El rol de la IA en la sociedad

Sofía Trejo

El futuro

La IA tiene el potencial de mejorar el mundo en términos de la medicina, el transporte, el uso de energía, educación, crecimiento económico y sustentabilidad. Sin embargo, esta misma tecnología posee la capacidad de exacerbar desigualdades, de crear una economía **oligopolista** (cuando unas cuantas compañías controlan el mercado) a nivel mundial, de reforzar el totalitarismo y de atentar en contra de los derechos humanos. Por ello es vital crear estrategias políticas y tecnológicas para garantizar que futuros avances en el área sean seguros y enfocados al bien común.

En esta sesión estudiaremos algunas de las preguntas y problemas de investigación que intentan encauzar el desarrollo de IA hacia un camino que beneficie a la humanidad. Para ello estudiaremos dos áreas:

- **Seguridad de IA** (AI Safety): se enfoca en la parte técnica sobre cómo construir IA.
- **Gobernanza de IA** (AI governance): entender el contexto y las instituciones donde las IA son construidas e implementadas. Intenta maximizar la probabilidad de que las personas construyendo dichos sistemas tengan los incentivos, metas, apoyo, tiempo, recursos etc. para desarrollar tecnologías que beneficien a la humanidad.

Seguridad de IA

Estos son algunos de los problemas en esta área son los siguientes

Inteligibilidad:

Poder interpretar y garantizar transparencia de los sistemas, como bajo la reducción de dimensiones. Entender qué características están siendo codificadas por los sistemas.

Especificación de valores:

Formalizar principios éticos, aprender preferencias humanas, medir y minimizar efectos extremos no deseados.

Garantías de seguridad y de acción:

Definir las probabilidades de comportamientos inseguros y restricciones en los mecanismos de exploración.

Ejemplo de problemática

En el 2002 se publicó el trabajo *The Evolved Radio and its Implications for Modelling the Evolution of Novel Sensors.*[1] En este trabajo se estudia la evolución de sensores que utilizan algoritmos evolutivos. En particular describe un experimento de hardware donde una red de transistores detectaron y utilizaron las ondas de radio que emanaban de una PC cercana.

Instituciones trabajando en Seguridad de IA

- Machine Intelligence Research Institute MIRI
- Future of Life Institute FLI
- Future of Humanity Institute FHI
- UC Berkeley
- Australian National University
- Centre For Human Compatible AI CHAI
- Google DeepMind
- Open AI

[1] Bird, Jon and Paul Layzell. “The Evolved Radio and its Implications for Modelling the Evolution of Novel Sensors.” Proceedings of the 2002 Congress on Evolutionary Computation, CEC'02 (Cat. No.02TH8600), 2002.

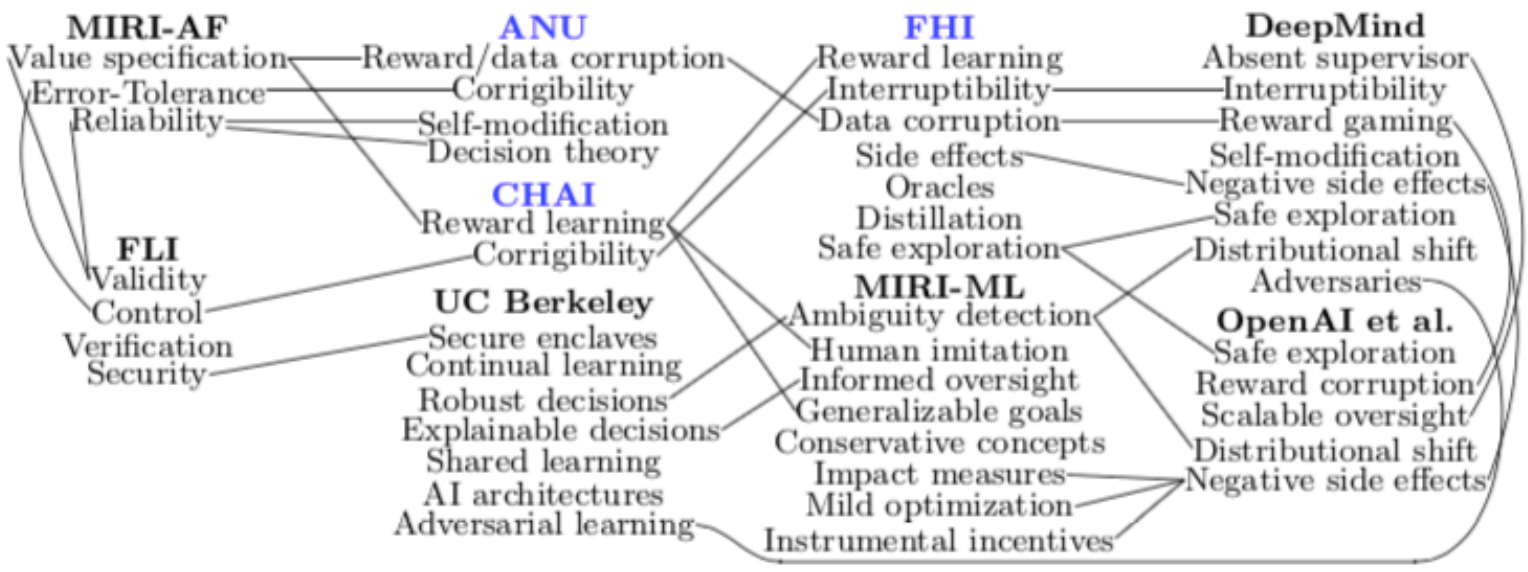

Figure 1: Connections between problems stated in different AGI safety research agendas (for ANU, CHAI, and FHI, the agendas are inferred from their recent publications).

Literatura sugerida

- T. Everitt, G. Lea, M. Hutter, *AGI Safety Literature Review*, Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence Survey track. (2018).
- D Amodei, C Olah, J Steinhardt, P Christiano, J Schulman, D Mané, *Concrete problems in AI safety,* arXiv (2016).

Gobernanza de IA

¿Qué rol jugará las IA en la seguridad internacional?; ¿la IA cambiará el panorama mundial en términos de armamento, inteligencia militar y estrategia?; ¿el desarrollo de IA generará nuevos medios de cooperación y mediación? Estas son algunas de las preguntas que debemos responder para garantizar que el desarrollo de IA no cause daños a la humanidad.

Dinámica de carrera

Es necesario investigar las circunstancias que podrían transformar el desarrollo de IA en una carrera armamentista. Por ejemplo es posible que si la IA se convierte en una herramienta militar crítica diversos actores intentarán controlarla y mantenerla en secreto.

Uno de los mayores peligros de que el desarrollo de tecnología se convierta en una carrera es que esta dinámica, muy probablemente, llevará a se tomen atajos y se descuiden factores fundamentales de seguridad.

¿Qué se puede hacer para prevenir esta dinámica? ¿Legislaciones o tratados? ¿Regulación de estándares de producción

Desigualdad, pérdida de empleos y redistribución del capital

Es probable que el desarrollo de IA incrementará las desigualdades sociales y entre países. Esta tecnología podría no sólo podrá automatizar trabajos de manufactura, sino que podría reemplazar a trabajadores de clase media,quienes terminarían trabajando por menos salario (siguiendo los esquemas actuales).
La IA parece estar generando nuevos monopolios internacionales. La presencia de dichos monopolios reduce la distribución del ingreso.
Existen propuestas de establecer un salario universal. ¿Es esto algo factible a nivel mundial?

Violación de derechos humanos

Actualmente varias IA están atentando en contra de garantías individuales. Por ello, varias organizaciones, como Amnistía Internacional, abdican por que los derechos humanos sean colocados al centro de las políticas en el área.

Los derechos humanos son una buena base y un buen marco legal para la política de IA, ya que son reconocidos internacionalmente y el costo en reputación para quienes los violan es muy alto.

No discriminación e igualdad:

Sistemas como el del Sistema de Crédito Social que será implementado en China evaluará y categoriza a las personas. Restringido su libertad (compra de trenes y vuelos) e influenciará toda su vida.

Participación política

El uso de IA abre oportunidades para socavar la democracia como no se había visto antes. Control de información, noticias falsas, manipulación de la realidad son usadas con esa finalidad.

Ejemplos de este tipo de problemática fueron los chatbots Rusos usados en Twitter y en FB para influenciar las elecciones presidenciales en EUA en el 2016. Twitter aceptó que más de 50,000 cuentas de Twitter asociadas con chatbots Rusos postearon contenido de manera automática durante dicho periodo.[2]

Privacidad

La privacidad en IA debería ser considerada como un derecho fundamental, en lugar de una preferencia ética. La privacidad es clave para garantizar otros derechos como la libertad de expresión, asociación, participación política e información.

Científicos de la Universidad de Stanford publicaron en el 2018 un estudio donde entrenaron redes neuronales para detectar la orientación sexual de las personas haciendo uso de imágenes sin consentimiento[3].

Libertad de Expresion

En el 2014 investigadores de Cornell colaboraron con FB en un estudio de contagio de emociones, examinando como emociones se propagan en redes sociales. Los investigadores manipularon la experience de 700,000 usuarios usando análisis de sentimientos para identificar posts y comentarios negativos. Los post negativos eran removidos del feed de los usuarios para investigar si alterando el feed de manera algorítmica los usuarios permanecerán más tiempo en el sitio[4] Esto muestra como las plataformas alteran el mundo para que perezca de una manera particular, aumentando y acallando realidades.

---

2 https://www.theguardian.com/technology/2018/jan/19/twitter-admits-far-more-russian-bots-posted-on-election-than-it-had-disclosed

3 https://datasociety.net/wp-content/uploads/2018/05/AI-Systems-and-Research-Revealing-Sexual-Orientation_Case-Study_Final_CC.pdf

4 Adam D. I. Kramer, et al., “Experimental Evidence of Massive-Scale Emotional Conta- gion Through Social Networks,” PNAS vol. 111 no. 24 (2014), http://www.pnas.org/content/ pnas/111/24/8788.full.pdf.

Instituciones trabajando en Gobernanza de IA

- Future of Humanity Institute
- Future of Life Institute
- Data and Society

Literatura sugerida

- M. Latonero, *Governing Artificial Intelligence: Upholding Human Rights & Dignity*, Data and Society (2018).
- *AI Policy Challenges And Recommendations*, Future of Life Institute website https://futureoflife.org/ai-policy-challenges-and-recommendations#Safety
- *How to Prevent Discriminatory Outcomes in Machine Learning*, World Economic Forum, (2018).
- *The Toronto Declaration: Protecting the Rights to Equality and Non-Discrimination in Machine Learning systems,* (2018).
- *Human Rights and Technology Issues Paper,* Australian Human Rights Commission, (2018).

Acciones que se pueden tomar ahora

Firmar de los principios de Asilomar

Principios bajo los cuales se debe desarrollar IA:

- Evitar dinamica de carrera
- Valores humanos
- Privacidad
- Bien común

Se encuentran en el site de The Future of Life Institute.[5]

Firmar petición para la prohibición de armas letales autónomas

Se encuentran en el site de The Future of Life Institute.[6]

---

[5] https://futureoflife.org/ai-principles/
[6] https://futureoflife.org/lethal-autonomous-weapons-pledge/

Bibliografía

- *AI Policy Challenges And Recommendations*, Future of Life Institute website, https://futureoflife.org/ai-policy-challenges-and-recommendations#Safety
- M. Latonero, *Governing Artificial Intelligence: Upholding Human Rights & Dignity*, Data and Society (2018).
- T. Everitt, G. Lea, and M. Hutter, *AGI Safety Literature Review*, In International Joint Conference on AI (IJCAI) & arXiv (2018).
- S. Russell, D. Dewey, M. Tegmark, *Research Priorities for Robust and Beneficial Artificial Intelligence,* Ai Magazine (2015), https://futureoflife.org/data/documents/research_priorities.pdf