Investigación Técnica sobre la Seguridad de la Inteligencia Artificial

En pocas palabras: Para mitigar los riesgos que plantea el desarrollo de la inteligencia artificial, es imperativo investigar cómo resolver los retos técnicos y los problemas de diseño para garantizar que los potentes sistemas de IA hagan lo que queremos - y sean beneficiosos - sin consecuencias catastróficas imprevistas.

Recomendado

Si se te da bien esta carrera, puede ser la mejor manera de que tengas un impacto social.

Estado de la revisión
Basado en una investigación de profundidad media

¿Por qué trabajar en Investigación de la Seguridad de la IA puede tener un Alto Impacto?

Como hemos argumentado, en las próximas décadas podríamos asistir al desarrollo de potentes algoritmos de aprendizaje automático con el potencial de transformar la sociedad. Esto podría tener importantes ventajas y desventajas, incluida la posibilidad de riesgos catastróficos.

Además de la labor estratégica y política que se analiza en esta revisión de carrera, otra forma clave de limitar estos riesgos es la investigación de los retos técnicos que plantean los potentes sistemas de IA, como el problema de la

alineación. En pocas palabras, ¿Cómo diseñamos sistemas de IA potentes para que hagan lo que queremos y no tengan consecuencias imprevistas?
Este campo de investigación ha empezado a despegar, y ahora hay importantes centros académicos y laboratorios de IA donde se puede trabajar en estas cuestiones, como Mila en Montreal, el Future of Humanity Institute (Instituto para el Futuro de la Humanidad) de Oxford, el Center for Human-Compatible Artificial Intelligence (Centro de Inteligencia Artificial Humanamente Compatible) de Berkeley, DeepMind en Londres y OpenAI en San Francisco. Hemos asesorado a más de 100 personas en este camino, y varias de ellas ya trabajan en las instituciones mencionadas. El Machine Intelligence Research Institute (Instituto de Investigación de Inteligencia de Máquinas) de Berkeley lleva trabajando en este campo desde 2005 y tiene una perspectiva y un programa de investigación poco convencionales en comparación con los demás laboratorios.

Hay mucha financiación disponible para investigadores con talento, incluidas becas académicas y donaciones filantrópicas de grandes entidades como Open Philanthropy (Filantropía Abierta). También es posible conseguir financiación para tu programa de doctorado. La principal necesidad del campo es más gente capaz de utilizar esta financiación para llevar a cabo la investigación.

¿En qué consiste esta vía?

En este camino, el objetivo es conseguir un puesto en uno de los principales centros de investigación sobre seguridad de la IA -ya sea en la industria, en organizaciones sin ánimo de lucro o en el mundo académico- y luego tratar de trabajar en las cuestiones más apremiantes, con el objetivo final de convertirse en un investigador principal que supervise la investigación sobre seguridad.

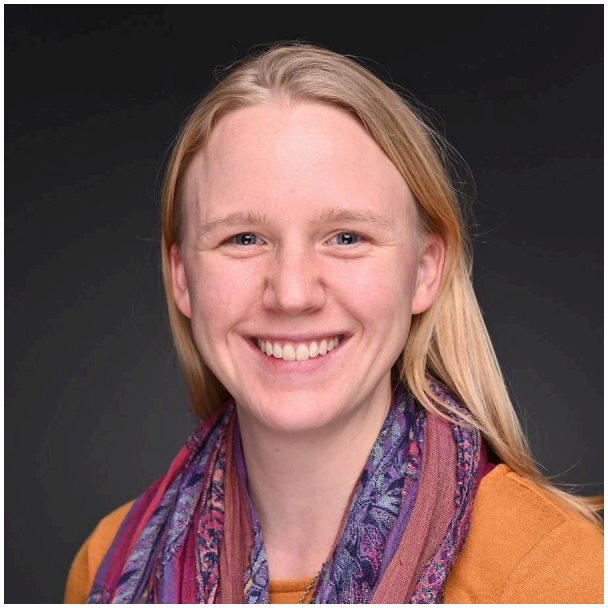
A grandes rasgos, los puestos técnicos en seguridad de la IA pueden dividirse en (i) Investigación e (ii) Ingeniería. Los investigadores dirigen el programa de investigación. Los ingenieros crean los sistemas y realizan los análisis necesarios para llevar a cabo la investigación.

Aunque los ingenieros tienen menos influencia sobre los objetivos de alto nivel de la investigación, sigue siendo importante que se preocupen por la seguridad, ya que comprenderán mejor los objetivos finales de la investigación (y así priorizarán mejor), estarán más motivados, cambiarán la cultura hacia la seguridad y utilizarán el capital de carrera que adquieran en beneficio de otros proyectos de seguridad en el futuro. Esto significa que la ingeniería puede ser una buena alternativa para quienes no quieren ser investigadores científicos.

También puede ser útil contar con personas que entiendan los retos de la seguridad de la IA trabajando en equipos de investigación de IA que no se centren directamente en la seguridad de la IA. Trabajar en estos equipos puede colocarte en una posición que te ayude a promover la preocupación por la seguridad en general, especialmente si acabas en un puesto directivo con influencia sobre las prioridades de la organización.

También nos entusiasmaría ver a más gente adquirir experiencia para trabajar en la seguridad de la IA en China o en relación con este país: lee más en nuestra revisión de carrera sobre las vías de seguridad y gobernanza de la IA relacionadas con China, algunas de las cuales adoptan la forma de investigación técnica.

Ejemplos de Personas que Siguen este Camino



Catherine Olsson

Catherine comenzó su doctorado en la Universidad de Nueva York trabajando en modelos computacionales de la visión humana. Con el tiempo, decidió trabajar directamente en seguridad de la IA, y consiguió un trabajo en OpenAI, y luego en Google Brain, antes de trasladarse a Anthropic.

SABER MÁS

Daniel Ziegler

Tras abandonar su doctorado en aprendizaje automático en Stanford, Daniel, que siempre había disfrutado construyendo cosas y quería ayudar a dar forma al desarrollo de la IA, decidió presentar su candidatura a OpenAI. Pasó seis semanas preparando la entrevista y consiguió el puesto. Su doctorado, en cambio, le habría llevado seis años. Daniel cree que este camino tan acelerado puede ser posible para muchos otros.

SABER MÁS

Chris Olah

Chris ha tenido una trayectoria fascinante y poco convencional. Chris no solo no tiene un doctorado, sino que ni siquiera tiene una licenciatura. Tras abandonar la universidad para ayudar a defender a un conocido que se enfrentaba a cargos penales injustos, Chris empezó a trabajar de forma independiente en la investigación del aprendizaje automático y, finalmente, consiguió unas prácticas en Google Brain.

SABER MÁS

Cómo evaluar tu idoneidad

Es probable que la investigación técnica sobre seguridad de la IA de mayor repercusión la lleven a cabo personas que ocupen los principales puestos de trabajo enumerados anteriormente. Así que para decidir si este camino es

adecuado para ti, es importante considerar si tienes una posibilidad razonable de conseguir esos puestos de trabajo.

- ¿Tienes posibilidades de entrar en una de las cinco mejores escuelas de posgrado en aprendizaje automático? Esto puede ser una buena prueba para saber si podrías conseguir un trabajo en un centro de investigación de IA de primer nivel, aunque no es un requisito.
- ¿Estás convencido de la importancia de la seguridad de la IA a largo plazo?
- ¿Eres ingeniero de software o de aprendizaje automático y has trabajado en alguna de las empresas que hacen parte de FAANG (Acrónimo que se refiere a las 5 empresas más grandes del sector tecnológico estadounidense, a saber: Facebook/Meta, Apple, Amazon, Netflix y Google/Alphabet)) y otras empresas competitivas? Es posible que puedas formarte para acceder a un puesto de investigador o, de lo contrario, ocupar un puesto de ingeniero.
- ¿Tienes posibilidades de contribuir a una cuestión de investigación relevante? Por ejemplo, ¿Estás muy interesado en el tema, tienes ideas sobre cuestiones que investigar y no puedes resistirte a llevarlas a cabo? Más información sobre cómo saber si eres apto para trabajar en investigación.

¿Cómo Entrar en Este Campo?

El primer paso en este camino suele ser realizar un doctorado en aprendizaje automático en una buena universidad. Es posible entrar en este campo sin un doctorado, pero es probable que se requiera en puestos de investigación en centros académicos y DeepMind, que constituyen una gran parte de los mejores puestos. Un doctorado en aprendizaje automático también abre opciones en

política de IA, IA aplicada y ganar para dar, por lo que este camino tiene buenas opciones de respaldo si más tarde decides que la seguridad técnica de la IA no es para ti.

Sin embargo, si prefieres la ingeniería a la investigación, el doctorado no es necesario. En su lugar, puedes hacer un máster o formarte en la industria.

También es posible entrar en este camino desde la neurociencia (sobre todo la computacional), así que si ya tienes formación en esa área, quizá no tengas que volver a estudiar.

Si ya estás muy familiarizado con la seguridad de la IA como área problemática, nuestra principal recomendación es que eches un vistazo a esta guía paso a paso para seguir una carrera en seguridad técnica de la IA, de Charlie Rogers-Smith.

Recientemente, también se han abierto oportunidades para que los científicos sociales contribuyan a la seguridad de la IA.

Encontrarás más información en los recursos que se indican a continuación.

Organizaciones recomendadas

- AI Safety Support (Apoyo a la seguridad de la IA) trabaja para reducir los riesgos existenciales y catastróficos de la IA apoyando a todos los que quieran trabajar en este problema, centrándose en ayudar a los nuevos y aspirantes investigadores en seguridad de la IA mediante el asesoramiento profesional y la creación de comunidades.

- Alignment Research Center (Centro de investigación sobre Alineación) es una organización de investigación sin ánimo de lucro que trabaja para alinear los futuros sistemas de aprendizaje automático con los intereses humanos. Su labor actual se centra en el desarrollo de una estrategia de alineación "de extremo a extremo" que pueda adoptarse hoy en la industria y, al mismo tiempo, escalarse con elegancia a futuros sistemas de aprendizaje automático. Consulta las vacantes actuales.
- Anthropic es una empresa de investigación y seguridad en IA que trabaja para crear sistemas de IA fiables, interpretables y dirigibles. Los intereses de investigación de su equipo multidisciplinar incluyen el lenguaje natural, la retroalimentación humana, las leyes de escalado, el aprendizaje por refuerzo, la generación de código y la interpretabilidad. Consulta las vacantes actuales.
- El Center for Human-Compatible Artificial Intelligence (Centro de Inteligencia Artificial Humanamente Compatible) tiene como objetivo desarrollar los medios conceptuales y técnicos para reorientar el impulso general de la investigación en IA hacia sistemas probadamente beneficiosos. Consulta las vacantes actuales.
- El Center on Long-term Risk (Centro de Riesgos a Largo Plazo) aborda los peores riesgos derivados del desarrollo y despliegue de sistemas avanzados de IA. Actualmente, se centra en escenarios de conflicto, así como en aspectos técnicos y filosóficos de la cooperación. Su labor incluye la realización de investigaciones interdisciplinarias, la concesión y recomendación de subvenciones y la creación de una comunidad de profesionales y otros investigadores en torno a estas prioridades. Consulta las vacantes actuales.
- DeepMind es probablemente el mayor grupo de investigación que desarrolla inteligencia artificial general en el mundo occidental. Solo confiamos en recomendar puestos en DeepMind que trabajen

específicamente en cuestiones de seguridad, ética, política y protección. Consulta las vacantes actuales.

- El Future of Humanity Institute (Instituto para el Futuro de la Humanidad) es un instituto de investigación multidisciplinar de la Universidad de Oxford. Los académicos del FHI utilizan las herramientas de las matemáticas, la filosofía y las ciencias sociales para abordar cuestiones de gran calado sobre la humanidad y sus perspectivas.
- El Machine Intelligence Research Institute (Instituto de Investigación de Inteligencia de Máquinas) fue uno de los primeros grupos en preocuparse por los riesgos de la inteligencia artificial a principios de la década de 2000, y ha publicado varios trabajos sobre cuestiones de seguridad y cómo resolverlas. Consulta las vacantes actuales.
- OpenAI se fundó en 2015 con el objetivo de investigar cómo hacer que la IA sea segura. Ha recibido más de mil millones de dólares en compromisos de financiación de la comunidad tecnológica. Solo confiamos en recomendar oportunidades en sus equipos de política, seguridad y protección. Consulta las vacantes actuales.
- Redwood Research lleva a cabo investigación aplicada para ayudar a alinear los futuros sistemas de IA con los intereses humanos. Consulta las vacantes actuales.

¿Deseas asesoramiento personalizado para seguir este camino?

Dado que esta es una de nuestras trayectorias prioritarias, si crees que puede ser una buena opción para ti, estaremos encantados de asesorarle personalmente sobre los pasos a seguir. Podemos ayudarte a considerar tus opciones, ponerte en contacto con otras personas que trabajen en el mismo campo e incluso ayudarte a encontrar trabajo u oportunidades de financiación.

Encuentra Trabajo como Investigador en Seguridad de la IA

Si crees que podrías encajar bien en este camino y estás listo para empezar a buscar oportunidades de trabajo, consulta nuestra lista **curada** de trabajos abiertos en este camino:

Becario, Verano 2023

Investigación Redwood
San Francisco, Área de la Bahía
Publicado hace 3 días

Responsable de Seguridad

Anthropic
San Francisco, Área de la Bahía
Publicado hace 3 días

Director de Operaciones de Seguridad, Seguridad Global

DeepMind

Londres, Reino Unido

Publicado hace 3 días

VER TODAS LAS OPORTUNIDADES

Más información

Lecturas Complementarias Clave:

- Para ayudarte a orientarte en este campo, te recomendamos el pack de iniciación a la seguridad en IA.
- Guía paso a paso de Charlie Rogers-Smith sobre carreras relacionadas con la seguridad de la IA
- Nuestro perfil de problema sobre el riesgo de la IA
- Este plan de estudios sobre seguridad de la IA (o, para algo más breve, esta secuencia de posts de Richard Ngo)
- Nuestra guía para convertirse en un ingeniero AM centrado en la seguridad de la IA

Otras lecturas:

- Revisión de la carrera de doctorado en aprendizaje automático
- Nuestra revisión de la carrera de seguridad técnica de IA de 2015
- Lista de lecturas del Center for Human-Compatible AI (Centro de Inteligencia Artificial Humanamente Compatible)
- Una colección de listas de lectura sobre la seguridad de la IA
- Ver todos nuestros artículos sobre carreras relacionadas con la seguridad de la IA

Podcasts:

- El Dr. Paul Christiano explica cómo OpenAI está desarrollando soluciones reales al "problema de la alineación de la IA" y su visión de cómo la

humanidad irá cediendo progresivamente la toma de decisiones a los sistemas de IA.

- Ingeniería de aprendizaje automático para la seguridad y robustez de la IA: guía de un ingeniero de Google Brain para entrar en este campo
- El mundo necesita investigadores de IA. Cómo convertirte en uno de ellos
- Chris Olah habla sobre cómo trabajar en los mejores laboratorios de IA sin un título universitario y qué demonios ocurre dentro de las redes neuronales.
- Un investigador de alineación de aprendizaje automático explica cómo convertirse en tal.
- Richard Ngo sobre grandes modelos lingüísticos, OpenAI y el esfuerzo por hacer que el futuro vaya bien