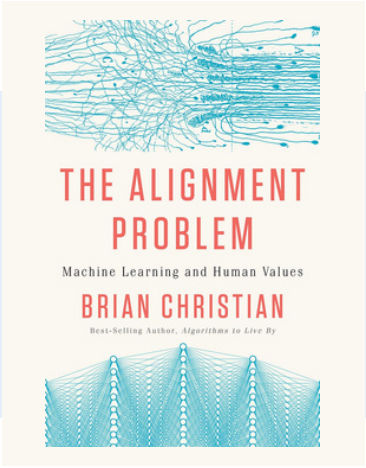

The Alignment Problem

Machine Learning and Human Values

Autor: Brian Christian

No se trata de un libro específico sobre gestión de proyectos, pero su manera de tratar el tema del alineamiento, así como problemas éticos y de seguridad, usa **ejemplos reales** que pueden servir de guía a **desarrolladores, gestores y emprendedores.**

Aunque es un libro valioso, algunos lo ven **«incompleto»** dadas las ramificaciones más amplias del problema del alineamiento. Otros libros como *Superinteligencia*, de Nick Bostrom, abordan ese tema con más profundidad.

¿Qué aporta?

Un repaso minucioso pero expresado de forma clara y divulgativa sobre la evolución de la inteligencia artificial **hasta 2020.** Resume cientos de entrevistas y conversaciones a lo largo de **4 años de investigación.**

Muchos lo consideran una lectura esencial. Aborda el tema del alineamiento con **casos reales y definiciones** que ayudan a entender el funcionamiento de estos modelos, y los problemas de **seguridad** y **éticos** que su uso y entrenamiento entrañan.

Ideas clave

La historia de la inteligencia artificial hace converger ramas del conocimiento desde las **matemáticas** y la **física** hasta la **psicología** y la **filosofía ética.**

Queda mucho por entender sobre **la mente humana.** La creación de máquinas que tratan de imitar sus funciones ha llevado a descubrimientos importantes sobre ésta, así como sobre **sesgos** individuales y culturales.

La evolución de esta tecnología es una historia accidentada, llena de serendipias, idealismo y casos reales que llaman a la **precaución.** Existe el riesgo de que un mal desarrollo cree profecías de cumplimiento inducido a gran escala, con su origen en modelos mal planteados. O de crear una sociedad que descuide y **pierda todo lo no quepa en un modelo;** que no sea cuantificable.

En el campo de la IA, los expertos advierten: no es suficiente con saber que funciona, sino que **hay que saber cómo funciona.**

Temas

Inteligencia artificial

Comprar libro

En inglés: Amazon

En español: no disponible

Formación

IA responsable para gestión ágil

Libros relacionados

Repaso histórico

Sesgos y transparencia

El libro elabora la evolución desde los perceptrones hasta las redes neuronales y modelos capaces de imitar lenguaje natural. Se explica su funcionamiento y por qué presentan sesgos importantes, reflejo de las características de sus datos de entrenamiento. Usar sets que representen de manera integral todos los escenarios relevantes es clave, pero no es una tarea trivial ni fácil.

Se presentan casos de uso de sistemas de IA que han puesto en relieve su falibilidad, así como serios problemas éticos y de seguridad. Entre ellos:

- Sistemas como COMPAS, para informar las decisiones de jueces y juntas de libertad condicional en algunos estados de EE.UU., perpetúan sesgos presentes en el cuerpo de policía y datos históricos.
- Un sistema de IA para el triaje en un hospital clasificaba pacientes asmáticos con síntomas de neumonía como «bajo riesgo», ya que al recibir tratamientos agresivos inmediatamente tras llegar al hospital, todos sobreviven.

Ante estos desafíos, el campo de la interpretabilidad de modelos ha ganado importancia, con soluciones como:

- **Técnicas de saliencia:** para visualizar qué partes de los datos de entrada influyen más en las decisiones.
- **Aprendizaje multitarea:** usar datos de salida adicionales y dar múltiples predicciones en lugar de una, para verificar que el modelo funciona como se espera.
- **Deconvolución:** para clarificar las operaciones dentro de las capas intermedias de redes neuronales.
- ***Concept Activation Vectors*** **(CAVs):** interpretar redes neuronales usando modelos de ML simples, para comprobar la importancia de ciertos conceptos en las decisiones.

Programando agentes

Más allá del procesamiento de lenguaje, está el desafío de programar máquinas capaces de llevar a cabo todo tipo de tareas de forma segura. Para ello, se estudia cómo condicionar su comportamiento, y con qué incentivos.

Aprendizaje por refuerzo (RL)
Este tipo de aprendizaje *(reinforcement learning)* se inspira en estudios sobre el comportamiento animal y el papel de la dopamina en la motivación, y es una alternativa al aprendizaje supervisado y no supervisado. Los sistemas entrenados por RL pueden **aprender y adaptarse** en tiempo real.

Un elemento crucial en la evolución de RL es el **«modelado»** ***(shaping):*** el refuerzo progresivo de comportamientos cada vez más cercanos al deseado. Sin embargo, uno de los desafíos del RL es determinar los incentivos correctos; identificar qué comportamientos reforzar.

Por ejemplo, se ha observado que puede ser más efectivo reforzar estados en lugar de acciones específicas (que el suelo esté limpio, en lugar del acto de limpiar).

Otro méodo eficaz es emular la «motivación intrínseca». Al igual que humanos y animales, que a menudo actúan motivados por el deseo de novedad, sorpresa y por curiosidad, los modelos de RL pueden **programarse para «buscar novedad».** Un caso son los sistemas que utilizan estrategias **«epsilon greedy»,** que alternan entre explorar nuevas posibilidades de forma aleatoria y explotar lo ya conocido. Otros se programan para realizar acciones que generen escenarios no vistos nunca antes por el sistema. Estos sistemas por ejemplo superan videojuegos difíciles «aprendiendo» sin supervisión directa.

Valores humanos y alineamiento

Imitación e inferencia

Cuando los entornos son demasiado complejos para especificar todo al sistema, una solución es que aprenda imitando. Se ha explorado, por ejemplo, en conducción autónoma, enseñando con vídeos de humanos al volante.

Uno de los desafíos de esta estrategia son los «errores en cascada»; la dificultad que tiene el modelo para recuperarse de errores no anticipados, situaciones no vistas. Esto se ha abordado parcialmente con métodos de aprendizaje profundo como la retropropagación.

Aprendizaje por refuerzo inverso (IRL)
El aprendizaje por refuerzo normal plantea la cuestión de: dado un sistema de incentivos, ¿qué comportamiento lo maximizará? En IRL, la pregunta es la inversa: dado el comportamiento observado, ¿qué incentivo, si lo hay, se está optimizando?

Un modelo de IRL infiere lo que se desea conseguir en base al comportamiento que percibe. Esto implica asumir que el sistema aprende de un experto infalible. Surgen cuestiones como hasta qué punto queremos que estos sistemas aprendan de nosotros, cuándo queremos que obedezcan o que desobedezcan. Los valores humanos en su estado actual parecen no ser suficientes. Se propone que puede ser necesario inculcar una **extrapolación coherente de nuestra voluntad: un ideal** de lo que desearíamos ser, no lo que somos.

El libro hace énfasis en las implicaciones a la hora de regular y legislar. Los **usuarios deben ser capaces de ver y modificar estos modelos,** para garantizar que respondan a sus deseos a largo plazo, no a acciones esporádicas o contextuales.

El consentimiento informado es fundamental para la seguridad y el bienestar, y puede entrar en conflicto con los intereses de las empresas que desarrollan los modelos.

Aprendizaje por refuerzo cooperativo (CRL)
Una posible solución es el aprendizaje por refuerzo cooperativo, en la que el sistema y el humano interactúan de forma que el humano pueda hacer correcciones y reconducir al modelo según sea necesario.

Incertidumbre

Las redes neuronales profundas son fáciles de engañar y no manejan bien la incertidumbre. En lugar de admitir que se les presenta un escenario desconocido y actuar con cautela, proceden de todas formas con gran confianza.

Cómo actuar cuando no se está seguro de qué es lo correcto es una cuestión moral muy compleja. Se está buscando la solución procurando que la IA no tome nunca acciones irreversibles, simulando incertidumbre y generando la necesidad de verificar la acción con un humano en ciertas ocasiones.

Para evitar acciones irreversibles, se usan técnicas como:

- ***Stepwise relative reachability:*** evaluar las consecuencias potenciales de una acción en términos de su accesibilidad relativa desde un estado base.
- ***Attainable utility preservation:*** diseñar los sistemas no sólo para lograr un único objetivo concreto, sino para mantener su capacidad de lograr otros objetivos auxiliares en el futuro.

Introducir incertidumbre mediante técnicas como el «dropout» o con grupos de modelos simultáneos que pueden entrar en conflicto permitiría al sistema reconocer que puede estar equivocándose. Pero sólo resuelve el problema si el sistema nunca llega a ganar demasiada confianza, o si el humano nunca actúa de manera que el sistema perciba como irracional o errónea.