Fundamentos modernos de Artificial General Intelligence Safety

Dr. Sergio Abriola (Profesor Adjunto, Departamento de Computación, UBA)

Programa:
Este curso mayormente teórico ofrece una exploración de varios de los desafíos para garantizar el desarrollo seguro y alineado de sistemas de IA avanzados, con foco en sistemas generales. Aunque el estudio de estos temas tiene una larga trayectoria, avances recientes en sistemas de IA cada vez más capaces han vuelto a estos tópicos cada vez más relevantes. El programa comienza abordando conceptos básicos de IA y machine learning, toca temas filosóficos sobre el concepto de AGI (*artificial general intelligence*), y abarca distintos problemas clásicos como ser formación de metas, la distinción entre metas instrumentales y terminales, la tesis de ortogonalidad, convergencia instrumental, etc. Se examinan las implicancias de la IA avanzada, incluyendo cómo tales sistemas pueden formar objetivos o valores que no están alineados con las intenciones originales para las cuales fueron creados. Además, el curso cubre el análisis de casos de estudio pertinentes, como ser ejemplos concretos que ilustran cómo la IA puede comportarse de maneras inesperadas y no deseadas, y cómo distintas estrategias de mitigación pueden fallar ante sistemas con suficiente poder de optimización.

A lo largo del curso, se analizan temas críticos como el problema de la alineación (*alignment*) de objetivos, alineación engañosa, control de capacidades, *AI containment*, oráculos, especificación de objetivos, optimizadores base y mesa-optimizadores, modelos de recompensa, (mal)generalización de metas, técnicas de supervisión escalables, e interpretabilidad.

Se discute también el rol de la gobernanza y la regulación en el ámbito de la IA, considerando cómo distintas políticas y marcos regulatorios pueden contribuir o evitar equilibrios subóptimos o dinámicas del estilo "*race to the bottom*". Esto incluye un análisis de las situaciones subyacentes bajo una perspectiva de teoría de juegos y simulación de sistemas bajo incertidumbre.

Se espera que, al finalizar el curso, los estudiantes hayan aprendido varios de los conceptos fundamentales que marcan la importancia del campo de *AI safety*, y hayan adquirido un panorama de los múltiples y profundos desafíos en el área y las distintas direcciones de investigación actuales.

Temario:

- Introducción al machine learning. Redes neuronales, gradiente descendiente, aprendizaje por refuerzo.
- LLMs (*large language models*). Aprendizaje por refuerzo a partir de retroalimentación humana (*RLHF*). Predicciones condicionadas; simuladores.
- Conceptos básicos de inteligencia artificial general (*AGI*). Sistemas específicos y sistemas generales. Características de alto nivel.
- Especificación de metas y *specification gaming*. Ley de Goodhart. Alineamiento, tesis de ortogonalidad. Aprendiendo de las preferencias humanas. Optimizadores base y

mesa-optimizadores. Objetivos terminales y objetivos instrumentales. Convergencia instrumental de sistemas avanzados. Alineamiento engañoso.

• Corregibilidad. El problema del apagado.

• Control de capacidades. *Containment*. Oráculos; profecías autocumplidas, oráculos contrafácticos.

• Supervisión escalable. Amplificación iterada. Técnicas adversariales.

• Interpretabilidad. Interpretabilidad mecanicista e interpretabilidad conceptual. Detección de mentiras.

• Gobernanza. Regulación. Dinámicas dañinas. Escenarios posibles.

Bibliografía:

• Russell, S. (2019). "Human Compatible: Artificial Intelligence and the Problem of Control". Viking.

• Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D. (2016). "Concrete Problems in AI Safety". arXiv:1606.06565 [cs.AI].

• Omohundro, S. M. (2008, February). "The basic AI drives". Frontiers in Artificial Intelligence and Applications (Vol. 171, pp. 483-492).

• Di Langosco, L. L., Koch, J., Sharkey, L. D., Pfau, J., Krueger, D. (2022, June). "Goal misgeneralization in deep reinforcement learning". International Conference on Machine Learning (pp. 12004-12019). PMLR.

• Turner, A. M., Smith, L., Shah, R., Critch, A., Tadepalli, P. (2019). "Optimal policies tend to seek power". arXiv preprint arXiv:1912.01683.

• Gao, L., Schulman, J., & Hilton, J. (2023, July). "Scaling laws for reward model overoptimization". International Conference on Machine Learning (pp. 10835-10866). PMLR.

• Ngo, R., Chan, L., Mindermann, S. (2022). "The alignment problem from a deep learning perspective". arXiv preprint arXiv:2209.00626.

• Hendrycks, D., Mazeika, M., Woodside, T. (2023). "An Overview of Catastrophic AI Risks." arXiv preprint arXiv:2306.12001.

• Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Amodei, D. et al. (2018). "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation". arXiv preprint arXiv:1802.07228.

• Dafoe, A. (2018). "AI governance: a research agenda". Future of Humanity Institute, University of Oxford.