



Key Terms for AI Governance

Key Terms for AI Governance

The field of AI is rapidly evolving across different sectors and disparate industries, leaving business, technology and government professionals without a common lexicon and shared understanding of terms and phrases used in AI governance. Even a search to define "artificial intelligence" returns a range of definitions and examples. From the cinematic, like HAL 9000 from "2001: A Space Odyssey," to the creative, like Midjourney and DALL-E generative art, to the common, like email autocorrect and mobile maps, the use cases and applications of AI continue to grow and expand into all aspects of life. This glossary is an update to the October 2023 release of IAPP's Key Terms for AI Governance.

Our glossary of key terms is an adaptation of the July 2024 release of the IAPP's "Key Terms for AI Governance." This adaptation was crafted to provide an accessible and practical reference on fundamental concepts in AI governance, while maintaining the quality and thoroughness of the original publication. The explanations aim to present both policy and technical perspectives, contributing to the ongoing, robust discourse on AI governance.

Accountability



The obligations and responsibilities of an AI system's developers and deployers to ensure the system operates in a manner that is ethical, fair, transparent and compliant with applicable rules and regulations (see also fairness and transparency). Accountability ensures the actions, decisions and outcomes of an AI system can be traced back to the entity responsible for it.

Accuracy



The degree to which an AI system correctly performs its intended task. It is the measure of the system's performance and effectiveness in producing correct outputs based on its input data. Accuracy is a critical metric in evaluating the reliability of an AI model, especially in applications requiring high precision, such as medical diagnoses.

Active learning



A subfield of AI and machine learning in which an algorithm selects some of the data it learns from. Instead of learning from all the data it is given, an active learning model requests additional data points that will help it learn the best.

Adaptive learning



A method that adjusts and tailors educational content to the specific needs, abilities and learning pace of individual students. The purpose of adaptive learning is to provide a personalized and optimized learning experience, catering to the diverse learning styles of students.

Adversarial attack



A safety and security risk to the AI model that can be instigated by manipulating the model, such as by introducing malicious or deceptive input data. Such attacks can cause the model to malfunction and generate incorrect or unsafe outputs, which can have significant impacts. For example, manipulating the inputs of a self-driving car may fool the model to perceive a red light as a green one, adversely impacting road safety.

AI assurance



A combination of frameworks, policies, processes and controls that measure, evaluate and promote safe, reliable and trustworthy AI. AI assurance schemes may include conformity, impact and risk assessments, AI audits, certifications, testing and evaluation, and compliance with relevant standards.

AI audit



A review and assessment of an AI system to ensure it operates as intended and complies with relevant laws, regulations and standards. An AI audit can help identify and map risks and offer mitigation strategies.

AI governance



A system of laws, policies, frameworks, practices and processes at international, national and organizational levels. AI governance helps various stakeholders implement, manage, oversee and regulate the development, deployment and use of AI technology. It also helps manage associated risks to ensure AI aligns with stakeholders' objectives, is developed and used responsibly and ethically, and complies with applicable legal and regulatory requirements.

Algorithm



A procedure or set of instructions and rules designed to perform a specific task or solve a particular problem using a computer.

Artificial general intelligence



AI that is considered to have human-level intelligence and strong generalization capability to achieve goals and carry out a broad range of tasks in different contexts and environments. AGI remains a theoretical field of research. It is contrasted with "narrow" AI, which is used for specific tasks or problems. → Acronym: AGI

Artificial intelligence

Artificial intelligence is a broad term used to describe an engineered system that uses various computational techniques to perform or automate tasks. This may include techniques, such as machine learning, in which machines learn from experience, adjusting to new input data and potentially performing tasks previously done by humans. More specifically, it is a field of computer science dedicated to simulating intelligent behavior in computers. It may include automated decision-making. → Acronym: AI

Automated decision-making

The process of making a decision by technological means without human involvement, either in whole or in part.

Bias

There are several types of bias within the AI field. Computational bias or machine bias is a systematic error or deviation from the true value of a prediction that originates from a model's assumptions or the data itself (see also input data). Cognitive bias refers to inaccurate individual judgment or distorted thinking, while societal bias leads to systemic prejudice, favoritism and/or discrimination in favor of or against an individual or group. Either or both may permeate the model or the system in numerous ways, such as through selection bias, i.e. biases in selecting data for model training. Bias can impact outcomes and pose a risk to individual rights and liberties.

Bootstrap aggregating

A machine learning method that aggregates multiple versions of a model (see also machine learning model) trained on random subsets of a dataset. This method aims to make a model more stable and accurate. → Sometimes referred to as bagging.

Chatbot



A form of AI designed to simulate human-like conversations and interactions that uses natural language processing and deep learning to understand and respond to text or speech.

Classification model



A type of model (see also machine learning model) used in machine learning that is designed to take input data and sort it into different categories or classes. → Sometimes referred to as classifiers.

Clustering



An unsupervised machine learning method in which patterns in the data are identified and evaluated, and data points are grouped accordingly into clusters based on their similarity.
→ Sometimes referred to as clustering algorithms.

Compute



The processing resources that are available to a computer system. This includes hardware components such as the central processing unit or graphics processing unit. Compute is essential for memory, storage, processing data, running applications, rendering graphics for visual media and powering cloud computing, among others.

Computer vision




A field of AI that uses computers to process and analyze images, videos and other visual inputs. Common applications of computer vision include facial recognition, object recognition and medical imaging.

Conformity assessment




An analysis, often performed by an entity independent of a model developer, on an AI system to determine whether requirements, such as establishing a risk management system, data governance, record-keeping, transparency and cybersecurity practices have been met.

Contestability




The principle of ensuring AI systems and their decision-making processes can be questioned or challenged by humans. This ability to contest or challenge the outcomes, outputs and actions of AI systems depends on transparency and helps accountability within AI governance.
→ Also called redress.

Corpus




A large collection of texts or data that a computer uses to find patterns, make predictions or generate specific outcomes. The corpus may include structured or unstructured data and cover a specific topic or a variety of topics.

Data leak




An accidental exposure of sensitive, personal, confidential or proprietary data. This can be a result of poor security defenses, human error, storage misconfigurations or a lack of robust policies around internal and external data sharing practices. Unlike a data breach, a data leak is unintentional and not done in bad faith.

Data poisoning



An adversarial attack in which a malicious user injects false data into a model to manipulate the training process, thereby corrupting the learning algorithm. The goal is to introduce intentional errors into the training dataset, leading to compromised performance and resulting in undesired, misleading or harmful outputs.

Data provenance



A process that tracks and logs the history and origin of records in a dataset, encompassing the entire life cycle from its creation and collection to its transformation to its current state. It includes information about sources, processes, actors and methods used to ensure data integrity and quality. Data provenance is essential for data transparency and governance, and it promotes better understanding of the data and eventually the entire AI system.

Data quality



The measure of how well a dataset meets the specific requirements and expectations for its intended use. Data quality directly impacts the quality of AI outputs and the performance of an AI system. High-quality data is accurate, complete, valid, consistent, timely and fit for purpose.

Decision tree



A type of supervised learning model used in machine learning (see also machine learning model) that represents decisions and their potential consequences in a branching structure.

Deep learning



A subfield of AI and machine learning that uses artificial neural networks. Deep learning is especially useful in fields where raw data needs to be processed, like image recognition, natural language processing and speech recognition.

Deepfakes



Audio or visual content that has been altered or manipulated using artificial intelligence techniques. Deepfakes can be used to spread misinformation and disinformation.

Diffusion model



A generative model used in image generation that works by iteratively refining a noise signal to transform it into a realistic image when prompted.

Discriminative model



A type of model (see also machine learning model) used in machine learning that directly maps input features to class labels and analyzes for patterns that can help distinguish between different classes. It is often used for text classification tasks, like identifying the language of a piece of text or detecting spam. Examples are traditional neural networks, decision trees and random forest.

Disinformation



Audio or visual content that is intentionally manipulated or created to cause harm. Disinformation can spread through deepfakes created by those who have malicious intentions.

Entropy



The measure of unpredictability or randomness in a set of data used in machine learning. A higher entropy signifies greater uncertainty in predicting outcomes.

Expert system



A form of rules-based AI that draws inferences from a knowledge base provided by human experts to replicate their decision-making abilities within a specific field, like medical diagnoses.

Explainability



The ability to describe or provide sufficient information about how an AI system generates a specific output or arrives at a decision in a specific context. Explainability is important in maintaining transparency and trust in AI.

Exploratory data analysis



Data discovery process techniques that take place before training a machine learning model to gain preliminary insights into a dataset, such as identifying patterns, outliers and anomalies and finding relationships among variables.

Fairness



An attribute of an AI system that prioritizes relatively equal treatment of individuals or groups in its decisions and actions in a consistent, accurate and measurable manner. Every model must identify the appropriate standard of fairness that best applies, but most often it means the AI system's decisions should not adversely impact, whether directly or disparately, sensitive attributes like race, gender or religion.

Federated learning



A machine learning method that allows models (see also machine learning model) to be trained on the local data of multiple edge devices. Only the updates of the local model, not the training data itself, are sent to a central location where they are aggregated to improve the global model — a process that is iterated until the global model is fully trained. This process enables better privacy and security controls for the individual user data.

Fine-tuning



The process of taking a pretrained deep learning model and training it further for a specialized task through supervised learning. It involves taking a foundation model that has already learned general patterns from a large dataset and training it for a specific task using a much smaller and labeled dataset.

Foundation model



A large-scale model that has been trained on extensive and diverse datasets to enable broad capabilities, such as language (see also large language model), vision, robotics, reasoning, search or human interaction, that can function as the base for use-specific applications.

→ Also called general purpose AI model and frontier AI.

Generalization



The ability of a model (see also machine learning model) to understand the underlying patterns and trends in its training data and apply what it has learned to make predictions or decisions about new, unseen data.

Generative AI



A field of AI that uses deep learning trained on large datasets to create content, such as written text, code, images, music, simulations and videos, in response to user prompts. Unlike discriminative models, generative AI makes predictions on existing data rather than new data.

Greedy algorithms



A type of algorithm that makes the optimal choice to achieve an immediate objective at a particular step or decision point based on the available information and without regard for the long-term optimal solution.

Ground truth



The absolute or objectively known state of a dataset against which the quality of an AI system can be evaluated. It serves as the real-world reference against which the outputs are measured for accuracy and reliability.

Hallucinations



Instances in which generative AI models create seemingly plausible but factually incorrect outputs under the appearance of fact.
→ Also called confabulations.

Human-centric AI



An approach to AI design, development, deployment and use that prioritizes human well-being, autonomy, values and needs. The goal is to develop AI systems that amplify and augment human abilities rather than undermine them.

Human-in-the-loop



A design paradigm that incorporates human oversight, intervention, interaction or control over the operation and decision-making processes of an AI system. → Acronym: HITL

Impact assessment



An evaluation process designed to identify, understand, document and mitigate the potential ethical, legal, economic and societal implications of an AI system in a specific use case.

Inference

A type of machine learning process in which a trained model (see also machine learning model) is used to make predictions or decisions based on input data.

Input data

Data provided to or directly acquired by a learning algorithm or model (see also machine learning model) for the purpose of producing an output. It forms the basis for machine learning models to learn, make predictions and carry out tasks.

Interpretability

The ability to explain or present a model's reasoning in human-understandable terms. Unlike explainability, which provides an explanation after a decision is made, interpretability emphasizes designing models that inherently facilitate understanding through their structure, features or algorithms. Interpretable models are domain-specific and require significant domain expertise to develop.

Large language model

A form of AI that utilizes deep learning algorithms to create models (see also machine learning model, foundation model and fine-tuning) pretrained on massive text datasets for the general purpose of analyzing and learning patterns and relationships among characters, words and phrases to perform text-based tasks. There are generally two types of LLMs: generative models that make text predictions based on the probabilities of word sequences learned from its training data (see also generative AI) and discriminative models that make classification predictions based on probabilities of data features and weights learned from its training data (see also discriminative model). The word large generally refers to the model's capacity measured by the number of parameters and to the enormous datasets it is trained on. → Acronym: LLM

Machine learning

A subfield of AI involving algorithms that iteratively learn from and then make decisions, recommendations, inferences or predictions based on input data. These algorithms build a model from training data to perform a specific task on new data without being explicitly programmed to do so. Machine learning implements various algorithms that learn and improve by experience in a problem-solving process that includes data collection and preparation, feature engineering, training, testing and validation. Companies and government agencies deploy machine learning algorithms for tasks such as fraud detection, recommender systems, customer inquiries, health care, and transportation and logistics. → Acronym: ML

Machine learning model

A learned representation of underlying patterns and relationships in data, created by applying an AI algorithm to a training dataset. The model can then be used to make predictions or perform tasks on new, unseen data.

Misinformation

False audio or visual content that is unintentionally misleading. It can be spread through deepfakes created by those who lack intent to cause harm.

Model card

A brief document that discloses information about an AI model, like explanations about intended use, performance metrics and benchmarked evaluation in various conditions, such as across different cultures, demographics or race (see also system cards).

Multimodal models

A type of model used in machine learning (see also machine learning model) that can process more than one type of input or output data, or "modality," at the same time. For example, a multimodal model can take both an image and text caption as input and then produce a unimodal output in the form of a score indicating how well the text caption describes the image. These models are highly versatile and useful in a variety of tasks, like image captioning and speech recognition

Natural language processing



A subfield of AI that helps computers understand, interpret and apply human language by transforming information into content. It enables machines to translate languages, read text or spoken language, interpret its meaning, measure sentiment, and determine which parts are important for understanding.

Neural networks



A type of model (see also machine learning model) used in deep learning that mimics the way neurons in the human brain interact with multiple processing layers, including at least one hidden layer. This layered approach enables machine-based neural networks to model complex nonlinear relationships and patterns within data. Artificial neural networks have a range of applications, such as image recognition and medical diagnoses.

Open-source software



A decentralized development model that provides the public with free and open access to source code, which can then be viewed, modified and redistributed according to the terms of its respective license. The goal is to promote innovation, transparency, shared collaboration and learning among researchers and technical experts.

Overfitting



A concept in machine learning that involves a model (see also machine learning model) becoming too specific to the training data and unable to generalize to unseen data, which means it can fail to make accurate predictions on new datasets.

Oversight



The process of effectively monitoring and supervising an AI system to minimize risks, ensure regulatory compliance and uphold responsible practices.

Oversight is important for effective AI governance, and mechanisms may include certification processes, conformity for

assessments and regulatory authorities responsible enforcement.

Post processing



Steps performed after a machine learning model has been run to adjust its output. This can include adjusting a model's outputs or using a holdout dataset — data not used in the training of the model — to create a function run on the model's predictions to improve fairness or meet business requirements.

Preprocessing



Steps taken to prepare data for training a machine learning model, which can include cleaning the data, handling missing values, normalizing the data, performing feature extraction and encoding categorical variables. Data preprocessing can play a crucial role in improving data quality, mitigating bias, addressing algorithmic fairness concerns and enhancing the performance and reliability of machine learning algorithms.

Prompt



An input or instruction provided to an AI model or system to generate an output.

Prompt engineering



Prompt engineering is the deliberate process of structuring a prompt or series of prompts to influence model behavior to generate more desirable outputs.

Random forest



A supervised machine learning (see also supervised learning) algorithm that builds multiple decision trees and merges them to get a more accurate and stable prediction. Each decision tree is built with a random subset of the training data (see also bootstrap aggregating), hence the name random forest. Random forests are helpful to use with datasets that are missing values or are very complex.

Red teaming

The process of testing the safety, security and performance of an AI system through an adversarial lens, typically through the simulation of adversarial attacks on the model to evaluate it against certain benchmarks, jailbreak it and try to make it behave in unintended or inappropriate ways. Red teaming reveals security risks, model flaws, biases, misinformation and other harms, and the results of such testing are passed along to the model developers for evaluation and remediation. Developers use red teaming to improve a model before and after releasing it to the public.

Reinforcement learning

A machine learning method that trains a model to optimize its actions within a given environment to achieve a specific goal, guided by feedback mechanisms of rewards and penalties. This training is often conducted through trial-and-error interactions or simulated experiences that do not require external data. For example, an algorithm can be trained to earn a high score in a video game by having its efforts evaluated and rated according to success toward the goal.

Reinforcement learning with human feedback

The process of combining the technique of reinforcement learning with human feedback during the training process. Human feedback is provided on the model's output, often by comparing different outputs for the same prompt and indicating which output aligns better with human preferences. In reinforcement learning, the model learns by receiving rewards or penalties. Combining this with human feedback provides an additional source of rewards and penalties and helps align the AI's behavior with human preferences and values. → Acronym: RLHF

Reliability

An attribute of an AI system that ensures it behaves as expected and performs its intended function consistently and accurately, even with new data that it has not been trained on.

Robotics



A multidisciplinary field that encompasses the design, construction, operation and programming of robots. Robotics allow AI systems and software to interact with the physical world.

Robustness



An attribute of an AI system that signifies the system's ability to be resilient to, overcome and withstand security attacks. Robustness ensures the system's functionality, performance and accuracy in a variety of environments and circumstances, even when faced with changed inputs or security attacks.

Safety



A broad term, which may refer to designing, developing and deploying AI systems that minimize AI harms from misinformation, disinformation, deepfakes, hallucinations and other unintended behaviors. It may also refer to mitigating and managing malicious use or rogue behavior. Safety also encompasses the prevention of existential or unexpected risks that may arise from advanced AI capabilities reflected in foundation models.

Semi-supervised learning




A subset of machine learning that combines both supervised learning and unsupervised learning by training the model on a large amount of unlabeled data and a small amount of labeled data. This avoids the challenges of finding large amounts of labeled data for training the model. Generative AI commonly relies on semi-supervised learning.

Small language models




A smaller version of their better-known and larger counterparts, large language models. Small is a reference to the size of the models. They have fewer parameters and require a much smaller training dataset, optimizing them for efficiency and better suiting them for deployment in environments with limited computational resources or for applications that require faster training and inference time.

Supervised learning




A subset of machine learning in which the model (see also machine learning model) is trained on labeled input data with known desired outputs. These two groups of data are sometimes called predictors and targets or independent and dependent variables respectively. This type of learning is useful for classification or regression. The former refers to training an AI to group data into specific categories and the latter refers to making predictions by understanding the relationship between two variables.

Synthetic data




Data generated by a system or model (see also machine learning model) that generally resembles the structure and statistical properties of real data but without any real-world, identifying information. It is often used for testing or training machine-learning models, particularly in cases with limited, unavailable or too sensitive real-world data.

System card




Similar to a model card, a system card is a brief document that discloses information about how various AI models work together within a network of AI systems, promoting greater explainability of the overall system.

Testing data




The dataset used to test and evaluate a trained model (see also machine learning model). It is used to assess the performance of the model with new data at the very end of the initial model development process and for future upgrades or variations to the model.

Training data




The dataset used to train a model (see also machine learning model) so it can accurately predict outcomes, find patterns or identify structures within the training data.

Transfer learning model




A type of model (see also machine learning model) used in machine learning in which an algorithm learns to perform one task, such as recognizing cats, and then uses that learned knowledge as a basis when learning a different but related task, such as recognizing dogs.

Transformer model



A neural network architecture that learns context and maintains relationships between sequence data, such as words in a sentence. It does so by leveraging the technique of attention, i.e., focusing on the most important and relevant parts of the input sequence. This helps to improve model accuracy. For example, in language learning tasks, by attending to the surrounding words, the model can comprehend the meaning of a word in the context of the whole sentence.

Transparency



A broad term that implies openness, comprehensibility and accountability in the way AI algorithms function and make decisions. However, the specific meaning of transparency may vary depending on context. May refer to the extent to which information regarding an AI system is made available to stakeholders, including disclosing if AI is used through techniques like watermarking, and explaining how the model works through model or system cards for example. It also refers to maintenance of technical and nontechnical documentation across the AI life cycle to keep track of processes and decision-making, which can also assist with auditability of the AI system. In the open-source context, transparency may refer to making the source code publicly accessible.

Trustworthy AI



In most cases, this term is used interchangeably with the terms responsible AI and ethical AI, which all refer to principle-based AI development and AI governance, including the principles of security, safety, transparency, explainability, accountability, privacy and nondiscrimination/nonbias (see also bias), among others.

Turing test



A test of a machine's ability to exhibit intelligent behavior equivalent to or indistinguishable from that of a human. Alan Turing (1912-1954) originally thought of the test to be an AI's ability to converse through a written text, such that a human reader would not be able to tell a computer-generated response from that of a human.

Under fitting



A neural network architecture that learns context and maintains relationships between sequence data, such as words in a sentence. It does so by leveraging the technique of attention, i.e., focusing on the most important and relevant parts of the input sequence. This helps to improve model accuracy. For example, in language learning tasks, by attending to the surrounding words, the model can comprehend the meaning of a word in the context of the whole sentence.

Unsupervised learning



A subset of machine learning in which the model is trained by looking for patterns in an unclassified dataset with minimal human supervision. The AI is provided with preexisting unlabeled datasets and then analyzes those datasets for patterns. This type of learning is useful for training an AI for techniques such as clustering data, outlier detection, dimensionality reduction, feature learning and principal component analysis.

Variables



In the context of machine learning, a variable is a measurable attribute, characteristic or unit that can take on different values. Variables can be numerical/quantitative or categorical/qualitative. → Also called features.

Variance



A statistical measure that reflects how far a set of numbers are spread out from their average value in a dataset. A high variance indicates the data points are spread widely around the mean. A low variance indicates the data points are close to the mean. In machine learning, higher variance can lead to overfitting. The trade-off between variance and bias is a fundamental concept in machine learning. Model complexity tends to reduce bias but increase variance. Decreasing complexity reduces variance but increases bias.

Watermarking



The process of embedding subtle or visually imperceptible patterns in AI-generated content or metadata that can only be detected by computers. Watermarking helps with the detection and labelling of AI generated content, promoting transparency.