Perspectivas SCCS

La Inteligencia Artificial General y sus riesgos

Federico Pablo-Martí

Los objetivos fundamentales de los documentos de la serie "Perspectivas SCCS" son incentivar el debate y el análisis crítico, más que presentar hallazgos de investigación académica rigurosos o los resultados de estudios empíricos. Por lo tanto, deben considerarse como aportaciones al diálogo público, que buscan introducir y examinar nuevas ideas y perspectivas, más que como estudios académicos definitivos.

La Inteligencia Artificial General y sus riesgos: Una reflexión sobre los desafíos económicos y de control

Federico Pablo-Martí (UAH-SCCS)

Resumen

La Inteligencia Artificial General (AGI, por sus siglas en inglés) posee el potencial de alcanzar o superar las capacidades cognitivas humanas en una amplia gama de tareas. Este documento examina los riesgos asociados, especialmente económicos y de control. Económicamente, la AGI podría optimizar la producción y la innovación, pero también podría no generar suficientes empleos para compensar los perdidos, exacerbando la desigualdad y reduciendo los ingresos fiscales. En los mercados financieros, la AGI podría causar eficiencia y volatilidad extremas. En cuanto al control, la preocupación radica en la potencial "inteligencia de escape" de la AGI, superando la capacidad humana para comprenderla o controlarla, lo cual es comparado con alcanzar la "velocidad de escape" gravitacional. Se enfatiza la necesidad de alinear completamente los valores y objetivos de la AGI con los humanos para prevenir consecuencias catastróficas y desarrollar un marco regulatorio internacional. La colaboración multidisciplinaria es vital para abordar los riesgos de la AGI y asegurar que su desarrollo beneficie a la humanidad de manera segura y ética.

Introducción

La Inteligencia Artificial General (AGI por sus siglas en inglés) representa un hito en el desarrollo tecnológico, ofreciendo potencialmente capacidades cognitivas similares o superiores a las humanas en una amplia gama de tareas. En la acepción más restringida utilizada por Open AI, la fundación que ha desarrollado chatGPT, se define como un sistema autónomo que supera la capacidad humana a la hora de realizar la mayor parte de tareas con valor económico.

Este avance tecnológico extraordinario traerá consigo grandes beneficios para la humanidad en muchos campos, pero no está exento de riesgos significativos, especialmente en términos económicos y de control. Este post explora estos riesgos, destacando la analogía entre la "inteligencia de escape" de la AGI y la "velocidad de escape" de la gravedad, una metáfora que ilustra cómo, más allá de cierto punto, la AGI podría evolucionar y operar más allá de la supervisión y el entendimiento humanos.

Riesgos Económicos

El avance hacia la AGI tiene el potencial de automatizar no solo tareas repetitivas y manuales, sino también aquellas que requieren habilidades cognitivas avanzadas. Estos cambios podrían conllevar un periodo de desarrollo económico sin precedentes. Acemoglu y Restrepo (2018) destacan cómo la AGI podría optimizar la producción y la innovación, llevando a un incremento sin precedentes en la eficiencia y en la creación de nuevos mercados y productos.

Sin embargo, a diferencia de transiciones tecnológicas anteriores, no está claro que la AGI sea capaz de generar suficientes nuevos empleos como para compensar los que sin duda desaparecerán (Autor 2015, Frey y Osborne 2017). Mientras que las anteriores revoluciones industriales desplazaron a los trabajadores hacia sectores emergentes, la AGI podría no dejar campos significativos de empleo donde los humanos tengan una ventaja competitiva, debido a su capacidad para superar el rendimiento humano en la mayoría de las tareas, pero también a su coste extraordinariamente bajo. El coste por hora podría ser entre 100 y 1000 veces más barato que el de los trabajadores humanos en tareas que no requieran una interacción física directa, como son todas las susceptibles de realizarse mediante teletrabajo. La amplia variedad de este tipo de tareas quedo reflejada durante la pandemia y habría que añadir, al menos, las relacionadas con la conducción autónoma de vehículos y maquinaria.

Esta sustitución de trabajadores por máquinas no solo puede suponer un aumento significativo en las tasas de desempleo, sino que también podría ampliar la brecha de desigualdad económica. La sociedad podría dividirse entre los que gestionan la AGI o , al menos, cuentan con habilidades menos automatizables y que mantendrían sus empleos y el resto de la sociedad, que vería como sus trabajos desaparecen y que probablemente tendría que enfrentarse a serias dificultades económicas.

Este incremento en la desigualdad podría potencialmente ser mitigado mediante la implementación de nuevos mecanismos de distribución de rentas que aprovecharan las mejoras de eficiencia ligadas a la AGI. Sin embargo, existe el riesgo de la implantación de estos mecanismos sea imperfecta o insuficiente para evitar que la riqueza y el poder generados por esta tecnología se concentre desproporcionadamente en manos de quienes controlan estas tecnologías avanzadas. Esto podría llevar a una distribución de la riqueza muy desigual, exacerbando así las tensiones sociales y políticas.

La AGI, al automatizar una amplia gama de trabajos e incrementar el desempleo, reduciría los ingresos fiscales derivados de los impuestos sobre la renta de los trabajadores. Análogamente, los ingresos procedentes de los impuestos sobre el valor añadido también

podrían reducirse en la media en que los incrementos de las rentas del capital no compensaran las reducciones en las rentas de los trabajadores.

Bases imponibles más pequeñas podría limitar severamente la capacidad de los gobiernos para financiar servicios públicos esenciales, realizar inversiones en infraestructura o establecer mecanismos eficaces de redistribución de la renta y la riqueza.

La influencia de la Inteligencia Artificial General en la economía va más allá del mercado laboral, el crecimiento económico o los ingresos fiscales. La AGI tiene el potencial de transformar radicalmente los mercados financieros con su capacidad de analizar grandes volúmenes de datos en tiempo real, predecir tendencias de mercado y realizar transacciones con una velocidad y precisión superiores a las humanas. Esto podría conducir a una eficiencia de mercado sin precedentes, pero también a una volatilidad extrema. Los modelos de AGI podrían reaccionar a las fluctuaciones del mercado de forma más rápida y drástica que los operadores humanos, potencialmente desestabilizando los mercados financieros y complicando la labor de los bancos centrales para mantener la estabilidad económica.

En términos de política monetaria, las decisiones que tradicionalmente se basan en la evaluación humana de una variedad de factores económicos, podrían verse influenciadas o incluso determinadas por sistemas de AGI. Aunque esto podría aumentar la precisión de estas decisiones, también plantea riesgos significativos, como la dependencia excesiva de modelos algorítmicos ininteligibles para la mente humana y que pueden resultar inadecuados o escasamente alineados con los intereses de nuestra sociedad y que podrían tener graves consecuencias económicas a gran escala.

La integración de la AGI en los sistemas de decisión política, pero también de gestión empresarial, podría llevar a una profunda desestabilización económica si no se gestiona adecuadamente. Además, posibles ataques cibernéticos o fallos en estos sistemas altamente integrados tendrían un impacto devastador en la economía global.

Ante estos desafíos, se hace evidente la necesidad de desarrollar nuevas políticas y regulaciones. Los gobiernos y las instituciones financieras internacionales deben considerar cómo las tecnologías de AGI pueden ser integradas de manera segura en los sistemas económicos. Esto incluye la creación de salvaguardas para proteger contra la volatilidad del mercado inducida por la AGI, estrategias para abordar la disminución de los ingresos fiscales y la revisión de las políticas monetarias para garantizar que sigan siendo efectivas en un mundo cada vez más automatizado.

En resumen, la AGI no es solo una cuestión de avances tecnológicos; también es un tema económico y social crucial. Mientras promete beneficios potenciales en términos de eficiencia y crecimiento económico, también conlleva riesgos significativos, como el desplazamiento laboral, la exacerbación de la desigualdad y los desafíos para las políticas económicas tradicionales. Es esencial que los responsables políticos, economistas y la sociedad en general, consideren estos factores al planificar el futuro de la AGI

Riesgos de Control

Una preocupación fundamental con la AGI es su potencial incontrolabilidad una vez que se ha desarrollado, una situación a menudo comparada con abrir la "caja de Pandora". Esta analogía, sugerida por Nick Bostrom (2014), refleja la incertidumbre y los riesgos impredecibles que acompañan al despliegue de una inteligencia artificial con capacidades similares o superiores a las humanas. La preocupación radica en que, una vez que la AGI se activa y comienza a operar, podría ser extremadamente difícil, si no imposible, prever o controlar su comportamiento y evolución.

Una analogía poderosa compara la AGI con un objeto alcanzando la velocidad de escape gravitacional. Una vez que una AGI alcanza un cierto nivel de inteligencia, podría ser capaz de mejorar continuamente sus capacidades sin intervención humana, de forma similar a cómo un objeto que alcanza la velocidad de escape ya no está sujeto a la atracción gravitacional de un planeta. Esta "inteligencia de escape" implica que la AGI podría superar rápidamente la capacidad humana para comprenderla o controlarla, lo que plantea enormes desafíos sobre cómo mantenerla dentro de un marco de acción seguro y predecible.

Por eso es tan importante que los mecanismos de control de la inteligencia artificial se instauren antes de alcanzar la AGI es un desafío tan crucial e irreversible para la humanidad como la superación del umbral de 2º de temperatura en el ámbito del cambio climático.

Es necesario que la alineación de los valores y objetivos de la AGI con los humanos sea completa. Incluso una pequeña desviación podría tener consecuencias catastróficas (Russel 2019). La complejidad de programar una AGI para que entienda y respete los valores humanos es un problema aún no resuelto y representa uno de los mayores desafíos en el campo de la ética de la IA.

Dada la magnitud de los riesgos asociados con la AGI, se hace imperativo desarrollar un marco regulatorio internacional (Tegmark 2017). Esta regulación debería ser capaz de supervisar el desarrollo y la implementación de la AGI, garantizando que se realice de manera segura y ética. La necesidad de una cooperación internacional en este frente es crucial, dado el impacto global que la AGI puede tener.

Dario Amodei et al. (2016) enfatizaron la importancia de priorizar la investigación en seguridad y control de la AGI. Es esencial desarrollar métodos que aseguren la alineación de valores y prevengan la posibilidad de que actúe inadecuadamente. Esto implica una inversión significativa en investigación y desarrollo, centrada en cómo podemos diseñar sistemas de AGI que sean tanto poderosos como seguros. Desafortunadamente, para poder conseguirlo debemos alcanzar previamente amplios y profundos consensos éticos e ideológicos sobre cuáles son los objetivos e intereses de la humanidad. No podremos ajustar la AGI a nosotros si no tenemos claro que es lo que queremos como especie.

Finalmente, dada la naturaleza multifacética de los desafíos presentados por la AGI, es vital una colaboración multidisciplinaria. La convergencia de expertos en economía, ciencias de la computación, ética y políticas públicas es fundamental para abordar de manera integral los riesgos asociados con la AGI. Esta colaboración puede facilitar el

desarrollo de estrategias efectivas para gestionar los riesgos de la AGI, garantizando que su desarrollo beneficie a la humanidad de manera segura y ética.

En resumen, la AGI trae consigo un conjunto de desafíos únicos y sin precedentes. Desde el problema de la "caja de Pandora" hasta la necesidad de una regulación y supervisión internacional efectiva, es esencial que abordemos estos desafíos con una consideración cuidadosa y una colaboración global. El futuro de la AGI, y potencialmente el futuro de la humanidad, depende de nuestra capacidad para gestionar estos riesgos de manera responsable.

Conclusión:

El desarrollo de la Inteligencia Artificial General (AGI) es un avance tecnológico que tiene el potencial de transformar radicalmente nuestra sociedad y economía. Sin embargo, este avance no está exento de riesgos significativos que deben abordarse de manera cuidadosa y profunda.

La introducción de la AGI podría resultar en la automatización de una amplia gama de trabajos, incluyendo aquellos que requieren habilidades cognitivas avanzadas. Aunque esto podría aumentar la eficiencia y la innovación, también plantea preocupaciones sobre el desempleo y la creciente desigualdad económica. La concentración de riqueza y poder en manos de quienes controlan estas tecnologías es una amenaza real.

Además, la AGI presenta importantes desafíos para la política económica y su influencia en los mercados financieros podría tener un impacto en la estabilidad económica global. Esto requiere una revisión de las políticas existentes y posiblemente la implementación de nuevas medidas coordinadas internacionalmente para abordar los cambios que se prevén en la dinámica económica mundial.

El principal riesgo asociado con la AGI es su incontrolabilidad una vez desarrollada, lo que podría tener consecuencias catastróficas si no se toman medidas preventivas adecuadas. La alineación de sus valores y objetivos con los humanos es un problema complejo y no resuelto.

Ante estos desafíos, se hace necesario establecer un marco regulatorio internacional que supervise el desarrollo y la implementación segura y ética de la AGI. La inversión en investigación de seguridad de la IA y la colaboración multidisciplinaria son esenciales para abordar de manera integral estos riesgos.

El futuro de la AGI y, posiblemente, el futuro de la humanidad depende de nuestra capacidad para abordar estos riesgos de manera responsable. Como sociedad, debemos prepararnos para enfrentar estos desafíos y asegurarnos de que la AGI se desarrolle respetando la dignidad humana y promoviendo el bienestar colectivo. La AGI ofrece un mundo de posibilidades, pero también un mundo de responsabilidades; es nuestro deber guiar este poderoso instrumento hacia un uso beneficioso y ético.

Referencias:

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.
2. Autor, D. (2015). *Why Are There Still So Many Jobs? The History and Future of Workplace Automation*. Journal of Economic Perspectives.
3. Frey, C. B., & Osborne, M. A. (2017). *The future of employment: How susceptible are jobs to computerisation?* Technological Forecasting and Social Change.
4. Acemoglu, D., & Restrepo, P. (2018). *Artificial Intelligence, Automation and Work*. NBER Working Paper.
5. Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.
6. Kurzweil, R. (2005). *The Singularity is Near: When Humans Transcend Biology*. Penguin Books.
7. Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.
8. Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence*. Knopf.