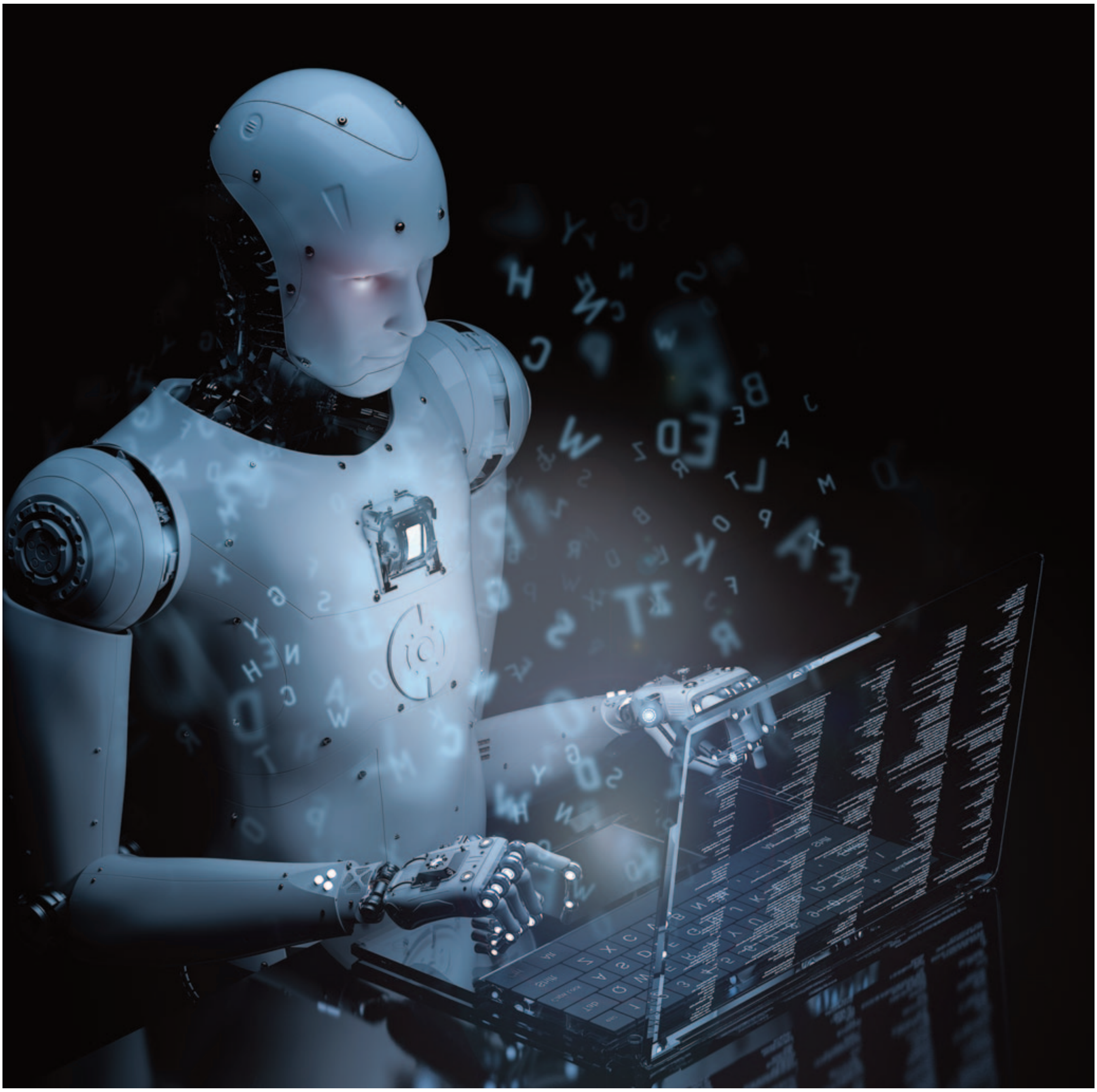
# Glossary



**Machine learning: s**ubfield of artificial intelligence focusing on the development of algorithms and models that enable machines to learn and improve their performance on specific tasks through experience.

**White box:** AI system or model whose inner workings are simple to understand and explain.

**Black box:** an AI system or model whose inner workings are unknown or difficult to understand.

**Right to an explanation:** legal concept holding that individuals have the right to know how automated decisions affecting them are made and to receive an understandable explanation of how the algorithms involved work.

**Explainability:** AI system´s ability to provide clear and understandable reasons for its predictions or decisions to users and stakeholders. This involves providing detailed and contextualized information on how and why an AI model arrives at a particular conclusion, which promotes trust and makes it easier for the technology to be adopted.

**GPT-4:** fourth generation of the Generative Pre-trained Transformer model, developed by the OpenAI Foundation, which is used for natural language processing and text generation tasks.

**Artificial intelligence (AI):** field of study that seeks to develop systems capable of performing tasks that normally require human intelligence, such as learning, reasoning, perception and decision making.

**Explainable artificial intelligence (XAI):** AI approach that seeks to make artificial intelligence models more understandable and transparent to humans.

**Interpretability:** ease with which humans can understand an AI model's decision-making process, as well as the relationships between input features and predictions or decisions. An interpretable model allows users to discern how a specific prediction or decision is arrived at.

**LIME (Local Interpretable Model-agnostic Explanations):** an explainability technique that helps to understand the individual predictions of an AI model by creating local interpretable approximations.

**Surrogate model:** interpretable model that is trained to mimic the predictions of a complex and less interpretable AI model, such as a deep neural network. The goal of a surrogate model is to provide a simplified and understandable explanation of how the original model makes decisions.

**Open AI Foundation:** an artificial intelligence research and development organization, currently owned by Microsoft, whose stated goal is to ensure that AI benefits all of humanity.

**Partial Dependence Plot (PDP):** a visualization technique that shows the average effect of a feature on the predictions of an AI model, holding all other features constant. It helps to understand the relationship between features and predictions, and to detect potential interactions and nonlinearities.

**Winograd Schema Test:** natural language understanding test that assesses an AI's ability to resolve ambiguities in language through the use of common knowledge and reasoning.

**General Data Protection Regulation (GDPR):** European Union legislation that lays down rules for the collection, storage and processing of personal data of EU citizens.

**AI bias:** systematic bias present in training data or in the design of an AI algorithm that can lead to unfair or discriminatory decisions.

**SHAP (SHapley Additive exPlanations):** explainability technique that uses Shapley values from cooperative game theory to attribute the importance of each variable in an AI model's prediction.

**Sparsity:** model property whereby the model only considers the subset of variables that are really relevant for calculation.

**Turing test:** test proposed by Alan Turing in 1950 that evaluates a machine's ability to imitate human intelligence to the point of being indistinguishable from a human in a conversation.

**Transformer:** neural network architecture introduced by Google Brain in 2017 that is primarily used in natural language processing (NLP) tasks. Transformers are known for their ability to handle long data sequences and for their training efficiency. They are based on attentional mechanisms, which allow the network to weigh the relative importance of words or items in a sequence over time. Transformers have driven the development of state-of-the-art language models, such as GPT and BERT, and have revolutionized the NLP field.

**Transparency:** an AI system's openness and accessibility in terms of its design, structure, and internal processes. A transparent system allows users and stakeholders to examine and understand its components, algorithms and decisions.

**Deep neural network:** machine learning algorithm that has multiple layers of artificial neurons and is capable of learning hierarchical representations of data.