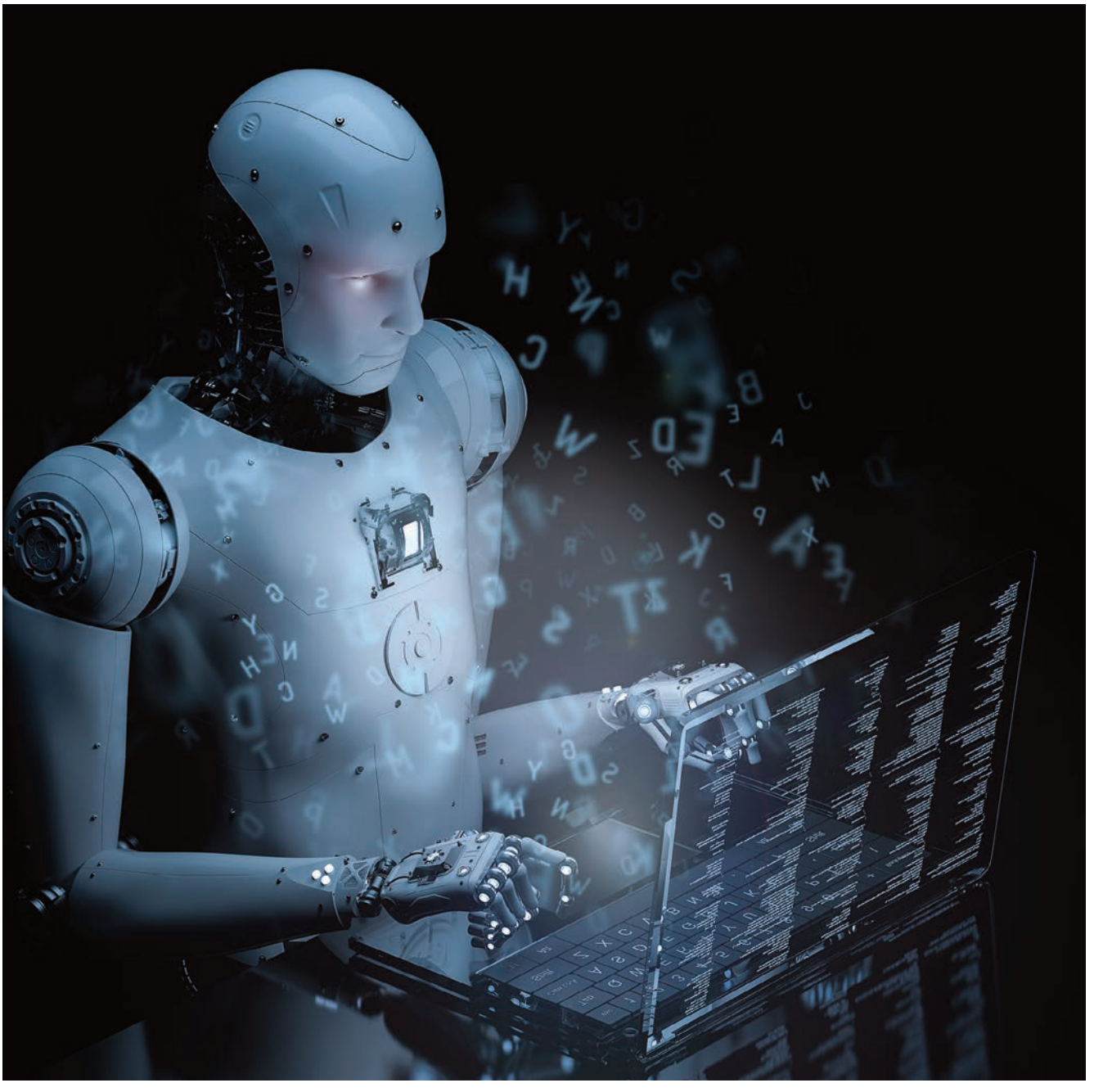
Glosario



Aprendizaje automático (machine learning): subcampo de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos que permiten a las máquinas aprender y mejorar su rendimiento en una tarea específica a través de la experiencia.

Caja blanca (white box): sistema o modelo de AI cuyo funcionamiento interno es sencillo de entender y explicar.

Caja negra (black box): sistema o modelo de AI cuyo funcionamiento interno es desconocido o difícil de entender.

Derecho a una explicación: concepto legal que sostiene que los individuos tienen derecho a saber cómo se toman las decisiones automatizadas que les afectan y a recibir una explicación comprensible de cómo funcionan los algoritmos involucrados.

Explicabilidad: capacidad de un sistema de AI para proporcionar justificaciones claras y comprensibles de sus predicciones o decisiones a los usuarios y partes interesadas. Implica ofrecer información detallada y contextualizada sobre cómo y por qué un modelo de AI llega a una conclusión particular, facilitando la confianza y la adopción de la tecnología.

GPT-4: cuarta generación del modelo Generative Pre-trained Transformer, desarrollado por la OpenAI Foundation, que se utiliza para tareas de procesamiento del lenguaje natural y generación de texto.

Inteligencia artificial (AI): campo de estudio que busca desarrollar sistemas capaces de realizar tareas que normalmente requieren inteligencia humana, como el aprendizaje, el razonamiento, la percepción y la toma de decisiones.

Inteligencia artificial explicable (XAI): enfoque de AI que busca hacer que los modelos de inteligencia artificial sean más comprensibles y transparentes para los humanos.

Interpretabilidad: facilidad con la que los humanos pueden comprender el proceso de toma de decisiones de un modelo de AI, así como las relaciones entre las características de entrada y las predicciones o decisiones. Un modelo interpretable permite a los usuarios discernir cómo se llega a una predicción o decisión específica.

LIME (Local Interpretable Model-agnostic Explanations): técnica de explicabilidad que ayuda a entender las predicciones individuales de un modelo de AI mediante la creación de aproximaciones locales interpretables.

Modelo subrogado: modelo interpretable que se entrena para imitar las predicciones de un modelo de AI complejo y menos interpretable, como una red neuronal profunda. El objetivo de un modelo subrogado es proporcionar una explicación simplificada y comprensible de cómo el modelo original toma decisiones.

Open AI Foundation: organización de investigación y desarrollo de la inteligencia artificial, actualmente participada por Microsoft, cuyo objetivo declarado es garantizar que la AI beneficie a toda la humanidad.

Partial Dependence Plot (PDP): técnica de visualización que muestra el efecto promedio de una característica en las predicciones de un modelo de AI, manteniendo constantes todas las demás características. Ayuda a comprender la relación entre las características y las predicciones, y a detectar posibles interacciones y no linealidades.

Prueba de esquemas de Winograd: prueba de comprensión del lenguaje natural que evalúa la capacidad de una IA para resolver ambigüedades en el lenguaje a través del uso de conocimiento y razonamiento común.

Reglamento General de Protección de Datos (GDPR): legislación de la Unión Europea que establece reglas para la recopilación, el almacenamiento y el procesamiento de datos personales de los ciudadanos de la UE.

Sesgos en AI: prejuicios sistemáticos presentes en los datos de entrenamiento o en el diseño de un algoritmo de AI que pueden llevar a decisiones injustas o discriminatorias.

SHAP (SHapley Additive exPlanations): técnica de explicabilidad que utiliza valores de Shapley, provenientes de la teoría de juegos cooperativos, para atribuir la importancia de cada variable en la predicción de un modelo de AI.

Sparsity: propiedad de un modelo por la que este solo considera el subconjunto de variables que son realmente relevantes para la estimación.

Test de Turing: prueba propuesta por Alan Turing en 1950 que evalúa la capacidad de una máquina de imitar la inteligencia humana al punto de ser indistinguible de un humano en una conversación.

Transformer: arquitectura de red neuronal introducida por Google Brain en 2017 que se utiliza principalmente en tareas de procesamiento del lenguaje natural (NLP). Los transformers son conocidos por su capacidad para manejar secuencias largas de datos y por su eficiencia en el entrenamiento. Se basan en mecanismos de atención, que permiten a la red ponderar la importancia relativa de las palabras o elementos en una secuencia a lo largo del tiempo. Los transformers han impulsado el desarrollo de modelos de lenguaje de vanguardia, como GPT y BERT, y han revolucionado el campo de NLP.

Transparencia: apertura y accesibilidad de un sistema de AI en términos de su diseño, estructura y procesos internos. Un sistema transparente permite a los usuarios y partes interesadas examinar y comprender sus componentes, algoritmos y decisiones.

Red neuronal profunda: tipo de algoritmo de aprendizaje automático que consta de múltiples capas de neuronas artificiales y es capaz de aprender representaciones jerárquicas de datos.