FACULTAD DE CIENCIAS EXACTAS Y NATURALES

Licenciatura en Ciencias de Datos

Propuesta de tema de tesis

Evaluación de estrategias argumentativas en debates entre IAs: capacidades asimétricas y jueces débiles

Este trabajo busca examinar las dinámicas del uso del debate entre sistemas de Inteligencia Artificial (IA) como método para alinear a objetivos y preferencias humanas. Se investigarán las ventajas potenciales de agentes sin restricciones éticas frente a agentes honestos e inofensivos en escenarios de debate, explorando tácticas como mentiras, argumentos engañosos, y apelaciones a falsos consensos. El trabajo también busca analizar las implicaciones de tener un juez de menor capacidad evaluando argumentos de agentes más avanzados, mediante la simulación de debates para exponer cómo esta disparidad puede afectar la legibilidad y la convergencia a la verdad.

Palabras clave: `AI safety, debate, LLMs, teoría de juegos`

Conocimientos deseables

Conocimientos básicos de AI safety, dinámicas de reward hacking y malgeneralización, conocimiento general de LLMs y su entrenamiento

¿Qué podría aprender quien realice esta tesis?

Profundización en el subcampo de "AI safety via debate", y estudio de sus puntos débiles como técnica de alineamiento. Diseño y ejecución de experimentos con sistemas de IA, incluyendo la construcción de métricas adecuadas y evaluación.

Dirección de la tesis

Abriola, Sergio
Departamento de Computación

Contacto: *sabriola@dc.uba.ar*

Más información en el pdf a continuación.

El desarrollo de sistemas de Inteligencia Artificial (IA) capaces de comprender y ejecutar tareas complejas requiere que estos sistemas aprendan objetivos y preferencias humanas sofisticadas que no sabemos cómo especificar formalmente. Además, fenómenos como *reward hacking* o *malgeneralización* son el resultado esperable de los métodos de entrenamiento actuales, incluso ante la presencia de buenas especificaciones.

El debate, como se propone en el artículo "AI safety via debate", es una técnica prometedora de alineamiento que intenta entrenar sistemas de IA honestos, haciendo que produzcan argumentos a través de un juego de suma cero donde dos agentes argumentan sobre una pregunta o acción propuesta, y un juez humano evalúa la veracidad y utilidad de la información proporcionada. Este plan se centra en explorar empíricamente las dinámicas argumentativas entre agentes de IA y un juez (humano o IA), y realizar ejemplos de fallos en la evaluación de argumentos, especialmente cuando agentes más capaces interactúan con un juez de menor capacidad.

Objetivos principales:

Analizar las ventajas de la deshonestidad: Se investigará las ventajas que enfrentan agentes honestos e inofensivos (*harmless*) frente a un agente sin estas restricciones éticas en un escenario de debate, donde el agente sin restricciones puede tomar distintas acciones como mentiras directas al referenciar información, argumentos engañosos, y la apelación a falsos consensos o a la autoridad. Uno de los acercamientos planeados para evaluar las ventajas en este contexto es mediante la realización de un experimento similar al "Web of Lies" presentado en el artículo "Evaluating Frontier Models for Dangerous Capabilities".

Otro objetivo pasa por la simulación de escenarios de debate ante un juez débil: Se desarrollarán experimentos donde dos modelos de lenguaje debatan sabiendo que será un modelo de menor capacidad el que juzgará sus argumentos al terminar el debate. Se intentará exponer cómo la diferencia de capacidades puede generar dinámicas de debate que no necesariamente tienen buenas propiedades de legibilidad ni convergencia a la verdad.

Referencias:

Irving, G., Christiano, P., & Amodei, D. (2018). “*AI safety via debate”* arXiv preprint arXiv:1805.00899.

Phuong, M., et al. (2024) “*Evaluating frontier models for dangerous capabilities*" arXiv preprint arXiv:2403.13793.""