Introducción a la seguridad de la IA, ética y sociedad

El rápido progreso en IA plantea interrogantes sobre cómo el despliegue de sistemas avanzados de IA afectará a la sociedad, tanto para bien como para mal. ¿Cómo podemos comprender y mitigar los riesgos de estos sistemas de IA? ¿Cómo garantizamos que la IA se implemente de manera ética y tenga un impacto social positivo?

Este curso, desarrollado por Dan Hendrycks, director del Centro para la Seguridad de la IA, tiene como objetivo proporcionar una introducción accesible a estudiantes, profesionales y otras personas que buscan comprender mejor estos problemas.

Garantizar que los sistemas de IA sean seguros es más que un simple problema de aprendizaje automático: es un desafío social que trasciende las fronteras disciplinarias tradicionales. Este curso adopta un enfoque holístico que se basa en conocimientos de la ingeniería, la economía y otros campos relevantes.

El curso tiene como objetivo fomentar una comprensión reflexiva y matizada de la seguridad de la IA, equipando a los participantes con las herramientas y los conocimientos necesarios para navegar en este campo en rápida evolución. Los temas clave cubiertos incluyen:

- Fundamentos de los sistemas de IA modernos y aprendizaje profundo, leyes de escalamiento y sus implicaciones para la seguridad de la IA
- Desafíos técnicos en la creación de una IA segura, incluida la opacidad, los juegos proxy y los ataques adversarios, y sus consecuencias para la gestión de los riesgos de la IA.
- Las diversas fuentes de riesgos a escala social derivados de la IA avanzada, como el uso malicioso, los accidentes, la IA no autorizada y el papel de la dinámica de carrera de la IA y los riesgos organizacionales.
- La naturaleza sistémica de la seguridad de la IA, la relevancia de la ingeniería de seguridad y la teoría de sistemas complejos, y la importancia de gestionar los eventos de cola y los cisnes negros
- Problemas de acción colectiva asociados con el desarrollo de la IA y desafíos con la construcción de sistemas cooperativos de IA
- Enfoques para la gobernanza de la IA, incluidos estándares de seguridad y tratados internacionales, y compensaciones entre el acceso centralizado y descentralizado a la IA avanzada.

Prefacio

La Inteligencia Artificial (IA) se está integrando rápidamente en los ejércitos, las economías y las sociedades, remodelando sus cimientos mismos. Dada la profundidad y amplitud de sus consecuencias, nunca ha sido más urgente comprender cómo garantizar que los sistemas de IA sean seguros, éticos y tengan un impacto social positivo.

Este libro de texto tiene como objetivo proporcionar un enfoque integral para comprender el riesgo de la IA. Nuestros objetivos principales incluyen consolidar el conocimiento fragmentado sobre el riesgo de la IA, aumentar la precisión de las ideas centrales y reducir las barreras de entrada al hacer que el contenido sea más simple y comprensible. El libro ha sido diseñado para ser accesible a lectores de diversos orígenes académicos. No es necesario haber estudiado IA, filosofía u otros temas similares. El contenido es amigable y algo modular, para que puedas elegir qué capítulos leer. Introducimos fórmulas en algunos lugares para especificar las afirmaciones con mayor precisión, pero los lectores deberían poder comprender los puntos principales sin ellas.

El riesgo de la IA es multidisciplinario. La mayoría de la gente piensa en los problemas relacionados con el riesgo de la IA en términos de modelos conceptuales en gran medida implícitos que tienen consecuencias significativas para el contenido de su pensamiento, y nuestro objetivo es reemplazar estos modelos implícitos con modelos explícitos y probados en el tiempo. Para comprender plenamente los riesgos que plantea la IA se requieren conocimientos en varias disciplinas académicas dispares, que hasta ahora no se han combinado en un solo texto. Este libro fue escrito para llenar ese vacío y equipar adecuadamente a los lectores para analizar el riesgo de la IA. Este libro de texto va más allá de los límites del ML para proporcionar una comprensión integral del riesgo de la IA. Nos basamos en ideas y marcos bien establecidos de los campos de la ingeniería, la economía, la biología, los sistemas complejos, la filosofía y otras disciplinas que pueden proporcionar información sobre los riesgos de la IA y cómo gestionarlos. Nuestro objetivo es dotar a los lectores de una comprensión sólida de los desafíos técnicos, éticos y de gobernanza que necesitaremos enfrentar para aprovechar la IA avanzada de manera beneficiosa.

Para tener una comprensión sólida de los desafíos de la seguridad de la IA, es importante considerar el contexto más amplio dentro del cual se están desarrollando y aplicando los sistemas de IA. Las decisiones y la interacción entre los desarrolladores de IA, los formuladores de políticas, militares y otros actores desempeñarán un papel importante en la configuración de este contexto. Dado que la IA influye en muchas esferas diferentes, hemos seleccionado deliberadamente marcos formales probados en el tiempo para proporcionar múltiples lentes para pensar sobre la IA, los actores relevantes y los impactos de la IA. Los marcos y conceptos que utilizamos son muy generales y útiles para razonar sobre diversas formas de inteligencia, desde seres humanos individuales hasta corporaciones, estados y sistemas de inteligencia artificial. Si bien algunas secciones del libro se centran más directamente en los riesgos de la IA que ya han sido identificados y discutidos hoy, otras establecen una introducción sistemática a ideas de la teoría de juegos, sistemas complejos,

relaciones internacionales y más. Esperamos que proporcionar estas herramientas conceptuales flexibles ayude a los lectores a adaptarse sólidamente al panorama en constante cambio de los riesgos de la IA.

Este libro no pretende ser la última palabra sobre todos los posibles riesgos de la IA, ya que la investigación sobre los riesgos de la IA es todavía relativamente nueva y se está desarrollando rápidamente. La variedad de temas tratados en este libro también significa que hay muchos detalles y matices que no tenemos espacio para abordar. Un libro que buscara el máximo rigor y profundidad sería mucho más largo que éste y probablemente también introduciría más jerga técnica y formalismo, reduciendo su accesibilidad. En lugar de esto, nuestro objetivo es presentar algunos conceptos y marcos que consideramos muy productivos para pensar en diversos riesgos de la IA. Dada la amplia gama de los problemas involucrados, es fácil desorientarse. Nuestra esperanza es que este libro de texto proporcione un punto de partida para otros a medida que construyen una imagen más detallada de estos riesgos y las posibles respuestas a ellos.

El contenido del libro de texto se divide en tres secciones: IA y riesgos a escala social, seguridad y ética y sociedad. En la sección IA y riesgos a escala social, describimos las principales categorías de riesgos de la IA e introducimos algunas características clave de los sistemas de IA modernos. En la sección Seguridad, analizamos cómo hacer que los sistemas de IA individuales sean más seguros. Sin embargo, si podemos hacerlos seguros, ¿cómo deberíamos dirigirlos? Para responder a esto, pasamos a la sección Ética y Sociedad y analizamos cómo crear sistemas de IA que promuevan nuestros valores más importantes. En esta sección, también exploramos los numerosos desafíos que surgen cuando hay múltiples sistemas de IA o múltiples desarrolladores de IA con intereses contrapuestos.

La sección La IA y los riesgos a escala social comienza con una descripción general informal de los riesgos de la IA, que resume muchas de las preocupaciones clave analizadas en este libro. Describimos algunos escenarios en los que los sistemas de IA podrían causar resultados catastróficos. Dividimos los riesgos en cuatro categorías: uso malicioso, dinámica de carrera armamentista de IA, riesgos organizacionales e IA no autorizadas. Estas categorías se pueden asignar de manera general a los riesgos discutidos con mayor profundidad en los capítulos de Gobernanza, Problemas de acción colectiva, Ingeniería de seguridad y Seguridad de agente único, respectivamente. Sin embargo, este mapeo es imperfecto ya que muchos de los riesgos y marcos discutidos en el libro de texto son más generales y abarcan todos los escenarios. No obstante, esperamos que los escenarios de este primer capítulo brinden a los lectores una imagen más concreta de los riesgos que exploramos en este libro. El capítulo Fundamentos de la IA ofrece una explicación accesible y no matemática de los sistemas de IA actuales, estableciendo conceptos de aprendizaje automático, aprendizaje profundo, leyes de escala, etc. Esto proporciona las bases necesarias para la discusión sobre la seguridad de los sistemas de IA individuales en la siguiente sección.

La sección Seguridad tiene como objetivo proporcionar una descripción general de los desafíos principales en la construcción segura de sistemas avanzados de

IA. Se basa en conocimientos tanto de la investigación sobre el aprendizaje automático como de las teorías generales de la ingeniería de seguridad y los sistemas complejos que proporcionan una lente poderosa para comprender estas cuestiones. En Seguridad de un solo agente, exploramos los desafíos para hacer que los sistemas de IA individuales sean más seguros, como el sesgo, la transparencia y la emergencia. En Ingeniería de seguridad, analizamos los principios para crear organizaciones más seguras y cómo estos pueden aplicarse a quienes desarrollan e implementan IA. La necesidad de una cultura de seguridad sólida en las organizaciones que desarrollan IA es crucial, para que las organizaciones no prioricen las ganancias a expensas de la seguridad. A continuación, en Sistemas complejos, mostramos que analizar las IA como sistemas complejos nos ayuda a comprender mejor la dificultad de predecir cómo responderán a las presiones externas o controlar los objetivos que pueden surgir en dichos sistemas. De manera más general, este capítulo nos proporciona un vocabulario útil para discutir diversos sistemas de interés.

La sección Ética y Sociedad se centra en cómo inculcar objetivos y limitaciones beneficiosos en los sistemas de IA y cómo permitir una colaboración eficaz entre las partes interesadas para mitigar los riesgos. En el capítulo IA beneficiosa y ética de las máquinas, presentamos el desafío de dar a los sistemas de IA objetivos que conduzcan de manera confiable a resultados beneficiosos para la sociedad, y discutimos varias propuestas junto con los desafíos que enfrentan. En Problemas de acción colectiva, utilizamos la teoría de juegos para ilustrar las muchas formas en que múltiples agentes (humanos, IA, grupos de humanos e IA) pueden no lograr buenos resultados y entrar en conflicto. También consideramos las dinámicas evolutivas que dan forma al desarrollo de la IA y cómo estas impulsan los riesgos de la IA. Estos marcos nos ayudan a comprender los desafíos de gestionar las presiones competitivas entre los desarrolladores de IA, los ejércitos o los propios sistemas de IA. Finalmente, en el capítulo sobre Gobernanza, analizamos variables estratégicas como el ritmo al que evolucionan los sistemas de IA y la amplitud con la que se distribuye el acceso a sistemas de IA potentes. Introducimos una variedad de caminos potenciales para gestionar los riesgos de la IA, incluida la gobernanza corporativa, la regulación nacional y la coordinación internacional.

El sitio web de este libro de texto (www.aisafetybook.com) incluye una variedad de contenido adicional. Contiene más recursos educativos, como vídeos, diapositivas, cuestionarios y preguntas de debate. Para los lectores interesados en contribuir a mitigar los riesgos de la IA, ofrece algunas sugerencias breves y enlaces a otros recursos sobre este tema. También se puede encontrar una variedad de apéndices en el sitio web con material adicional que no se pudo incluir en el libro de texto en sí.

Dan Hendrycks
Centro para la seguridad de la IA