Consejos para una carrera en seguridad de AGI

Richard Ngo
Traducción J.A Muñoz

Publicación en el foro AE el 2 de mayo 2023

A menudo, las personas me piden consejos sobre una carrera profesional relacionada con la seguridad de la AGI. Esta publicación resume los consejos que suelo dar más frecuentemente. Lo he dividido en tres secciones: mentalidad general, investigación de alineación y trabajo de gobernanza. Para cada una de las dos últimas, comienzo con un consejo de alto nivel dirigido principalmente a estudiantes y a aquellos que están en las primeras etapas de sus carreras, luego profundizo en más detalles del campo. Vean también esta publicación que escribí hace dos años, que contiene un montón de consejos de carrera bastante generales.

Mentalidad general

Para tener un gran impacto en el mundo, necesitas encontrar una gran palanca. Este documento asume que piensas, como yo, que la seguridad de la AGI (es la mayor de esas palancas). Sin embargo, hay muchas formas de accionar esa palanca, desde la investigación y la ingeniería hasta las operaciones y la construcción del campo, pasando por la política y las comunicaciones. Te animo a elegir entre estas opciones principalmente en base a tu adaptación personal, una combinación de lo que realmente eres bueno y lo que realmente disfrutas. En mi opinión, la diferencia entre ser una gran adaptación versus una adaptación mediocre anula otras diferencias en el impacto de la mayoría de los pares de trabajos relacionados con la seguridad de la AGI (Inteligencia Artificial General).

¿Cómo debes encontrar tu adaptación personal? Para empezar, debes centrarte en encontrar trabajo donde puedas obtener ciclos de retroalimentación rápidos. Eso normalmente implicará poner manos a la obra o hacer algún tipo de proyecto concreto (en lugar de simplemente leer y aprender) y ver cuán rápido puedes progresar. Eventualmente, una vez que hayas tenido una gran cantidad de experiencia, podrías notar un sentimiento de confusión o frustración: ¿por qué todos los demás no entienden la cuestión, o lo están haciendo tan mal ? (Aunque cabe mencionar que algunos investigadores de alto nivel comentaron en un borrador que ellos no tuvieron esta experiencia.) Para algunas personas eso implica investigar un tema específico (para mí, la pregunta "¿cuál es el mejor argumento de que la AGI será desalineada?"); para otros se trata de aplicar habilidades como la conciencia (por ejemplo, "¿por qué los demás no pueden simplemente pasar por todos los pasos obvios?"). Ser excelente rara vez se siente como si fueras excelente, porque tus propias habilidades establecen tu línea de base de lo que se siente normal.

¿Y si tienes esa experiencia para algo que no disfrutas haciendo? Espero que esto sea bastante raro, porque ser bueno en algo suele ser muy placentero. Pero en esos casos, sugeriría intentarlo hasta que observes que incluso una serie de éxitos no te emociona sobre lo que estás haciendo; y en ese punto, probablemente intenta cambiar (aunque esto depende bastante de los detalles específicos).

Por último: la seguridad de la AGI es un campo joven y pequeño; hay mucho por hacer y todavía muy pocas personas para hacerlo. Te animo a tener agencia cuando se trata de hacer que las cosas pasen y se hagan: la mayoría de las veces la respuesta a "¿por qué no está ocurriendo esta cosa aparentemente buena?" o "¿por qué no somos 10 veces mejores en esta cosa en particular?" es "porque nadie se ha ocupado de ello todavía". Y las calificaciones más importantes para poder resolver un problema son normalmente la capacidad de notarlo y la voluntad de intentarlo. Una anécdota para ayudar a ilustrar este punto: una amiga mía ha tenido cuatro trabajos en cuatro organizaciones de investigación de alineación de primer nivel; ninguno de esos trabajos existía antes de que ella se acercara a los grupos relevantes para sugerir que deberían contratar a alguien con su conjunto de habilidades. Y esto es solo lo que es posible *dentro* de las organizaciones existentes; si estás lanzando tu propio proyecto, hay muchas más oportunidades para hacer cosas totalmente novedosas. (La principal excepción es cuando se trata de divulgación y recomendaciones políticas. La alineación es un campo inusual porque la base de fans y seguidores es mucho mayor que el número de investigadores, y por lo tanto debemos tener cuidado de evitar que el discurso de alineación sea dominado por defensores que tienen poca familiaridad con los detalles técnicos, y parecen demasiado seguros de sí mismos. Puedes ver la discusión aquí para leer más sobre esto.)

Investigación de alineación

Comenzaré con algunas recomendaciones de alto nivel, luego daré una breve descripción de cómo veo el campo.

1. **La alineación está limitada por la mentoría.** Si tienes poca experiencia en investigación, tu principal prioridad debería ser encontrar al mejor mentor posible para ayudarte a adquirir habilidades de investigación, por ejemplo, a través de la realización de investigaciones en el laboratorio de un profesor, o pasantías en laboratorios de IA. La mayoría de los mejores investigadores y mentores aún no están trabajando en alineación, por lo que la mejor opción para la mentoría puede estar fuera de la alineación, pero los doctorados son lo suficientemente largos, y los plazos de tiempo lo suficientemente cortos. Debes asegurarte de que tu mentor esté emocionado de supervisar algún tipo de investigación relevante a la alineación. Ocasionalmente, las personas pueden comenzar a hacer un gran trabajo sin ninguna mentoría; si esto te emociona, no dudes en intentarlo, pero céntrate en los tipos de investigación donde tienes ciclos de retroalimentación rápidos.

2. **Necesitarás involucrarte activamente.** La mejor investigación en ML y alineación se involucra intensamente con las redes neuronales (con solo algunas excepciones). Incluso si eres más teórico, debes planear interactuar regularmente con los modelos y adquirir las habilidades de codificación relevantes. En particular, veo a muchos investigadores junior que quieren hacer "investigación conceptual". Pero deberías asumir que dicha investigación es inútil hasta que se concreta en la escritura de código o la demostración de teoremas, y que necesitarás concretarla tú mismo (siendo el modelado de amenazas la principal excepción, ya que fuerza un tipo diferente de concreción). Quizás una vez que seas un investigador senior con intuiciones obtenidas de la experiencia práctica, podrás dar un paso atrás y pensar principalmente en soluciones potenciales a un nivel alto, pero ese no puede ser tu plan como investigador junior, predeciblemente te alejará de hacer un trabajo útil.

3. **Puedes empezar rápidamente**. Las personas que vienen de campos como la física y las matemáticas a menudo no se dan cuenta de cuánto más superficial es el aprendizaje profundo como campo, y por lo tanto piensan que necesitan pasar mucho tiempo comprendiendo primero los fundamentos teóricos. No es necesario, puedes comenzar a hacer investigación en aprendizaje profundo con nada más que las matemáticas de primer año de pregrado, y aprender las cosas que te faltan a medida que avanzas. (Aunque la habilidad de codificación es un requisito previo mucho más importante.) También puedes aprender muchas de las bases conceptuales de la alineación a medida que avanzas, especialmente en roles más pesados en ingeniería. Si bien recomiendo que todos los investigadores de alineación eventualmente se familiaricen con las ideas cubiertas en el currículo de Fundamentos de Alineación, mejorar las habilidades en investigación empírica debería ser una prioridad mayor para la mayoría de las personas que ya han decidido seguir una carrera en investigación de alineación y que aún no son investigadores de ML.
   Algunas formas recomendadas de mejorar habilidades en investigación empírica (aproximadamente en orden)
   a. MLAB
   b. ARENA°
   c. Jacob Hilton's deep learning curriculum
   d. Neel Nanda's guide to getting started with mechanistic interpretability
   e. Replicar artículos. Cada uno de estos te enseña habilidades importantes para una buena investigación: cómo implementar algoritmos, cómo depurar código y experimentos, cómo interpretar resultados, etc. Una vez que hayas implementado un algoritmo o replicado un artículo, puedes intentar extender los resultados mejorando de alguna manera las técnicas.

4. **La mayoría de las investigaciones no tendrán éxito.** Esto es cierto tanto a nivel de proyectos individuales, como a nivel de todas las direcciones de investigación: la investigación es un dominio muy pesado . Deberías estar buscando arduamente las intuiciones centrales sobre por qué una dirección de investigación en particular tendrá éxito, cuya ausencia puede estar oculta bajo matemáticas o algoritmos complicados (como argumento aquí). (Puedes pensar en esto como un tipo de investigación conceptual, pero destinado a dirigir tu propio trabajo empírico o teórico, en lugar de estar destinado como un resultado de investigación en sí mismo.) En la siguiente sección expongo algunas de mis opiniones sobre cuáles direcciones de investigación son y no son prometedoras.

Direcciones de investigación en alineación

Desde mi punto de vista, la investigación de alineación más prometedora se divide en tres categorías principales. Las describo a continuación, así como tres categorías secundarias que considero valiosas. Ten en cuenta que espero que los límites entre todas estas se difuminen con el tiempo a medida que la investigación sobre ellas avance, y a medida que automatizamos más y más cosas.

1. **Supervisión escalable:** encontrar formas de aprovechar modelos más potentes para producir mejores señales de recompensa. La investigación de supervisión escalable puede ser particularmente de alto impacto si termina siendo adoptada ampliamente, por ejemplo, como una herramienta para prevenir alucinaciones (como el trabajo de los equipos de alineación en RLHF que ahora ha sido adoptado muy ampliamente).

a. El artículo teórico al que más a menudo dirijo a las personas es el artículo de debate de Irving et al.
b. El artículo empírico al que más a menudo dirijo a las personas es el artículo de críticas de Saunders et al., que puede verse como el caso más simple del algoritmo de debate; Bowman et al. (2022) también es útil desde una perspectiva metodológica.
c. Los otros dos algoritmos conocidos en esta área son la amplificación iterada y el modelado de recompensas recursivas. En mi opinión, las **personas a menudo sobreestiman las diferencias entre estos algoritmos,** y las presentaciones estándar de ellos oscurecen las formas en las que son estructuralmente similares. Personalmente, encuentro el debate el más fácil de razonar (y parece que otros están de acuerdo, ya que más trabajos se basan en él que en los demás), por lo tanto, por eso recomiendo a menudo que las personas trabajen en eso.
d. ¿La supervisión escalable solo llevará a más avances en capacidades? Esta es una pregunta importante; una forma en que lo pienso es en términos de la brecha generador-discriminador-crítica del artículo de críticas de Saunders y sus colegas. Específicamente, mientras espero que cerrar la brecha generador-discriminador sea un avance de doble propósito (y podría ser bueno o malo dependiendo de tus otras opiniones), cerrar la brecha discriminador-crítica al producir explicaciones correctas comprensibles para humanos definitivamente debería ser visto como un avance de alineación.

2. **Interpretabilidad mecanicista:** encontrar formas de entender cómo funcionan internamente las redes. Aunque todavía es un subcampo pequeño de ML, lo veo como una forma de empujar todo el campo de ML desde una perspectiva "conductista" que solo se centra en las entradas y salidas hacia un marco "cognitivista" que estudia lo que está sucediendo dentro de las redes neuronales. También es mucho más fácil de hacer fuera de los laboratorios de la industria que el trabajo de supervisión escalable. Para comenzar, consulta los 200 problemas concretos abiertos en interpretabilidad mecanicista de Nanda.

a. Tres conceptos de trabajo en la interpretabilidad mecanicista

i. Estudios de caso: encontrar algoritmos dentro de las redes que implementen capacidades específicas. Mis trabajos favoritos aquí son Olsson et al. (2022), Nanda et al. (2023), Wang et al. (2022) y Li et al. (2022); estoy emocionado de ver más trabajos que se basen en el último en particular para encontrar modelos del mundo y metas representadas internamente dentro de las redes.

ii. Resolviendo la superposición: encontrar formas de entrenar redes para tener menos conceptos superpuestos dentro de las neuronas individuales. El recurso clave aquí es Elhage (2022) (así como otros trabajos en el hilo de Circuitos del Transformador).

iii. Interpretabilidad escalable: encontrar algoritmos para identificar o modificar automáticamente representaciones internas. Mis trabajos favoritos: Meng et al. (2022) y Burns et al. (2023) (aunque algunos consideran que este último está más cerca del trabajo de supervisión escalable).

3. **Teoría de alineación:** encontrar marcos formales que podamos usar para razonar acerca de la IA avanzada. Quiero señalar que el éxito en este tipo de investigación es aún más pesado en la cola que las otras direcciones de investigación que he descrito: parece requerir habilidades matemáticas excepcionales, una comprensión profunda de la teoría de ML y intuiciones filosóficas matizadas. No soy optimista de que alguna de las direcciones de investigación que se enumeran aquí vaya a funcionar, pero están intentando abordar problemas tan fundamentales que incluso los éxitos parciales podrían ser algo importante.

a. Estoy especialmente emocionado por el trabajo de Christiano sobre la formalización de argumentos heurísticos, la agenda teórica del aprendizaje de Kosoy (en particular la infra-bayesianismo) y varios trabajos de Scott Garrabrant (por ejemplo, racionalidad geométrica, conjuntos factorizados finitos y marcos cartesianos).
b. Históricamente, la mayoría del trabajo en esta categoría ha sido realizado por MIRI (por ejemplo, trabajo en teoría de decisión funcional e inducción de Garrabrant). Sin embargo, su producción ha disminuido significativamente últimamente; por lo tanto, principalmente los considero como un grupo de investigadores persiguiendo sus intereses individuales, en lugar de una agenda de investigación unificada.
c. ¿Por qué creo que vale la pena seguir la teoría de alineación? En gran parte porque el conocimiento científico suele estar muy interconectado. La teoría de alineación a menudo parece desconectada del ML moderno, pero los movimientos de las estrellas alguna vez parecieron totalmente desconectados de los eventos en la tierra. ¿Y quién podría haber adivinado que entender la variación en los picos de los pinzones avanzaría nuestra comprensión de... bueno, básicamente todo en la biología? En muchos dominios existen principios clave que explican una enorme variedad de fenómenos, y la principal dificultad radica en encontrar un ángulo de ataque tratable. Por eso, hacer las preguntas correctas suele ser más importante que obtener resultados concretos. Por ejemplo, preguntar "¿cuál es la estrategia óptima en esta formalización específica de un juego de 2 jugadores?" representa una gran parte del trabajo de inventar la teoría de juegos.

Tres otras áreas de investigación que parecen importantes, pero menos centrales:

1. Evaluaciones: encontrar formas de medir qué tan peligrosos y/o desalineados están los modelos.

    a. Hasta ahora se ha publicado poco sobre esto; lo principal a observar son las evaluaciones ARC (también discutidas en la sección 2.9 de la tarjeta del sistema GPT-4). En general, parece que las evaluaciones de alineación son muy difíciles, por lo que la mayoría de las personas se están enfocando en evaluaciones para medir capacidades peligrosas.

    b. Mi propia opinión es que las evaluaciones vivirán o morirán dependiendo de cuán simples y escalables sean. Las mejores evaluaciones serían fácilmente implementables incluso por personas sin ningún conocimiento en alineación, y seguirían significativamente las mejoras desde los sistemas actuales hasta las superinteligencias. En resumen, esto se debe a que el propósito principal de las evaluaciones es facilitar la toma de decisiones y la coordinación, y ambos se benefician enormemente de métricas legibles y predecibles.

2. Entrenamiento adversarial sin restricciones: encontrar formas de generar entradas en las que los sistemas desalineados se comportarán incorrectamente.

    a. Parece que hay razones principiadas fuertes para esperar que esto sea difícil: en general, solo se puede generar datos falsos que engañen a un modelo usando un modelo mucho más poderoso. Pero puede ser posible encontrar ejemplos adversarios sin restricciones aprovechando la interpretabilidad mecanicista, como se exploró en esta publicación de Christiano.
    b. El documento empírico al que más frecuentemente dirijo a las personas es Ziegler et al. (2022) (vease también los otros documentos que citan).

3. Modelado de amenazas: comprendiendo y previendo cómo la IA General (AGI) podría conducir a resultados catastróficos.

    a. A menudo, dirijo a las personas a mi propio artículo reciente (Ngo et al., 2022). Otros trabajos buenos incluyen informes de Joe Carlsmith y Ajeya Cotra. (Cohen et al. (2022) hacen un caso revisado por pares sobre el riesgo existencial de la AGI, pero está demasiado enfocado en la alineación externa para que yo lo acepte.)

    b. Una dirección de investigación de modelado de amenazas que parece valiosa es comprender el hacking de gradientes (y entender la cooperación entre diferentes modelos en general). Otra es explorar las formas específicas en que los AGI son más propensos a ser desplegados en el mundo real, y qué tipos de vulnerabilidades podrían ser capaces de explotar.

Por el contrario, algunas líneas de investigación que creo que están sobrevaloradas por muchos recién llegados al campo, junto con algunas críticas a ellas:

1. Aprendizaje por refuerzo inverso cooperativo (la dirección que Stuart Russell defiende en su libro Compatible con Humanos); críticas aquí y aquí.

2. El trabajo de John Wentworth sobre abstracciones naturales; exposición y crítica aquí, y otra aquí.

3. Trabajos que dependen de que los agentes actúen de forma miope, incluyendo solo hacer predicciones de paso-tiempo siguiente (por ejemplo, trabajos sobre la abstracción de simuladores, o sobre condicionar modelos predictivos); crítica aquí.

Trabajo de gobernanza

Los he divido mentalmente en tres categorías: investigación de gobernanza, gobernanza de laboratorios y trabajos de política. Estas son algunas conclusiones de alto nivel para cada una:

1. Investigación de gobernanza
    a. El principal consejo que doy a las personas que quieren entrar en este campo: **elige un tema relevante e intenta convertirte en un experto en él**. Hay alrededor de dos docenas de temas en los que desearía que existiera un experto mundial en la aplicación de este tema para hacer que la IA de AG sea beneficioso, y tal persona no existe; he hecho una lista de esos temas a continuación. Para aprender sobre ellos, recomiendo encarecidamente no solo leer y absorber ideas, sino también escribir sobre ellas. Es muy plausible que, comenzando sin ningún conocimiento en el campo, en seis meses podrías

escribir una publicación o un artículo que avance la frontera de nuestro conocimiento sobre cómo uno de esos temas es relevante para la gobernanza de la IA de AG.

b. No necesariamente necesitas mantenerte en lo que elijas a largo plazo; mi afirmación principal es que es importante tener *algún* tema concreto para investigar. A medida que lo hagas, gradualmente te ramificarás a otros temas que son tangencialmente relevantes y adquirirás un conocimiento más amplio del campo (el curso de Fundamentos de Gobernanza es una buena manera de hacerlo). Eventualmente, podrás hacer "investigación estratégica" con implicaciones mucho más amplias. Pero tratar de hacer eso desde el principio es un mal plan: te irá mucho mejor con una base de experiencia detallada desde la que trabajar.

c. En general, creo que la gente **sobrevalora el "análisis" y subestima las "propuestas"**. Hay muchos factores de alto nivel que afectarán la gobernanza de la AGI, y podríamos pasar el resto de nuestras vidas tratando de analizarlos. Pero, en última instancia, lo que necesitamos son mecanismos concretos que realmente muevan la aguja, los cuales actualmente son escasos. Por supuesto, necesitas hacer análisis para entender los factores que influirán en el éxito de las propuestas, pero siempre debes tener en mente el objetivo de intentar materializarlo en algo útill.

d. En relación con esto, personalmente no creo que el modelado cuantitativo sea muy valioso. Aún no he visto tal modelo de una pregunta de gran envergadura (por ejemplo, proyecciones de cálculo, velocidad de despegue, cronogramas) cuyas conclusiones cambien sustancialmente mis opiniones sobre cuáles son las mejores propuestas de gobernanza. Si tal modelo tiene un gran éxito, puede cambiar mis creencias de, digamos, un 25% a un 75% en una proposición dada. Pero esa es solo una diferencia de factor 3, mientras que un plan para resolver la gobernanza podría ser uno o dos órdenes de magnitud más efectivo que otro. Y en general, los modelos rara vez me mueven tanto, porque incluso unos pocos parámetros libres permiten a las personas sobreajustar dramáticamente a sus intuiciones; típicamente preferiría tener un breve resumen de las principales ideas que la persona que hace el modelado aprendió durante ese proceso. Así que prioriza los planes primero, las ideas en segundo lugar y los modelos en último lugar.

e. No te limites demasiado por la viabilidad política, especialmente al formular las primeras versiones de un plan. Casi nadie en el mundo tiene buenas intuiciones sobre cómo funciona realmente la política, y buenas intuiciones sobre cuán loco será el progreso hacia la AGI. Todo tipo de posibilidades se abrirán en el futuro, solo necesitamos estar listos con propuestas concretas cuando lo hagan. Sin embargo, una comprensión profunda de los impulsores fundamentales de las decisiones políticas de hoy será útil para navegar cuando las cosas comiencen a cambiar mucho más rápido.

2. Gobernanza de laboratorios de IA
    a. Los laboratorios líderes a menudo están dispuestos a llevar a cabo propuestas que no comprometen mucho su trabajo principal de capacidades; el cuello de botella suele ser la agencia y el trabajo requerido para implementar realmente la propuesta. Por lo tanto, las intervenciones del tipo "dile a los laboratorios que se preocupen más por la seguridad" generalmente no funcionan muy bien, mientras que las intervenciones del tipo "aquí hay una petición concreta, estos son los pasos específicos que tendrías que tomar, aquí hay una persona que ha acordado liderar el esfuerzo" tienden a ir bien. Esta publicación transmite esa idea particularmente bien.

    b. Es difícil para las personas fuera de los laboratorios conocer suficientes detalles sobre lo que está sucediendo dentro de los laboratorios para poder hacer propuestas concretas, pero espero que haya algunos casos importantes en los que sea posible. Esto probablemente se parezca bastante al camino que esbocé en la sección sobre

investigación de gobernanza, primero adquiriendo experiencia en un tema específico, luego generando propuestas específicas.

c. Hay una habilidad específica para lograr que las cosas se hagan dentro de las organizaciones grandes que la mayoría de los AEs carecen (debido a la falta de experiencia corporativa, además de la falta de orientación hacia las personas), pero que es particularmente útil al impulsar propuestas de gobernanza de laboratorio. Si la tienes, el trabajo de gobernanza de laboratorio puede ser adecuado para ti.

3. Trabajos relacionados con políticas
    a. Trabajos relacionados con políticas pueden implicar trabajar en posiciones relacionadas con el gobierno, con el objetivo de intentar ocupar un puesto en el que puedas contribuir a una regulación gubernamental efectiva. Aunque no es mi área de especialización, puedo ofrecerte algunos consejos generales para tener una carrera exitosa en este campo y cómo acelerar tu progreso considerando que las personas cada vez están más preocupadas por la inteligencia artificial. Obtener una maestría corta y participar en programas de becas en política son formas rápidas de avanzar hacia roles de políticas de nivel medio en menos tiempo. Además, adquirir aunque sea un mínimo nivel de experiencia legible en inteligencia artificial (por ejemplo, cualquier título o empleo relacionado con ciencias de la computación/inteligencia artificial) también es útil.

Lista de temas de gobernanza

Aquí hay algunos temas en los que desearía contar con un experto mundial en para aplicarlo a a la seguridad de la IA generalizada (AGI). Un ejemplo de cómo sería un trabajo destacado en uno de estos temas es el artículo de Baker sobre lecciones del control de armas nucleares (un tema que habría estado en esta lista si el no hubiera escrito eso).

Un conjunto de temas puede describirse de manera general como "cualquier cosa mencionada en el artículo de Yonadav Shavit sobre gobernanza informática", en particular:

1. Registro a prueba de manipulaciones en las GPU
2. Seguimiento global de las GPU
3. Algoritmos de prueba de aprendizaje
4. inspecciones in situ de los modelos
5. Detección de centros de datos
6. Construcción de un conjunto para inferencia verificable
7. Medición del uso efectivo de cómputo (por ejemplo, mediante la medición y el control del progreso algorítmico)
8. Regulación del entrenamiento descentralizado a gran escala (si se vuelve competitivo con el entrenamiento centralizado)

Otro conjunto de temas relacionados con la seguridad, como:

1. Prevenir la extracción de parametros de redes neuronales (por parte de terceros o por parte de la propia IA)
2. Evaluar la posibilidad de replicación autónoma a través de Internet
3. Ascenso de privilegios desde sistemas seguros (por ejemplo, si tu asistente de codificación no está alineado, ¿qué podría lograr?)

4. Monitoreo de centros de datos (por ejemplo, si se estuvieran ejecutando copias no autorizadas de un modelo en tus servidores, ¿cómo lo sabrías?)
5. Detectar canales de comunicación no autorizados entre diferentes copias de un modelo
6. Detectar manipulaciones (por ejemplo, si se hubiera modificado tu ejecución de entrenamiento, ¿cómo lo sabrías?)
7. ¿Qué tan vulnerables son los sistemas de comando y control nuclear?
8. Monitoreo de comportamiento escalable (por ejemplo, ¿cómo podemos agregar información de registros de monitoreo de millones de IA?)

Y una tercera categoría más miscelánea (y menos técnica):

1. ¿Qué aparato regulatorio dentro del gobierno de los Estados Unidos sería más efectivo para regular las grandes ejecuciones de entrenamiento?
2. ¿Qué herramientas y métodos tiene el gobierno de los Estados Unidos para auditar a las empresas de tecnología?
3. ¿Cuáles son las mayores brechas en los controles de exportación de los Estados Unidos a China, y cómo podrían cerrarse?
4. ¿A qué aplicaciones o demostraciones de IA reaccionará más fuertemente la sociedad?
5. ¿Qué interfaces utilizarán los humanos para interactuar con las IA en el futuro?
6. ¿Cómo se desplegará más probablemente la IA para tareas sensibles (por ejemplo, asesorar a líderes mundiales) dadas las preocupaciones sobre la privacidad?
7. ¿Cómo podría polarizarse el discurso político en torno a la IA, y qué podría mitigar eso?
8. ¿Qué se necesitaría para automatizar infraestructuras cruciales (fábricas, armas, etc)?