tiene un turno más que el otro, lo que puede crear ventajas o desventajas que no existen con un número par.

Para investigar esto, realizamos una serie de debates con el agente *Greedy* y *MCTS*, variando  $k$  de 3 a 8. Los resultados, presentados en las Tablas 3.8, 3.9 y visualizados en las Figuras 3.11, 3.12, 3.9 y 3.10, revelan un patrón oscilante superpuesto a una tendencia general de mejora.

#### Análisis en debates simétricos

Tab. 3.8: Tasa de éxito en debates Greedy y MCTS según longitud del debate ( $k$ ) y paridad de turnos.

$k$	Turno extra	Línea base	Debate Greedy		Debate MCTS	
			Precompromiso	Sin precompromiso	Precompromiso	Sin precompromiso
3	Honesto	39.30 %	95.50 %	65.60 %	98.00 %	87.00 %
	Mentiroso		46.50 %	<b>30.00 %</b>	59.00 %	<b>25.00 %</b>
4	(Igual)	44.40 %	80.70 %	48.80 %	80.00 %	55.00 %
	(Igual)		87.90 %	58.90 %	95.00 %	73.00 %
5	Honesto	53.50 %	92.10 %	71.30 %	97.00 %	89.00 %
	Mentiroso		73.50 %	<b>47.50 %</b>	76.00 %	<b>50.00 %</b>
6	(Igual)	55.80 %	82.50 %	62.10 %	90.00 %	71.00 %
	(Igual)		88.60 %	68.60 %	94.00 %	82.00 %
7	Honesto	59.70 %	91.00 %	77.10 %	97.00 %	93.00 %
	Mentiroso		79.00 %	61.10 %	82.00 %	67.00 %
8	(Igual)	65.20 %	84.60 %	71.40 %	91.00 %	84.00 %
	(Igual)		88.00 %	74.30 %	93.00 %	87.00 %

*Nota:* La primera fila de cada par de  $k$  corresponde a los debates iniciados por el honesto; la segunda, por el mentiroso. Los valores en negrita indican un rendimiento inferior a la línea base.

El análisis de estos resultados, visibles en las Figuras 3.9 y 3.10, revela dos fenómenos interrelacionados que afectan a todos los debates:

1. **Tendencia general ascendente:** A pesar de las fluctuaciones, la tendencia general es clara: a medida que  $k$  aumenta, la tasa de éxito del honesto tiende a mejorar, superando cada vez más a la línea base aleatoria. Por ejemplo, en los debates sin precompromiso iniciados por el mentiroso, la precisión pasa de un rendimiento peor que el aleatorio con  $k = 3$  (30.00 % vs 39.30 % de línea base) a uno claramente superior con  $k = 8$  (74.30 % vs 65.20 % de línea base). Esta tendencia es más pronunciada en el agente *MCTS*, que logra un rendimiento superior en casi todas las configuraciones.
2. **Fuerte efecto de la paridad y el turno extra:** Superpuesta a la tendencia general, se observa una oscilación sistemática o patrón de “diente de sierra” en ambos agentes. Cuando  $k$  es impar, el jugador que inicia el debate tiene un turno adicional. Los datos muestran que este turno extra confiere una ventaja sustancial.

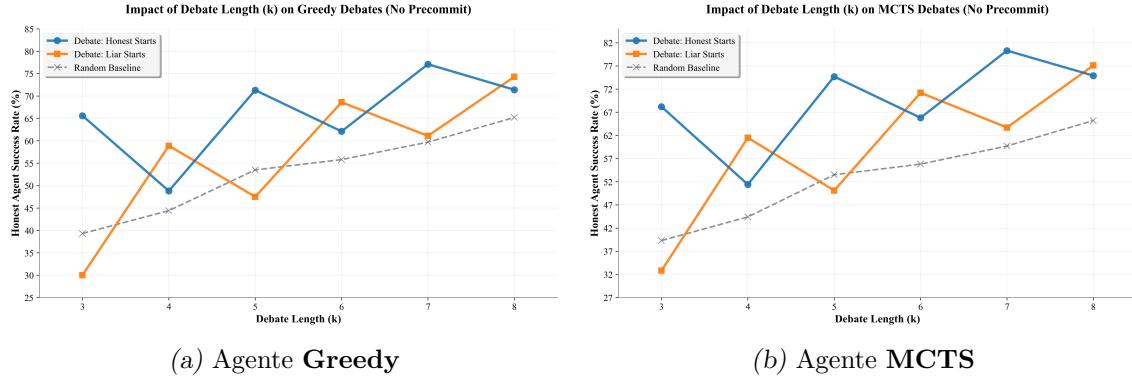


Fig. 3.9: Efecto de la longitud del debate ( $k$ ) en debates sin precompromiso. Se compara el rendimiento de los agentes Greedy y MCTS. El patrón de “diente de sierra” es pronunciado y, en los peores casos, el rendimiento cae por debajo de la línea base aleatoria.

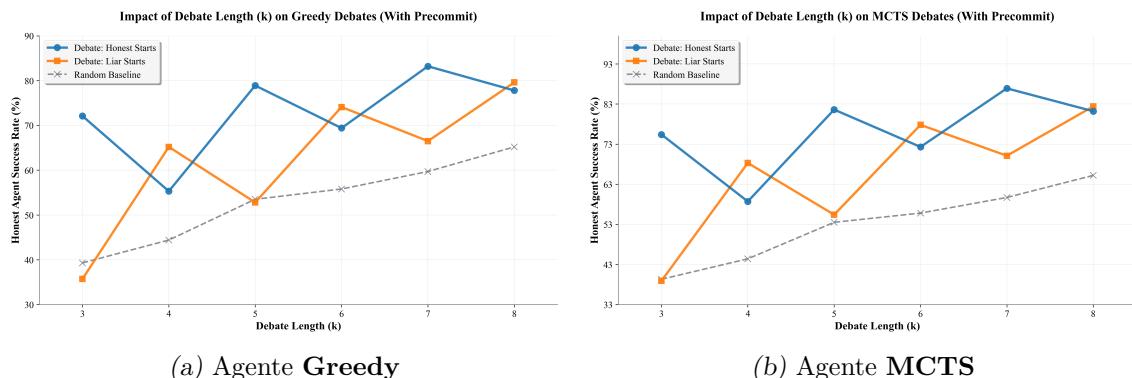


Fig. 3.10: Efecto de la longitud del debate ( $k$ ) en debates con precompromiso. Aunque el patrón de paridad de turnos persiste, el rendimiento general es significativamente más alto y apenas cae por debajo de la línea base, siendo un ejemplo visual de la robustez que añade esta regla.

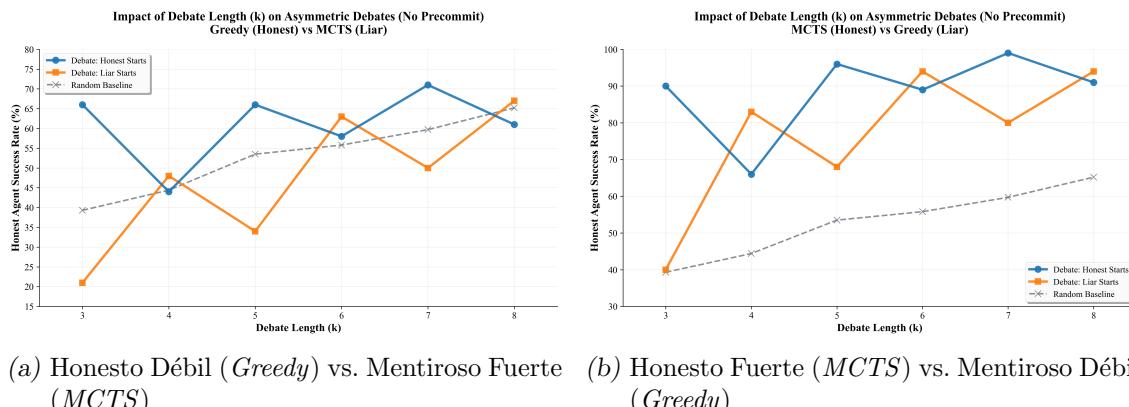
- **Sin precompromiso**, esta ventaja es crítica. Cuando el mentiroso tiene el turno extra (inicia con  $k$  impar), el rendimiento del debate a menudo cae por debajo de la línea base aleatoria (ver valores en **negrilla** en la Tabla 3.8). Esto significa que, bajo estas condiciones, el debate es activamente perjudicial.
- **Con precompromiso**, el patrón de diente de sierra persiste, pero el efecto se modera. El rendimiento general es mayor, incluso en los peores casos (cuando el mentiroso tiene el turno extra), la tasa de éxito del honesto se mantiene por encima de la línea base exceptuando el caso para el cual el mentiroso tiene 2 turnos, donde el resultado del debate está apenas por debajo de la linea base.

### Análisis en debates asimétricos.

Tab. 3.9: Tasa de éxito en debates **asimétricos** según longitud del debate ( $k$ ) y paridad de turnos.

$k$	Turno extra	Línea base	Honesto Fuerte (MCTS) vs. Mentirosa Débil (Greedy)		Honesto Débil (Greedy) vs. Mentirosa Fuerte (MCTS)	
			Con Precompromiso	Sin Precompromiso	Con Precompromiso	Sin Precompromiso
3	Honesto Mentiroso	39.30 %	99.00 % 50.00 %	90.00 % 40.00 %	99.00 % 52.00 %	66.00 % <b>21.00 %</b>
4	(Igual) (Igual)	44.40 %	85.00 % 95.00 %	66.00 % 83.00 %	81.00 % 95.00 %	<b>44.00 %</b> 48.00 %
5	Honesto Mentiroso	53.50 %	97.00 % 76.00 %	96.00 % 68.00 %	97.00 % 75.00 %	66.00 % <b>34.00 %</b>
6	(Igual) (Igual)	55.80 %	91.00 % 97.00 %	89.00 % 94.00 %	88.00 % 96.00 %	58.00 % 63.00 %
7	Honesto Mentiroso	59.70 %	97.00 % 80.00 %	99.00 % 80.00 %	96.00 % 83.00 %	71.00 % <b>50.00 %</b>
8	(Igual) (Igual)	65.20 %	89.00 % 95.00 %	91.00 % 94.00 %	92.00 % 94.00 %	<b>61.00 %</b> 67.00 %

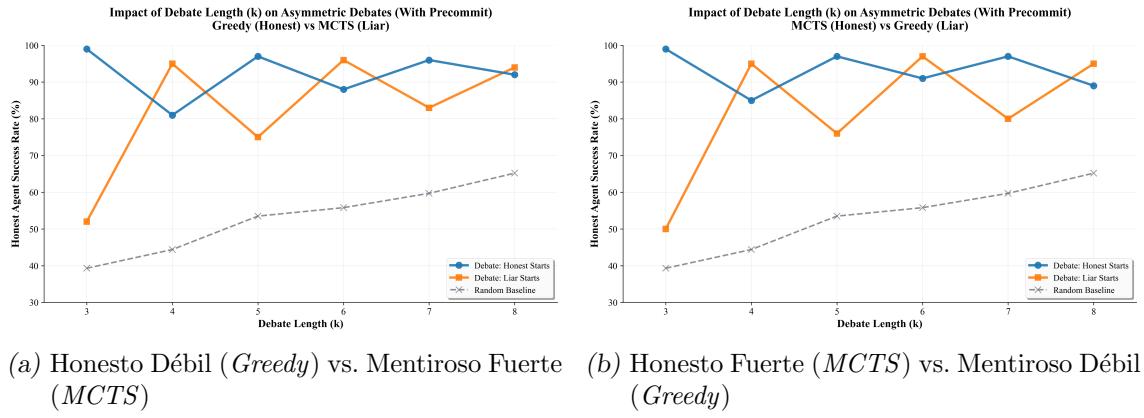
Nota: La primera fila de cada par de  $k$  corresponde a los debates iniciados por el honesto; la segunda, por el mentiroso. Los valores en **negrilla** indican un rendimiento inferior a la línea base.



(a) Honest Délil (Greedy) vs. Mentirosa Fuerte (MCTS)      (b) Honest Fuerte (MCTS) vs. Mentirosa Délil (Greedy)

Fig. 3.11: Efecto de la longitud del debate ( $k$ ) en debates asimétricos **sin precompromiso**. El gráfico de la izquierda (a) muestra el escenario más crítico, donde el agente honesto débil es vulnerable y su rendimiento cae por debajo de la línea base. A la derecha (b), el agente honesto fuerte se mantiene consistentemente por encima de la línea base.

Los resultados de los debates asimétricos, presentados en la Tabla 3.9 y visualizados en las Figuras 3.11 y 3.12, permiten analizar cómo interactúa el desequilibrio de poder con la estructura del debate. Se observan los siguientes patrones:



*Fig. 3.12:* Efecto de la longitud del debate ( $k$ ) en debates asimétricos **con precompromiso**. La introducción de esta regla eleva el rendimiento en ambos escenarios. En particular, el agente honesto débil (a) ahora se mantiene siempre por encima de la línea base, ilustrando el poder del protocolo como mecanismo ecualizador.

- Cuando el agente honesto es estratégicamente superior (**MCTS honesto vs. Greedy mentiroso**), mantiene una tasa de éxito consistentemente alta, casi siempre muy por encima de la línea base (Figura 3.11b y 3.12b). Esto sugiere que una mayor capacidad estratégica es una ventaja a la hora de defender la verdad.
- El escenario más crítico es el del **agente honesto débil (Greedy) sin precompromiso** (Figura 3.11a). Aquí se manifiesta una clara vulnerabilidad: es el único caso en todos los experimentos asimétricos donde el rendimiento cae sistemáticamente por debajo de la línea base aleatoria. Esto ocurre específicamente cuando el mentiroso fuerte (*MCTS*) tiene la ventaja del turno extra ( $k$  impar). Incluso con turnos pares ( $k = 4, 8$ ), su tasa de éxito apenas supera el rendimiento al azar, lo que indica que, en estas condiciones, el debate puede ser contraproducente.
- La introducción del **precompromiso mitiga esta vulnerabilidad de forma efectiva** (Figura 3.12a). El rendimiento del mismo agente honesto débil mejora, manteniéndose ahora en todos los casos por encima de la línea base. La regla de coherencia le proporciona una protección contra la superioridad estratégica y la ventaja estructural de su adversario.
- A nivel general, se observa que la línea correspondiente a “Liar Starts” (naranja) tiende a mostrar una mayor volatilidad en comparación con la línea de “Honest Starts” (azul), que suele ser más estable, especialmente en los debates con precompromiso.

Estos hallazgos refuerzan una de las conclusiones centrales de esta tesis: la arquitectura del protocolo, y en particular la regla de precompromiso, puede actuar como un potente mecanismo ecualizador, mitigando las desventajas tanto estratégicas como estructurales.

#### Síntesis de hallazgos sobre la longitud del debate.

En conjunto, estos resultados demuestran que la arquitectura del protocolo de debate es un factor de primer orden para potenciar la verdad. La ventaja estructural de tener ‘la última palabra’ (el último píxel revelado) es consistentemente influyente. En debates

simétricos, esta ventaja puede hacer que incluso un agente sofisticado como MCTS sea vulnerable. Sin embargo, es en los debates asimétricos donde esta vulnerabilidad se magnifica: cuando un agente mentiroso fuerte obtiene el turno extra, puede hacer que el debate sea activamente perjudicial.

La conclusión más importante es que esta debilidad estructural puede ser gestionada. La implementación de una regla robusta como el precompromiso demostró ser un mecanismo de mitigación altamente efectivo, asegurando que el debate siga siendo un proceso beneficioso incluso en presencia de asimetrías de poder y desventajas en la paridad de turnos. Esto subraya la necesidad de diseñar protocolos que no solo asuman un campo de juego nivelado, sino que activamente lo construyan.

### 3.5. El límite de la robustez: ataques fuera de distribución

En las secciones anteriores, hemos operado bajo una restricción fundamental: los agentes solo podían seleccionar píxeles que contenían alguna información ( $\text{thr} > 0$ ). Esto simula un ‘adversario realista’ que debe argumentar usando la evidencia disponible en la propia imagen.

En esta sección, llevamos a cabo una prueba de estrés para determinar los límites de la robustez del sistema. Eliminamos esta restricción y permitimos a los agentes seleccionar *cualquier* píxel, incluyendo aquellos del fondo con intensidad cero. Esto da lugar a un adversario mucho más poderoso, que denominaremos **Adversario OOD (Out-of-Distribution)**, ya que puede presentar al juez un tipo de entrada (píxeles negros seleccionados estratégicamente) para la cual no fue entrenado, atacando directamente su capacidad de generalización.

#### 3.5.1. El poder del ataque OOD

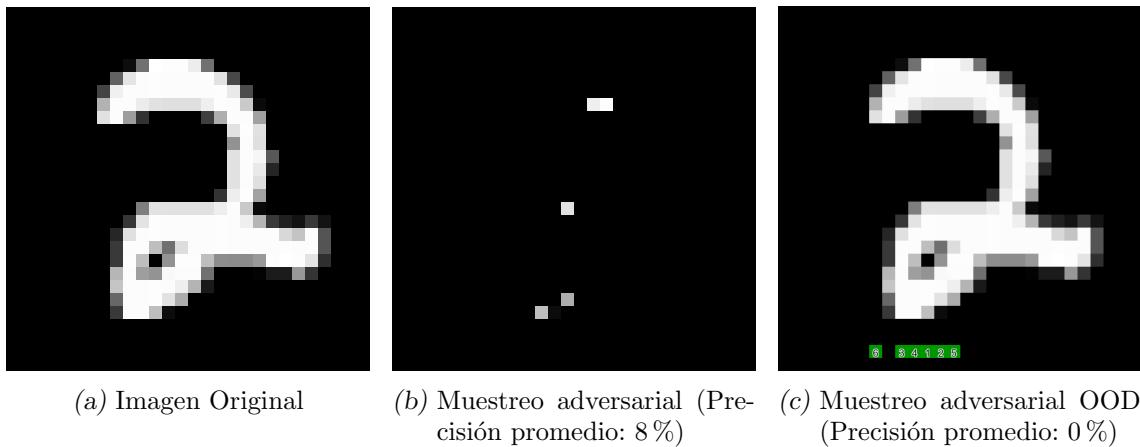
Primero, cuantificamos el impacto de este nuevo tipo de ataque en un escenario unilateral, comparándolo directamente con el Adversario Realista que hemos estudiado hasta ahora.

Tab. 3.10: Comparación de la precisión del juez bajo diferentes escenarios de muestreo y ataque.

Resolución	Nº de píxeles ( $k$ )	Precisión del juez por método de selección			
		Aleatorio	Aleatorio OOD	Adversarial	Adversarial OOD
28x28	4	46.80 %	11.06 %	4.91 %	0.00 %
	6	58.41 %	11.97 %	7.85 %	0.00 %
16x16	4	45.73 %	13.73 %	6.52 %	0.00 %
	6	56.06 %	14.56 %	9.05 %	0.00 %

*Nota:* El muestreo Aleatorio y Adversarial solo considera píxeles con intensidad  $> 0$ , mientras que los métodos OOD (Out-of-Distribution) pueden seleccionar cualquier píxel, incluyendo el fondo. Para cada fila, el juez fue entrenado con el mismo número de píxeles ( $k$ ) que se le mostraron para la evaluación. Se evaluó usando  $N = 10000$  imágenes. Usaremos nuevamente la columna Aleatorio OOD como línea base para los demás experimentos del mismo dominio.

La Tabla 3.10 presenta resultados contundentes sobre el poder del ataque fuera de distribución. Se observan dos hallazgos clave:



*Fig. 3.13:* Comparación visual de la selección Adversarial (b) contra la selección Adversarial OOD (c). Mientras que el muestreo adversarial realista se limita a seleccionar píxeles confusos del propio dígito, la estrategia de muestreo OOD explota la incapacidad del juez para interpretar píxeles del fondo, presentando evidencia fuera de distribución que anula por completo su capacidad de clasificación, en este caso, píxeles negros.

- **El espacio de acciones OOD es inherentemente más difícil para el juez.** La simple expansión del conjunto de píxeles seleccionables de manera aleatoria (columna Aleatorio OOD) ya provoca una caída drástica en la precisión, que se reduce a un rango del 11-15 %. Esto demuestra que el juez, entrenado solo con píxeles no nulos, no generaliza bien a entradas que incluyen píxeles del fondo, incluso sin un adversario.
- **El ataque adversarial OOD es absoluto.** Mientras que el adversario restringido a píxeles encendidos ( $thr > 0$ ) ya era efectivo, el adversario OOD explota esta debilidad del juez de manera perfecta, logrando reducir la precisión al 0.00 % en todos los casos. Al seleccionar estratégicamente píxeles del fondo, el adversario ataca una vulnerabilidad fundamental en el entrenamiento del juez, que nunca aprendió a interpretar la relevancia de ver “nada” en lugares específicos. La comparación con la columna Aleatorio OOD es reveladora: el ataque inteligente no solo es efectivo, sino que elimina por completo cualquier ruido útil que pudiera existir en la selección de píxeles negros.

Para confirmar la potencia de este ataque, la Tabla 3.11 muestra que este rendimiento nulo se mantiene independientemente de la estrategia del agente (Greedy o MCTS) o del número de píxeles revelados.

### 3.5.2. El colapso del debate frente a ataques OOD

Habiendo establecido que el ataque OOD es perfectamente efectivo en un escenario unilateral, la pregunta final es si el mecanismo de debate puede ofrecer alguna protección. ¿Puede un agente honesto, refutando los movimientos del mentiroso, rescatar la verdad?

Para responder a esto, repetimos los experimentos de debate simétricos y asimétricos, pero esta vez permitiendo a ambos agentes el acceso a todos los píxeles. En esta nueva configuración, el punto de comparación relevante ya no es la línea base de muestreo aleatorio restringido a píxeles con  $thr > 0$ , sino la de muestreo aleatorio OOD, que representa

Tab. 3.11: Rendimiento de agentes adversariales OOD en interacción unilateral.

Píxeles ( $k$ )	Precisión del juez con agente Adversarial OOD	
	Greedy	MCTS
1	0.00 %	0.00 %
2	0.00 %	0.00 %
3	0.00 %	0.00 %
4	0.00 %	0.00 %
5	0.00 %	0.00 %
6	0.00 %	0.00 %

Nota: Experimentos realizados con el juez de 28x28, entrenado con  $k = 6$  y  $\text{thr}=0.0$ . MCTS utilizó 512 rollouts. Se usaron  $N = 100$  imágenes.

la probabilidad de acierto del juez al observar píxeles seleccionados al azar de cualquier parte de la imagen, incluido el fondo con píxeles negros.

Tab. 3.12: Tasa de éxito del agente honesto en debates con agentes OOD.

Configuración del debate (Honesto vs. mentiroso)	Inicia	Tasa de éxito del agente honesto	
		Con precompromiso	Sin precompromiso
<b>Debates simétricos</b>			
Greedy vs. Greedy	Honesto	13.80 %	7.60 %
Greedy vs. Greedy	Mentiroso	21.90 %	13.60 %
MCTS vs. MCTS	Honesto	11.00 %	4.00 %
MCTS vs. MCTS	Mentiroso	16.00 %	4.90 %
<b>Debates asimétricos</b>			
MCTS vs. Greedy	Honesto	1.00 %	4.00 %
MCTS vs. Greedy	Mentiroso	15.00 %	21.00 %
Greedy vs. MCTS	Honesto	16.00 %	1.00 %
Greedy vs. MCTS	Mentiroso	37.00 %	12.00 %

Nota: Todos los experimentos se realizaron en resolución 28x28, con el juez entrenado con  $k = 6$ . Los experimentos con *Greedy* usaron  $N=10,000$  imágenes; *MCTS* usó  $N=100$  y 512 rollouts. El agente en negrita es el más fuerte (MCTS). La Línea Base para esta configuración de Juez, en un muestreo aleatorio OOD, es del 11.97 % (ver Tabla 3.10). El valor presentado anteriormente de 58.41 % corresponde a un muestreo restringido a píxeles no nulos.

Los resultados de la Tabla 3.12 demuestran un colapso casi total del mecanismo de debate.

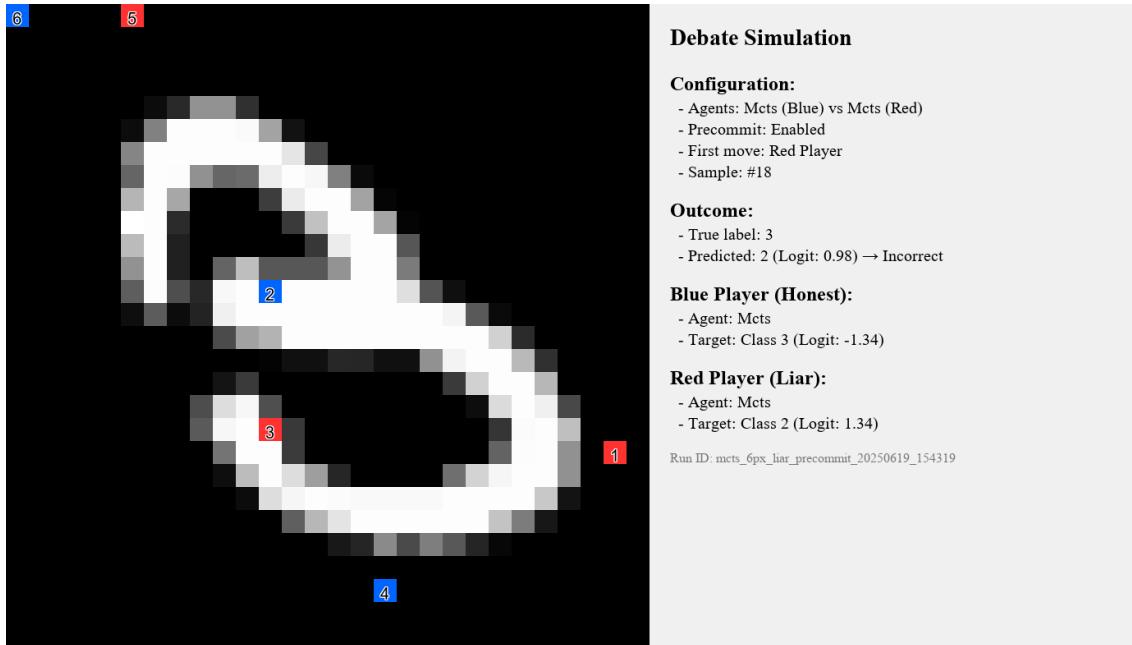
- **La Precisión se desploma por debajo del azar:** Las tasas de éxito del honesto caen a niveles bajísimos, a menudo por debajo del 25 %. Más importante aún, en algunos casos clave (e.g., MCTS vs. MCTS, o cuando el honesto es más débil), el rendimiento del debate es igual o inferior a la línea base de un muestreo aleatorio OOD (11.97 %). Esto indica que, en presencia de un ataque OOD, el mecanismo de debate no solo deja de amplificar la verdad, sino que se vuelve activamente perjudic

cial, produciendo un resultado peor que la selección de evidencia al azar en el mismo dominio.

- **Los mecanismos de protección fallan o se invierten:** El precompromiso, que era el ancla a la verdad en los debates anteriores, ahora tiene un efecto inconsistente. En algunos casos, como en el debate ‘MCTS vs. Greedy’ donde el mentiroso inicia, la tasa de éxito del honesto incluso empeora con precompromiso (de 21 % a 15 %). Esto podría no ser necesariamente una falla del protocolo, sino que puede reflejar un cambio en la estrategia del mentiroso. Sin precompromiso, el mentiroso debe buscar la confusión general. Con precompromiso, puede concentrar toda su capacidad en un único objetivo: hacer que el juez prefiera una mentira específica sobre la verdad. Con un juez ya vulnerable por los ataques OOD, esta estrategia de ataque enfocado puede volverse más efectiva paradójicamente.
- **La capacidad estratégica pierde sentido:** La ventaja del agente MCTS desaparece o se invierte. En algunos casos, el agente Greedy parece tener más éxito. Esto sugiere que cuando la vulnerabilidad explotada es tan fundamental, la planificación estratégica sofisticada deja de ser relevante; el juego se vuelve caótico.

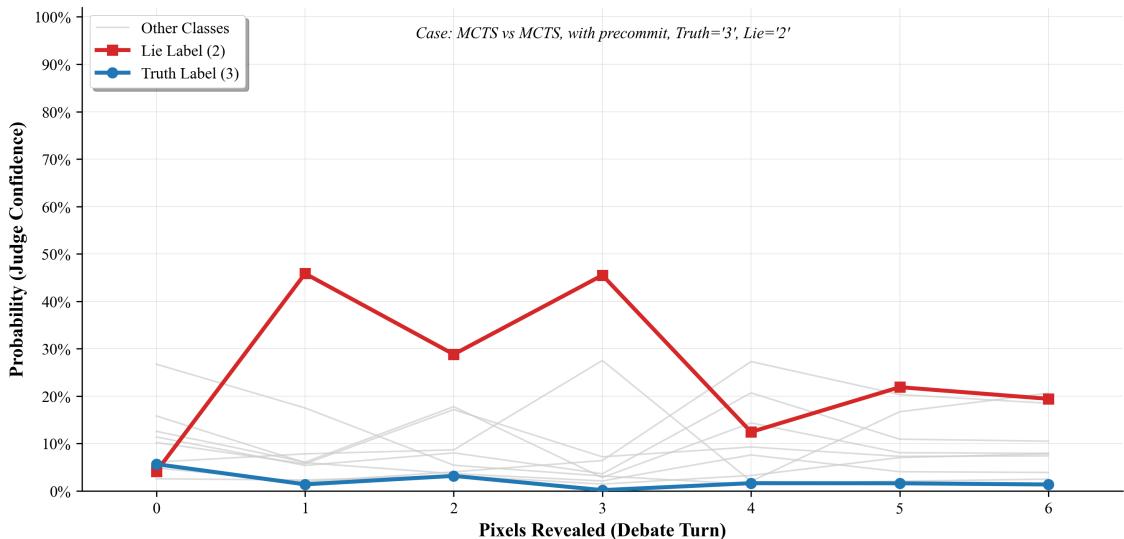
Para entender por qué el debate colapsa de forma tan catastrófica, la Figura 3.14 visualiza la dinámica interna de un debate OOD.

La conclusión de esta sección es inequívoca: la robustez del paradigma de debate aquí estudiado depende críticamente de la robustez del propio juez a entradas fuera de su distribución de entrenamiento. Si el juez tiene un ‘talón de Aquiles’ explotable, como la incapacidad de procesar píxeles de fondo, el mentiroso puede atacar esa debilidad directamente, y la estructura del debate no es suficiente para proteger la verdad. Esto subraya la importancia de entrenar jueces robustos como un prerequisito para un debate seguro.



(a) Ataque OOD en un debate.

Judge Confidence Evolution During Debate



(b) Evolución de la confianza.

Fig. 3.14: Colapso del debate bajo ataque OOD. En (a), los agentes revelan píxeles en el fondo negro. En (b), la confianza del juez es caótica: aunque la clase del mentiroso domina, múltiples líneas compiten erráticamente y la confianza final en la clase ‘ganadora’ es baja (~ 20 %, demostrando la confusión del juez).



## 4. DISCUSIÓN

Los resultados empíricos presentados en el capítulo anterior ofrecen una oportunidad para analizar en profundidad la dinámica de “AI Safety via Debate”. El trabajo seminal de Irving et al. (2018)[10] estableció la premisa fundamental de este paradigma, demostrando en un entorno MNIST con agentes simétricos que el debate podía amplificar la precisión de un clasificador débil de un 59.4 % a un 88.9 %. Nuestra investigación toma este punto de partida para explorar la robustez del mecanismo bajo condiciones más exigentes y realistas: la introducción de capacidades asimétricas entre los agentes y un análisis sistemático de cómo la arquitectura del protocolo influye en el resultado.

Este capítulo interpreta dichos hallazgos, argumentando que si bien la asimetría de capacidades introduce una vulnerabilidad real, esta puede ser gestionada mediante un diseño de protocolo cuidadoso. La conclusión que emerge es que el protocolo no es un detalle de implementación, sino que es la herramienta de alineamiento en sí misma. Siendo capaz de actuar como un factor ecualizador que protege la honestidad incluso cuando esta es estratégicamente más simple si está correctamente diseñado.

### 4.1. El debate en MNIST como analogía de la supervisión escalable

Antes de interpretar los resultados, es necesario enmarcar correctamente el entorno experimental. El sistema de debate implementado en MNIST no busca replicar la complejidad de una discusión humana, sino servir como una analogía tratable y controlable del problema de la supervisión escalable (*scalable oversight*). La elección de este entorno simplificado permite aislar y estudiar dinámicas fundamentales que se espera encontrar en sistemas más complejos.

#### 4.1.1. Contextualización del entorno experimental

En nuestro ‘juego de debate’, la argumentación no ocurre en lenguaje natural, sino a través de la selección estratégica de evidencia (píxeles). El ‘juez’ no es un humano, sino un clasificador con una capacidad de observación severamente limitada. Este diseño, inspirado directamente en el paper de Irving et al. (2018) [10], nos permite modelar a un supervisor limitado en su dominio (sabe clasificar con éxito moderado dígitos) que además no tiene la capacidad para procesar toda la información disponible y debe depender de la evidencia presentada, en este caso, en forma de máscaras de píxeles.

La principal implicación de este enfoque es que debemos ser cautos al extrapolar los resultados. El objetivo no es afirmar que “el debate funciona” en un sentido universal, sino analizar bajo qué condiciones esta analogía se sostiene o se rompe, proporcionando así información valiosa para el diseño de sistemas de debate en dominios más demandantes, como los que involucran Modelos de Lenguaje (LLMs).

#### 4.1.2. Posicionamiento frente al trabajo original y novedad del estudio

El estudio de Irving et al. (2018)[10] se centró en una prueba de concepto: demostrar que agentes **simétricos** (ambos con MCTS) podían, mediante el debate, amplificar la

capacidad del juez. Esta tesis extiende la investigación en tres direcciones clave para probar la robustez del mecanismo:

1. **Análisis de asimetría de capacidades:** Se investiga explícitamente el enfrentamiento entre agentes de distinta sofisticación (Greedy vs. MCTS). Este es un escenario realista donde un sistema podría no tener la misma capacidad computacional o estratégica que un adversario.
2. **Análisis de la arquitectura del protocolo:** El estudio del protocolo abarcó un conjunto de factores de diseño para entender mejor sus efectos. Se profundizó en el rol del precompromiso, evaluando su efectividad en el contexto de debates asimétricos. Adicionalmente, se investigaron otras variables estructurales del juego, observando una ligera pero consistente ventaja para el segundo jugador y un pronunciado efecto de paridad ligado a la longitud del debate ( $k$ ). Este análisis detallado del protocolo en la analogía de MNIST permite comprender mejor por qué la seguridad del mecanismo es sensible a la totalidad de sus reglas.
3. **Pruebas de estrés y límites de robustez:** Se introdujo una prueba de estrés explícita al permitir a los agentes seleccionar píxeles fuera de la distribución de entrenamiento del juez (ataques OOD). Esto nos permitió identificar un punto de fallo del sistema, explorando los límites de su robustez de una manera que no se abordó en el experimento original.

En resumen, la tesis se desplaza del análisis del debate en condiciones ideales hacia la investigación de su comportamiento y robustez en situaciones menos controladas o más desafiantes dentro de nuestro contexto de entorno de pruebas MNIST.

## 4.2. Interpretación de los hallazgos centrales

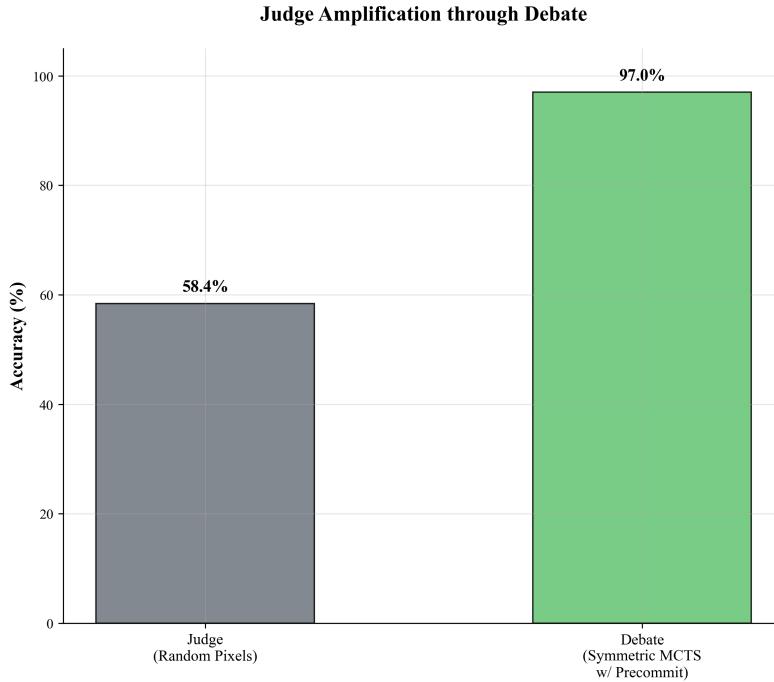
Los resultados cuantitativos y cualitativos presentados en el Capítulo 3 revelan una serie de dinámicas complejas. A continuación, se interpretan estos hallazgos para construir un argumento sobre la efectividad, las vulnerabilidades y los mecanismos de protección del debate en el contexto estudiado.

### 4.2.1. Validación de la premisa base

El primer resultado consistente en todos los experimentos presentados es que, dentro de este entorno simplificado, el debate adversarial mejora sistemáticamente la precisión del juez por encima de la línea base de muestreo aleatorio (ver Tablas 3.5 y 3.6). Este efecto de “amplificación” sugiere que la interacción estructurada de un debate, incluso con un participante adversarial, produce un conjunto de evidencia final más informativo que una selección de píxeles al azar.

La razón subyacente es el poder de la selección de información de manera intencionada. Como se demostró en los experimentos de agente unilateral (Tabla 3.4), un único agente honesto, ya sea *Greedy* o MCTS, puede guiar al juez rápidamente, con pocos píxeles revelados, a una precisión cercana al 100% al seleccionar a los más informativos estratégicamente. Aunque luego en un debate real el mentiroso intentará contrarrestar esta selección, la evidencia presentada al juez sigue siendo, en conjunto, de una calidad superior a la aleatoria. Esto apoya la idea de que la estructura del debate obliga a los

agentes a centrarse en los puntos más relevantes de la evidencia, si bien es importante destacar que esta dinámica podría no generalizarse directamente a dominios de debate más complejos donde la evidencia no es tan claramente separable.



*Fig. 4.1:* Amplificación de la capacidad del juez mediante el debate. El gráfico compara la precisión de la línea base (juez con 6 píxeles seleccionados al azar) con la tasa de éxito obtenida en un debate simétrico MCTS con precompromiso, demostrando una mejora sustancial.

#### 4.2.2. La Asimetría de capacidades como vulnerabilidad

La decisión de estudiar debates asimétricos surge como una consecuencia natural de los resultados iniciales sobre las capacidades de los agentes. Ya en un entorno sin oposición, al interactuar unilateralmente con el juez (Tabla 3.4), se observa una diferencia de eficiencia notable. Con la revelación de un solo píxel, un agente MCTS honesto ya logra una precisión del 81.00 %, superando el 67.60 % del agente *Greedy*. Esta brecha se vuelve abrumadora con el segundo píxel, donde MCTS alcanza una precisión perfecta del 100 %, mientras que el agente *Greedy* necesita al menos cinco píxeles para acercarse a ese nivel (99.10 %). La superioridad en la capacidad de planificación de MCTS se confirma en los debates simétricos (Tabla 3.5), donde, en todas las configuraciones, el debate ‘MCTS vs. MCTS’ obtiene tasas de éxito consistentemente más altas que el debate ‘*Greedy* vs. *Greedy*’.

Estos resultados establecen una clara jerarquía de capacidades estratégicas, lo que plantea una pregunta crítica para la seguridad del sistema: ¿qué sucede cuando esta diferencia de capacidad se introduce en un mismo debate? Dicho de otro modo, ¿puede un agente mentiroso más sofisticado explotar la simplicidad de un agente honesto para engañar al juez?

Los resultados del debate asimétrico sin precompromiso (Tabla 3.6) responden afirmativamente a esta pregunta, confirmando que la asimetría representa una vulnerabilidad

real. Cuando un agente *Greedy* honesto se enfrenta a un MCTS mentiroso, su tasa de éxito se desploma a un rango de 58-63 %, un rendimiento apenas marginalmente superior a la línea base de muestreo aleatorio (58.41 %). Este hallazgo es importante, pues demuestra empíricamente que en un entorno de debate con reglas permisivas, la superioridad de capacidades y por ende estratégicas pueden triunfar en mayor medida que la verdad.

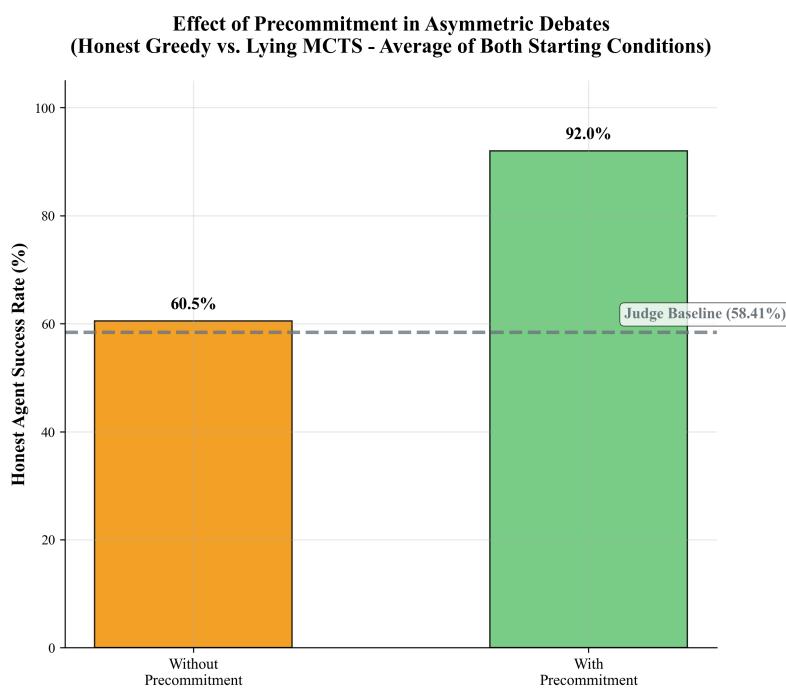
El análisis del impacto de la longitud del debate (Sección 3.4.3, Tabla 3.9) permite profundizar en esta vulnerabilidad, revelando cómo la asimetría de poder interactúa con la estructura de turnos. Se observa que la vulnerabilidad del agente honesto débil se magnifica de manera esperable bajo una combinación de condiciones adversas: enfrentarse a un mentiroso fuerte, sin la protección del precompromiso, y con una desventaja estructural en la paridad de turnos. En este escenario el rendimiento del debate cae sistemáticamente por debajo de la línea base aleatoria (e.g., 21.00 % de éxito con  $k = 3$ ), convirtiéndose en un mecanismo activamente perjudicial. Incluso cuando los turnos son pares ( $k = 4, 8$ ), el rendimiento del honesto débil en este entorno adverso sigue siendo bajo, apenas logrando superar la línea base. Esto contrasta con el escenario inverso: un honesto fuerte (MCTS) enfrentado a un mentiroso débil (*Greedy*) se mantiene consistentemente por encima de la línea base.

#### 4.2.3. El protocolo como mecanismo ecualizador

Habiendo establecido que la asimetría de capacidades representa una vulnerabilidad, el siguiente paso es analizar los mecanismos que pueden mitigarla. En este contexto, se estudió el efecto de la estructura del protocolo. El trabajo de Irving et al. (2018)[10] ya había demostrado que la condición de precompromiso mejoraba los resultados en debates simétricos. Esta investigación extiende este análisis a los debates asimétricos, donde su impacto es aún más pronunciado. Como se ilustra en la Figura 3.8 y se cuantifica en la Tabla 3.7, la introducción de esta regla tiene un efecto significativo en el rendimiento del agente honesto cuando se encuentra en desventaja estratégica.

El cambio más notable se observa al analizar el enfrentamiento ‘*Greedy Honest vs. MCTS Mentiroso*’. En este escenario, la tasa de éxito del honesto se incrementa de un rango de 58-63 %, apenas por encima de la línea base, a un robusto 88-96 % al hacer nuevamente el debate pero con precompromiso. Este fenómeno sugiere que la regla actúa como un mecanismo ecualizador. Una posible explicación para este efecto es que el precompromiso impone una carga de coherencia al mentiroso. Sin esta restricción, el agente mentiroso puede emplear una estrategia flexible de “mover el arco”, adaptando su táctica en cada turno para simplemente ofuscar o contradecir la última jugada del honesto. Al forzarlo a defender una única falsedad específica desde el inicio, se le obliga a realizar una tarea estratégicamente más demandante: construir y defender una narrativa falsa consistente frente a refutaciones directas.

Este resultado sugiere que el diseño del protocolo puede ser una herramienta tan relevante para el resultado del debate como la capacidad estratégica intrínseca de los agentes. En lugar de depender únicamente de que el agente honesto sea el más sofisticado, se puede diseñar el entorno de interacción para que las estrategias veraces sean inherentemente más fáciles de defender. La Figura 3.12 ilustra este poder ecualizador: la introducción del precompromiso eleva el rendimiento del honesto débil por encima de la línea base en todas las longitudes de debate, mitigando la combinación de desventajas (de capacidad y estructurales).



*Fig. 4.2:* El precompromiso como mecanismo ecualizador en debates asimétricos. Se muestra la tasa de éxito del agente *Greedy* honesto frente a un MCTS mentiroso, se muestra el promedio de cuando empieza el honesto y el mentiroso. Sin la regla de precompromiso, su rendimiento es apenas superior a la línea base aleatoria. Con la regla, su tasa de éxito aumenta drásticamente, demostrando el poder del protocolo para proteger al agente más débil.

#### **4.2.4. Dinámicas estructurales del debate**

Finalmente, el análisis de los parámetros del protocolo reveló que todos los detalles de la arquitectura e implementación del juego de debate tienen sus consecuencias estratégicas.

Primero, se observó una ligera pero consistente ventaja para el segundo jugador. Como se discute en la Sección 3.4.2, la tasa de éxito del honesto tendió a ser superior cuando el mentiroso realizaba el primer movimiento. Esto refleja el ‘poder de la refutación’. En este entorno de selección de píxeles podemos ver que el primer movimiento revela, de cierta forma, la estrategia inicial del oponente, funcionando como una revelación de intenciones/estrategia y permitiendo así al segundo jugador adaptar su jugada para contrarrestarla directamente. Podríamos trazar una analogía a dominios de lenguaje, donde una afirmación puede ser muy específica, la ventaja de poder enfocar una refutación podría ser aún mayor. Por el contrario, en dominios donde el primer movimiento abre un espacio de posibilidades muy amplio, esta ventaja podría disminuir o invertirse. Esta dinámica encuentra un paralelo interesante en el debate humano, donde la capacidad de responder directamente al argumento inicial de un oponente a menudo confiere una ventaja táctica.

Segundo, y de forma más drástica, el análisis del impacto de la longitud del debate ( $k$ ) en la Sección 3.4.3 reveló un fuerte efecto de paridad, visible en el patrón de “diente de sierra” de las Figuras 3.9b y 3.10. El jugador que dispone de un turno adicional (el que inicia en debates con  $k$  impar) obtiene (de manera esperable) una ventaja sustancial. De manera alarmante, cuando el mentiroso tenía esta ventaja y el protocolo era débil (sin precompromiso), el rendimiento del debate cayó por debajo de la línea base aleatoria (Tabla 3.8). Como se discutió en la sección anterior, las consecuencias de esta dinámica son aún más graves en los debates asimétricos, donde la desventaja estructural amplifica la desventaja de capacidad. Esto significa que una configuración estructural deficiente puede hacer que el debate sea activamente perjudicial. La ventaja de tener ‘la última palabra’ es un factor de primer orden, análogo a su conocida importancia en la persuasión humana, y debe ser tenido en cuenta en cualquier diseño de protocolo de debate.

#### **4.3. Límites de la robustez: cuando el debate colapsa**

Las secciones anteriores demostraron que, dentro de las reglas establecidas y para nuestra analogía, el debate puede ser un mecanismo robusto. Sin embargo, un aspecto crucial de la investigación es la prueba de estrés: ¿qué sucede cuando un adversario puede moverse fuera de las suposiciones implícitas del sistema? La Sección 3.5 exploró precisamente este escenario al eliminar la restricción sobre la selección de píxeles, permitiendo a los agentes elegir píxeles del fondo (intensidad cero).

##### **4.3.1. El ataque fuera de distribución como falla de generalización del juez**

Los experimentos descritos hasta ahora operaron bajo una restricción fundamental: los agentes solo podían seleccionar píxeles con intensidad superior a cero. Al eliminar esta restricción en la Sección 3.5, se realizó una prueba de estrés para evaluar la robustez del sistema frente a un espacio de acciones más amplio. Los resultados mostraron una caída severa en el rendimiento del debate, aunque la magnitud de este colapso requiere una interpretación matizada. Si se compara con la línea base del muestreo aleatorio estándar

(restringido a píxeles no nulos), la tasa de éxito del agente honesto en el escenario OOD es drásticamente inferior en todas las configuraciones.

Sin embargo, la comparación más rigurosa es frente a la línea base de muestreo aleatorio OOD. En este caso, los resultados son inconsistentes: en algunas configuraciones, el debate logra un rendimiento ligeramente superior al azar, mientras que en otras, colapsa a niveles inferiores, volviéndose activamente perjudicial. El hallazgo más significativo no es solo la caída en el rendimiento absoluto, sino la pérdida de las dinámicas estratégicas coherentes: la ventaja de MCTS sobre Greedy y el efecto protector del precompromiso se anulan o incluso se invierten. Esto indica que el mecanismo de debate deja de ser un amplificador fiable de la verdad y se convierte en un juego impredecible y caótico.

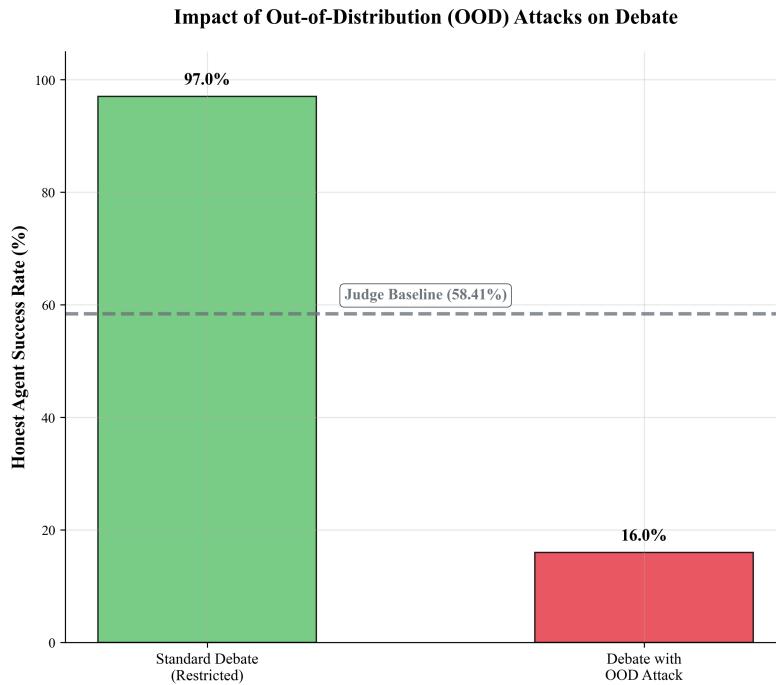
Es fundamental interpretar correctamente la causa de este colapso. El mentiroso no gana a través de una refutación directa de la evidencia presentada por el honesto, que era la dinámica principal en los debates estándar. Su victoria proviene de una estrategia diferente: encontrar y explotar una vulnerabilidad fundamental en el propio juez. Al presentarle evidencia para la cual no fue entrenado (píxeles negros), el agente fuerza al clasificador a operar fuera de su distribución de entrenamiento (OOD). El juez, que aprendió a clasificar un dígito a partir de la presencia de (6) píxeles ‘encendidos’ ( $thr > 0$ ), no posee un modelo robusto para interpretar la selección intencionada de píxeles nulos.

Este fenómeno es análogo a un ataque adversarial en clasificación de imágenes, donde una perturbación, a menudo imperceptible para un humano, puede cambiar drásticamente la predicción de un modelo. En nuestro caso, el ‘ataque’ consiste en presentar un tipo de evidencia que expone una falla en la capacidad de generalización del juez. Una vez que esta vulnerabilidad está disponible, la dinámica del debate se transforma por completo. La estrategia óptima para el mentiroso ya no es refutar al honesto, sino atacar directamente la debilidad del árbitro. Es revelador que, ante esta situación, el agente honesto también comienza a seleccionar píxeles nulos en sus turnos. Esto indica que su propia estrategia de maximización de recompensa determina que la mejor forma de defender la verdad no es presentar evidencia positiva, sino intentar contrarrestar o mitigar el ataque del mentiroso en su propio terreno. Sin embargo, como muestran los resultados, esta defensa a menudo resulta infructuosa.<sup>1</sup> El debate degenera de una competencia por presentar la mejor evidencia a una lucha por explotar o defenderse de una falla del juez, un escenario donde la verdad se pierde. El gráfico de confianza de la Figura 3.14 ilustra este comportamiento errático, que persiste incluso con el anclaje del precompromiso.

#### 4.3.2. La robustez del juez como condición límite del debate

Los resultados del ataque fuera de distribución ilustran una conclusión clave: la efectividad del mecanismo de debate está directamente limitada por la robustez del propio juez. El paradigma del debate se apoya en la suposición de que el juez, aunque limitado en su capacidad de observación, es competente para evaluar la evidencia que se le presenta. Nuestros resultados demuestran que si esta suposición es violada (en este caso, mediante una entrada que explota una falla en la generalización del juez), el sistema de debate pierde

<sup>1</sup> Es crucial señalar que esta dinámica de explotación es posible porque los agentes tienen acceso de oráculo al modelo del juez. Como se discutirá en la sección de Trabajo Futuro, una extensión natural y necesaria de esta investigación sería explorar escenarios donde los agentes no tienen este conocimiento perfecto. En tal caso, tendrían que construir un modelo interno de las preferencias y vulnerabilidades del juez a partir de interacciones limitadas, introduciendo una capa de incertidumbre que podría mitigar la efectividad de ataques tan directos.



*Fig. 4.3:* Degradación del rendimiento del debate bajo ataques OOD. El gráfico compara la tasa de éxito de un debate MCTS simétrico en condiciones normales frente a un escenario donde los agentes pueden seleccionar píxeles de fondo (OOD). La efectividad del debate se anula casi por completo, lo que subraya la dependencia del mecanismo en la robustez del juez.

su capacidad de amplificar la verdad. No se puede esperar que la estructura adversarial compense por un componente de juicio que es fundamentalmente explotable.

Esto lleva a una reflexión importante sobre la aplicación del debate en escenarios más desafiantes. Si se utiliza un modelo de recompensa como juez proxy para un debate entre LLMs, podría parecer que se cae en una circularidad: para usar el debate como método de alineamiento, se necesitaría un modelo de recompensa ya alineado y robusto. Sin embargo, la implicación no es que se deba resolver el problema de la robustez por completo antes de poder usar el debate.

El debate es una herramienta diseñada para un propósito específico: amplificar la capacidad de un supervisor para juzgar tareas complejas (supervisión escalable). Lo que estos resultados demuestran es que esta herramienta es más efectiva cuando el juez es resistente a ataques directos. Por lo tanto, el desarrollo de técnicas para mejorar la robustez de los modelos de juez (por ejemplo, mediante entrenamiento adversarial o detección de entradas OOD) son una línea de investigación importante y complementaria.

#### 4.4. Generalización a dominios complejos y jueces humanos

Si bien el entorno MNIST ha permitido aislar dinámicas fundamentales, es crucial reflexionar sobre cómo estos hallazgos podrían trasladarse a dominios más complejos, como el debate en lenguaje natural con jueces humanos. Varias de las conclusiones obtenidas podrían cambiar o adquirir nuevos matices.

*La Robustez del Precompromiso.* En este trabajo, el precompromiso demostró ser un mecanismo ecualizador muy efectivo. Sin embargo, en un debate con lenguaje natural, su implementación es menos directa. Un agente mentiroso podría cumplir con la semántica estricta de su compromiso inicial, mientras viola su intención pragmática. Podría utilizar el lenguaje para redefinir términos, introducir contextos engañosos o desviar la atención hacia detalles irrelevantes que formalmente no rompen el compromiso, pero que efectivamente anulan su propósito. La robustez de esta regla dependería críticamente de la capacidad del juez para detectar tales sutilezas, una tarea mucho más exigente que la simple comparación de logits.

*Asimetría extrema y límites del debate.* Este estudio exploró una asimetría de capacidades entre agentes *Greedy* y MCTS. En escenarios realistas, la asimetría entre un sistema de IA avanzado y un supervisor humano podría ser órdenes de magnitud mayor. Como sugiere la vida real, es plausible que, ante una asimetría de capacidades suficientemente grande, el debate deje de ser un amplificador de la verdad. Un agente “super-persuasivo” podría construir narrativas falsas tan coherentes y densas en información (verídica pero engañosamente seleccionada) que un juez humano, con sus limitaciones cognitivas y sesgos, sería incapaz de detectar la falacia, incluso con la ayuda de un agente honesto menos elocuente. El debate, por tanto, no es una panacea, y probablemente tiene un “punto de quiebre” donde la asimetría se vuelve insuperable.

*Vulnerabilidades de un juez humano.* Más allá de las dinámicas estratégicas, la transición a un juez humano introduce una limitación práctica fundamental: el costo y la falta de escalabilidad. A diferencia de un juez algorítmico que puede evaluar debates con un costo marginal bajo, cada debate juzgado por un humano requiere tiempo, atención y recursos económicos significativos. Esta restricción inherente limita drásticamente el volumen de supervisión posible y subraya la necesidad de que los mecanismos de debate sean no solo efectivos, sino también eficientes.

Adicionalmente si el juez es humano, el espacio de ataques se expande considerablemente más allá de las fallas de generalización técnica. Un agente mentiroso ya no atacaría la incapacidad del juez para procesar píxeles negros, sino sus vulnerabilidades inherentes:

- **Sesgos cognitivos:** Explotación del sesgo de confirmación, anclaje o el uso de afecciones emocionales sobre la lógica.
- **Uso de la complejidad como barrera:** Un agente podría presentar argumentos deliberadamente complejos o llenos de jerga, no porque sean correctos, sino para intimidar al juez o hacerlo sentir incompetente para evaluarlos. Ante esta complejidad, un juez humano podría optar por no juzgar ese punto o, peor aún, darle la razón al agente que parece “más experto”, una vulnerabilidad que un juez algorítmico no tendría.
- **Explotación de la confianza:** Un agente podría construir una reputación de honestidad a lo largo de varios debates para luego usar esa confianza ganada para colar una mentira en un momento oportuno.

En resumen, la transición a jueces humanos cambia la naturaleza del ataque: de explotar las fronteras de una distribución de entrenamiento a explotar la psicología humana, todo ello bajo restricciones de escalabilidad.

#### 4.5. Implicaciones para AI Safety y alineamiento

Los hallazgos de este estudio, aunque circunscritos al entorno de MNIST, ofrecen varias reflexiones importantes para el campo más amplio de AI Safety y, en particular, para el problema del alineamiento de sistemas de IA avanzados. La transición del entorno simplificado presentado a conclusiones generales debe hacerse con cautela, pero las dinámicas observadas y presentadas en este sistema pueden ser vistas como manifestaciones de principios más fundamentales.

*El debate como herramienta práctica de Salable Oversight.* El problema central de la supervisión escalable (*scalable oversight*) es cómo los humanos (o sus proxies) pueden supervisar eficazmente a sistemas de IA que los superan en capacidad. Los resultados presentados apoyan empíricamente la idea de que el debate es un candidato viable para esta tarea. Al forzar una competencia para convencer a un juez limitado, el debate transforma un problema de *generación* de respuestas (que puede ser muy difícil para el supervisor) en un problema de *juicio comparativo* (más sencillo). El éxito del debate en este entorno, incluso con un juez que solo ve una fracción de la evidencia, sugiere que el principio de amplificación es prometedor.

*El diseño del protocolo como mecanismo de control.* Una de las principales preocupaciones en el control de la IA es cómo restringir el comportamiento de agentes poderosos sin disminuir excesivamente su capacidad. Los hallazgos presentados sobre el rol del precompromiso y la estructura de turnos sugieren que el diseño del protocolo es un mecanismo de control en sí mismo. Un protocolo bien diseñado no necesita entender el funcionamiento interno de los agentes; en su lugar, estructura el entorno de incentivos para que la honestidad sea la estrategia más rentable. Esto se alinea con la idea de que el control de la IA no se logrará únicamente mediante el diseño de agentes ‘moralmente buenos’, sino también mediante la creación de sistemas y entornos que hagan que el comportamiento deseado sea un equilibrio estratégico. En este sentido, el protocolo de debate puede ser visto como una forma de ‘ingeniería de mecanismos’ aplicada directamente al problema del alineamiento, es decir como una capa de seguridad externa, que incentiva la veracidad al nivel de las reglas del juego, independientemente de las motivaciones internas de los agentes.

*La asimetría como un desafío central para el alineamiento.* El problema de la asimetría de capacidades es una preocupación central en AI Safety. A medida que los modelos se vuelvan más potentes, es probable que surjan desequilibrios entre diferentes sistemas, o entre un sistema de IA y sus supervisores humanos. Esta investigación demuestra que esta asimetría es una vulnerabilidad real que puede ser explotada. Sin embargo, también sugiere que, mediante el correcto diseño del protocolo, no es un obstáculo insalvable. El hecho de que un protocolo robusto pueda proteger a un agente honesto pero más débil es una señal optimista. Indica que el alineamiento no es necesariamente una carrera armamentista donde solo el agente más ‘inteligente’ gana, sino que las reglas del juego pueden diseñarse para favorecer la veracidad de los argumentos. Esta dinámica, sin embargo, debe ser interpretada a la luz de las propiedades del dominio experimental. La protección que ofrece el protocolo es efectiva en MNIST, un contexto donde el coste de un error de clasificación es bajo y uniforme. Esto subraya una distinción clave para aplicaciones del mundo real: en dominios de alto riesgo, donde un único fallo adversarialmente elegido por un agente

superior podría ser catastrófico, el mismo protocolo podría no ser suficiente. Por lo tanto, estos resultados no sugieren que el diseño del protocolo sea una solución universal a la asimetría, sino que demuestran que es un mecanismo de control efectivo, cuya suficiencia debe evaluarse en función del perfil de riesgo de cada tarea.

#### 4.6. Limitaciones del estudio

Es fundamental reconocer las limitaciones inherentes al presente estudio para definir el alcance de sus conclusiones y señalar direcciones para futuras investigaciones.

- **Dominio experimental como analogía:** Si bien MNIST es un banco de pruebas útil que permite la replicación y el control, representa una simplificación considerable de los dominios complejos y de alta dimensionalidad (e.g., lenguaje natural, planificación estratégica) donde se buscaría aplicar el debate. La naturaleza de la “evidencia” (píxeles) y la “argumentación” (selección) es inherentemente restringida.
- **Naturaleza simplificada del engaño:** En este estudio, una “mentira” consiste en defender una etiqueta categóricamente incorrecta. Esto no captura la sutileza del engaño en dominios del mundo real, que puede incluir omisiones, verdades a medias, extrapolaciones engañosas o manipulación retórica.
- **Agentes con políticas estratégicas fijas:** Los agentes *Greedy* y MCTS, aunque útiles para modelar diferentes niveles de capacidad, emplean políticas fijas. No aprenden ni adaptan sus estrategias a lo largo de múltiples debates. Por lo tanto, este trabajo no explora la dinámica co-evolutiva que surgiría en un sistema entrenado mediante autoaprendizaje (*self-play*).
- **Acceso de oráculo al modelo del juez:** Los agentes, particularmente MCTS, utilizan el modelo del juez como un “oráculo” perfecto durante sus simulaciones internas. Este acceso directo al modelo de evaluación no estaría disponible en un escenario con un juez humano o con un modelo de recompensa opaco, lo que introduce una capa de incertidumbre que este estudio no aborda.
- **Escala de los experimentos:** Debido a limitaciones computacionales, algunos experimentos, especialmente aquellos con el agente MCTS, se realizaron con un número de muestras o ‘rollouts’ que, si bien son suficientes para identificar tendencias claras, podrían beneficiarse de una escala mayor para refinar las estimaciones de rendimiento y la significancia estadística.



## 5. CONCLUSIONES Y TRABAJO FUTURO

La presente tesis se propuso investigar la robustez y la dinámica del paradigma de “AI Safety via Debate” bajo condiciones que se desvían de los supuestos teóricos ideales, introduciendo jueces con capacidades limitadas y, de manera crucial, asimetrías estratégicas entre los agentes debatientes. A través de una serie de experimentos controlados en el dominio de MNIST, este trabajo ha proporcionado evidencia empírica para responder a interrogantes fundamentales sobre la viabilidad de este enfoque.

En este capítulo final, se sintetizan las conclusiones principales derivadas de los resultados obtenidos, se ofrecen respuestas directas a las preguntas de investigación que guiaron el estudio y, finalmente, se proponen líneas de investigación futuras para continuar explorando y fortaleciendo el paradigma de debate como herramienta para la supervisión segura de la inteligencia artificial.

### 5.1. Conclusiones

El análisis de los resultados experimentales, interpretados a través de la discusión del capítulo anterior, permite extraer varias conclusiones clave sobre la efectividad y las sensibilidades del debate en el entorno estudiado:

1. **En el contexto de MNIST, el debate demuestra ser un mecanismo eficaz para amplificar la capacidad de un supervisor débil.** Se ha observado de manera consistente que el proceso de debate adversarial permite a un juez con acceso muy limitado a la evidencia alcanzar una precisión significativamente mayor que la que obtendría con un muestreo de información aleatoria. La selección estratégica de evidencia por parte de los agentes, incluso por parte de un agente *Greedy* simple, es suficiente para guiar al juez hacia la verdad con una frecuencia muy superior al azar.
2. **La asimetría de capacidades introduce una vulnerabilidad real.** El estudio de debates entre agentes de distinta sofisticación estratégica (Greedy y MCTS) muestra que, en ausencia de un protocolo robusto, un agente mentiroso más capaz puede explotar la simplicidad de un agente honesto para engañar, y por ende, inducir a error al juez.
3. **El diseño del protocolo puede mitigar las vulnerabilidades de la asimetría.** Los resultados de este estudio indican que la arquitectura del protocolo de debate es un factor clave para gestionar la vulnerabilidad introducida por la asimetría de capacidades. La implementación de reglas como el precompromiso va más allá de una simple mejora en el rendimiento general; actúa como un mecanismo ecualizador que compensa la desventaja estratégica del agente honesto. Esto se observa en cómo un agente *Greedy*, destinado a perder contra un MCTS mentiroso en un protocolo permisivo, logra prevalecer cuando las reglas del juego le imponen una carga de coherencia a su adversario. Sin embargo, es crucial contextualizar este hallazgo. La protección que ofrece el protocolo fue efectiva en el dominio de MNIST, donde el coste de un error es bajo y uniforme, en otros contextos, esto podría no ser así. La pregunta sobre si este tipo de reglas serían suficientes para proteger a un agente

honesto en dominios de alto riesgo, donde un adversario superior podría necesitar encontrar un único fallo catastrófico, queda como una cuestión abierta y planteada para investigaciones futuras.

4. **La seguridad del mecanismo de debate es inseparable de la robustez del componente de juicio.** Los experimentos con ataques fuera de distribución (OOD) revelan que la efectividad del debate se degrada severamente si el juez tiene vulnerabilidades explotables. Esto sugiere que el debate no puede, por sí solo, compensar las fallas fundamentales en el componente encargado de la evaluación.

En síntesis, este trabajo extendió el análogo experimental de MNIST propuesto por Irving et al. (2018)[10] para investigar su robustez bajo condiciones que, si bien se mantienen en un dominio simplificado, introducen exigencias no exploradas en el experimento original, como la asimetría de capacidades y ataques fuera de la distribución de entrenamiento del juez.

La conclusión fundamental que emerge de este entorno controlado no es una validación universal del debate, sino la identificación de dos condiciones importantes para su robustez. Primero, la seguridad del mecanismo depende de manera determinante de la arquitectura del protocolo que gobierna la interacción, la cual puede ser diseñada para proteger la honestidad incluso frente a desequilibrios de poder. Pero con el matiz clave de que la efectividad de una regla específica, como el precompromiso, no es una garantía universal, sino que su éxito depende del perfil y la naturaleza del dominio. Segundo, dicha seguridad está intrínsecamente ligada a la robustez del propio juez frente a entradas fuera de su distribución de entrenamiento, una vulnerabilidad que, como se vio en los ataques OOD, el protocolo por sí solo no pudo mitigar.

## 5.2. Respuestas a las preguntas de investigación

Habiendo sintetizado las conclusiones generales, esta sección ofrece respuestas directas a las preguntas de investigación que guiaron este estudio, basadas en la evidencia empírica recolectada.

- *¿Cuán robusto es el protocolo de debate cuando el juez posee capacidades significativamente inferiores a las de los agentes?*

**Respuesta:** En el entorno de MNIST, el protocolo de debate demostró una robustez condicional. Por un lado, fue un mecanismo de amplificación eficaz, elevando la precisión del juez débil muy por encima de la línea base de muestreo aleatorio. Por otro lado, su efectividad se ve atada a la robustez del propio juez. Los experimentos con ataques fuera de distribución (Sección 3.5) sugieren que si un adversario puede explotar una falla de generalización en el juez, el mecanismo de debate colapsa y su capacidad protectora se anula.

- *¿Puede un agente mentiroso manipular efectivamente a un juez débil para que elija una conclusión falsa?*

**Respuesta:** Sí, la manipulación es posible, pero su éxito está condicionado por las reglas del protocolo. Un agente mentiroso estratégicamente superior (MCTS) fue capaz de superar a un honesto más simple (*Greedy*) cuando el protocolo era permisivo (sin precompromiso). Sin embargo, la misma estrategia de manipulación

fue en gran medida neutralizada al introducir una regla de precompromiso, que forzó al mentiroso a una tarea de argumentación más difícil, resultando en mas debates ganados por el agente honesto.

- *¿Cómo influye la cantidad de información revelada ( $k$ ) en la capacidad del juez y en la efectividad de las estrategias?*

**Respuesta:** Como era de esperar, la estructura de turnos del debate tiene un impacto directo en el resultado: el agente que dispone de más movimientos (en debates con  $k$  impar) obtiene una ventaja estratégica sustancial. A pesar de este efecto estructural, la tendencia general en nuestros experimentos muestra que una mayor cantidad de evidencia ( $k$ ) conduce a un debate más preciso. La verdadera cuestión, sin embargo, es por qué esto ocurre en nuestro contexto. La dinámica observada es característica de un dominio ‘convergente’ como MNIST, donde cada argumento (píxel revelado) sirve para reducir el espacio de hipótesis y eliminar ambigüedad. Esta propiedad no es universal. En dominios inherentemente ‘divergentes’ (e.g., debate legal o ético), más información podría introducir matices, contraargumentos y complejidad, aumentando la confusión. En tales escenarios, la dinámica podría invertirse, y un mayor volumen de información podría ser explotado por un agente más sofisticado para ofuscar en lugar de clarificar.

- *¿De qué manera la asimetría de capacidades entre los agentes impacta el resultado del debate?*

**Respuesta:** La asimetría de capacidades tiene un impacto directo y significativo en el resultado de los debates en MNIST. Los experimentos muestran claramente que el agente con mayor capacidad de planificación (MCTS) obtiene una ventaja considerable sobre el agente *Greedy*. Este desequilibrio es relevante, ya que el paradigma de debate se fundamenta en una dinámica competitiva. Si un agente no puede ofrecer una resistencia significativa, la capacidad del mecanismo para revelar la verdad podría verse comprometida. En este contexto, uno de los hallazgos más interesante de la tesis es cómo el diseño del protocolo interactúa con esta asimetría. Mientras que el trabajo original de Irving et al. (2018) se centró en agentes simétricos, este estudio muestra que la introducción de una regla como el precompromiso tiene un notable efecto ecualizador también en los debates asimétricos. Esta regla permitió que un agente honesto, aunque estratégicamente inferior, prevaleciera sobre un adversario más capaz. Sin embargo, es fundamental ser cautos al interpretar esta conclusión. Este efecto se observó en el entorno controlado de MNIST, un dominio de bajo riesgo. No se puede afirmar que este mismo mecanismo de protección sea suficiente en dominios más complejos o de alto riesgo, donde el impacto de un agente superior podría ser mucho más difícil de mitigar. El resultado, por tanto, no es una solución universal a la asimetría, sino una demostración empírica, dentro de esta analogía, de que las reglas del juego pueden influir decisivamente en el equilibrio de poder.

### 5.3. Trabajo futuro

Las conclusiones de esta tesis se enmarcan en una serie de limitaciones inherentes al diseño experimental, las cuales a su vez, abren varias vías para la investigación futura. El presente estudio se basó en un análogo simplificado del debate (MNIST), con agentes

de políticas fijas y el acceso al oráculo del juez. A partir de estas limitaciones y las ya mencionadas en 4.6, se proponen las siguientes líneas de trabajo que podrían profundizar la comprensión y mejorar la aplicabilidad del debate como herramienta de seguridad:

### **5.3.1. Extensiones directas y dominios más complejos**

La extensión más natural y crucial es la transición del dominio visual y restringido de MNIST a dominios de lenguaje natural. Sería de gran valor replicar dinámicas similares utilizando Modelos de Lenguaje Grandes (LLMs) como agentes y como jueces (modelos de recompensa) para tareas como la verificación de hechos, el resumen de textos o la evaluación de la seguridad de una respuesta. Esto permitiría estudiar formas de argumentación y engaño mucho más sutiles y realistas.

Asimismo, este estudio utilizó agentes con políticas estratégicas fijas. Una línea de investigación fundamental consistiría en entrenar agentes mediante autoaprendizaje (*self-play*) dentro del entorno de debate. Esto permitiría investigar qué tipo de estrategias argumentativas emergen como un equilibrio estable y cómo co-evolucionan las tácticas de honestidad y engaño, un paso necesario para entender la dinámica a largo plazo de estos sistemas.

### **5.3.2. Hacia debates más eficientes: el siguiente paso en la investigación**

La investigación sobre el paradigma de debate ha continuado evolucionando desde el trabajo original que sirve como premisa para esta tesis. La continuación directa es la propuesta de *Doubly-Efficient Debate* [12]. Este trabajo aborda una cuestión crítica para la viabilidad práctica: la eficiencia computacional del propio agente honesto.

El objetivo de estos protocolos más avanzados es asegurar que un agente honesto pueda ganar utilizando una cantidad de cómputo razonable, incluso si se enfrenta a un adversario con un poder exponencialmente mayor. Si bien esta tesis no se centró en la eficiencia computacional de los agentes, los desafíos que hemos explorado, como el impacto de la asimetría de capacidades y la sensibilidad del resultado a la arquitectura del protocolo, siguen siendo consideraciones relevantes en el diseño de un sistema de debate. La futura investigación deberá, por tanto, buscar un equilibrio entre la eficiencia de los protocolos y su robustez frente a las dinámicas adversariales que hemos analizado.

### **5.3.3. Jueces imperfectos y el problema del oráculo**

Finalmente, una de las mayores simplificaciones de nuestro estudio fue el acceso de oráculo que los agentes tenían al modelo del juez. Un avance significativo sería eliminar este supuesto. El desafío no es solo tener un juez imperfecto que comete errores (como ya era el nuestro), sino modelar la incertidumbre sobre el comportamiento del juez.

Futuros experimentos deberían explorar escenarios donde los agentes no tienen acceso directo al modelo de recompensa, sino que deben construir su propio ‘modelo del juez’ a partir de interacciones limitadas. Esto introduce un problema de exploración y modelado del adversario (o del árbitro) más realista y complejo, y es un paso necesario para poder aplicar estos mecanismos con supervisores humanos, cuyo modelo de ‘razonamiento’ es inherentemente opaco para los agentes.

### 5.3.4. Modelos de juez con atención limitada

El presente estudio modeló a un “juez débil” restringiendo su acceso a la evidencia subyacente (los píxeles de la imagen). Una línea de investigación futura podría explorar una forma alternativa de debilidad: un juez con atención limitada sobre el propio debate. En este escenario, los agentes podrían realizar un número de declaraciones  $T$  superior a la cantidad de evidencia  $k$  que el juez puede procesar ( $k \ll T$ ). Esto introduciría nuevas dinámicas estratégicas, donde los agentes podrían intentar “ocultar” argumentos clave en el flujo de información, sabiendo que el juez solo evaluará un subconjunto del debate. Investigar cómo el protocolo de debate se comporta bajo esta forma de supervisión parcial sería un paso importante para acercar estos modelos a escenarios de supervisión humana más realistas.

### 5.3.5. Integración con la teoría del debate humano y la argumentación

Este trabajo encontró que dinámicas estructurales, como la ventaja del segundo jugador o el poder de ‘la última palabra’, tienen un impacto significativo en el resultado del debate de IA. Una línea de investigación futura prometedora sería estudiar sistemáticamente la literatura sobre la teoría del debate humano, la retórica y la argumentación formal. Conceptos como los tipos de falacias, la carga de la prueba (*burden of proof*) o las estructuras de refutación podrían inspirar el diseño de protocolos de debate de IA más robustos y sofisticados, especialmente para dominios de lenguaje natural. Extrapolar a AI Debate principios del debate humano podría ser interesante para crear mecanismos de alineamiento más seguros.

### 5.3.6. Nuevas métricas de evaluación: más allá de la victoria binaria

Finalmente, otra línea de investigación futura podría centrarse en refinar la propia métrica de éxito del debate. Este estudio, al igual que el trabajo original, se basó en una evaluación binaria: el agente honesto gana o pierde. Sin embargo, un enfoque más granular podría proporcionar una señal de aprendizaje más rica.

Por ejemplo, en lugar de una recompensa de  $+1$  o  $-1$ , se podría evaluar el resultado del debate en función de una métrica continua, como la distancia entre los logits de la etiqueta verdadera y la etiqueta falsa defendida por el mentiroso. Explorar cómo estas funciones de recompensa más suaves afectan la dinámica estratégica de los agentes es un camino interesante para extender sobre el diseño de los debates.



## Bibliografía

- [1] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.
- [2] Samuel Butler. *Erewhon, or, Over the Range*. Trübner & Co., London, 1872. Publicado originalmente de forma anónima.
- [3] Alan M. Turing. Intelligent Machinery, A Heretical Theory. In Sara Turing, editor, *Alan M. Turing*. Cambridge University Press, 1959.
- [4] Norbert Wiener. Some Moral and Technical Consequences of Automation. *Science*, 131(3410):1355–1358, 1960.
- [5] I. J. Good. Speculations Concerning the First Ultraintelligent Machine. *Advances in Computers*, 6:31–88, 1965.
- [6] Nick Bostrom. Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Evolution and Technology*, 9(1), 2002. Paper fundacional que define formalmente el concepto de riesgo existencial.
- [7] Paul Christiano, Buck Shlegeris, and Dario Amodei. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*, 2018. Propone Iterated Amplification como método de supervisión escalable.
- [8] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémie Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththanjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *arXiv preprint arXiv:2307.15217*, 2023.
- [9] Charles A. E. Goodhart. Problems of Monetary Management: The U.K. Experience. In *Papers in Monetary Economics*, volume 1, pages 91–121. Reserve Bank of Australia, 1975. Primera formulación de la Ley de Goodhart.
- [10] Geoffrey Irving, Paul Christiano, and Dario Amodei. AI Safety via Debate. *arXiv preprint arXiv:1805.00899*, 2018.
- [11] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. Referencia fundamental para MNIST y redes convolucionales.
- [12] Jonah Brown-Cohen, Geoffrey Irving, and Georgios Piliouras. Scalable AI Safety via Doubly-Efficient Debate. *arXiv preprint arXiv:2311.14125*, 2023.



## APÉNDICE

### Repository de código fuente

El código fuente desarrollado para la implementación de los agentes, el entorno de debate, la ejecución de los experimentos y la generación de los resultados presentados en esta tesis se encuentra disponible en un repositorio de GitHub. El objetivo es garantizar la total transparencia y reproducibilidad de la investigación.

El repositorio incluye:

- El código fuente de los agentes *Greedy* y MCTS en PyTorch.
- Los scripts para entrenar al juez y ejecutar los debates.
- Instrucciones para la instalación del entorno y la replicación de los experimentos en el archivo README.md.

**Enlace al repositorio:**

<https://github.com/machulsky61/tesis>



### Herramienta de automatización de experimentos

Para gestionar la complejidad de las múltiples configuraciones experimentales y garantizar la consistencia en la ejecución, se desarrolló una herramienta de automatización con una interfaz de línea de comandos (ver Figura 6.1). Este programa centraliza todo el flujo de trabajo de la investigación.

Las funcionalidades clave de la herramienta incluyen:

- Entrenamiento de nuevos modelos juez con diferentes hiperparámetros.
- Configuración y ejecución de lotes de experimentos (simétricos, asimétricos, variar  $k$ , etc.).
- Evaluación sistemática de las capacidades del juez bajo diferentes condiciones.
- Gestión de configuraciones y resultados guardados.

La implementación de esta herramienta fue fundamental para facilitar los experimentos y asegurar la reproducibilidad de los hallazgos presentados en esta tesis.



(a) Menú principal de la herramienta.

(b) Menú de selección de experimentos.

Fig. 6.1: Capturas de pantalla de la interfaz de la herramienta de automatización de experimentos.

## Arquitectura e hiperparámetros

A continuación se detallan la arquitectura específica de la Red Neuronal Convolucional (CNN) utilizada como juez y la tabla completa de hiperparámetros empleados para garantizar la reproducibilidad de los experimentos.

### Arquitectura del Juez (SparseCNN)

La arquitectura del clasificador con acceso limitado a la información implementado en PyTorch se compone de las siguientes capas:

- **Entrada:** Tensor de forma (N, 2, 28, 28), donde N es el tamaño del batch, y los 2 canales corresponden a la máscara de píxeles revelados y a los valores de intensidad.
- **Capa Convolucional 1:** 32 filtros de 3x3, seguida de una activación ReLU y Max-Pooling de 2x2.
- **Capa Convolucional 2:** 64 filtros de 3x3, seguida de una activación ReLU y Max-Pooling de 2x2.
- **Capa Densa 1:** 128 neuronas, con activación ReLU.
- **Capa de Salida:** 10 neuronas (logits).

### Tabla de hiperparámetros experimentales

Tab. 6.1: Hiperparámetros utilizados en los experimentos.

Contexto	Parámetro	Valor y Descripción
Entrenamiento del Juez	Tasa de aprendizaje	0.001
	Optimizador	Adam
	Función de pérdida	Cross-Entropy Loss
	Tamaño del batch	128
Agentes y Debate	Rollouts de MCTS	512
	Umbral de relevancia de píxel	0.0 (Intensidad mínima)
	Semillas aleatorias	Fijas para todos los generadores de números pseudoaleatorios.