Key Terms for AI Governance

The field of artificial intelligence 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 June 2023 release of IAPP's Key Terms for AI Governance. The updated version has been developed with reference to various materials and with valuable feedback from experts on IAPP's AI Governance Center Advisory Board. It includes new terms and modifications to existing terms. To learn more about the new version, please refer to the explainer article on the Updates to the Key Terms for AI Governance.

The original glossary from June 2023 was developed with reference to numerous materials and designed to provide succinct, but nuanced, definitions and explanations for some of the most common terms related to AI today. The same methodology has been retained for the new updates as well. The explanations aim to present both policy and technical perspectives and add to the robust discourse on AI governance. Although there are some shared terms and definitions, this glossary is separate from the official IAPP Glossary of Privacy Terms.

A | B | C | D | E | F | G | H | I | L | M | N | O | P | R | S | T | U | V | Key terms

TERM	DEFINITION
Accountability[1]	The obligation and responsibility of the creators, operators and regulators of an AI system to ensure the system operates in a manner that is ethical, fair, transparent and compliant with applicable rules and regulations (see fairness and transparency). Accountability ensures the actions, decisions and outcomes of an AI system can be traced back to the entity responsible for it.
Active learning	A subfield of AI and machine learning where an algorithm can select 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. → Also called query learning.

TERM	DEFINITION
Adversarial machine learning	A machine learning technique that raises a safety and security risk to the model and can be seen as an attack. These attacks 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 governance	A system of laws, policies, frameworks, practices and processes at international, national and organizational levels. AI governance helps various stakeholders implement, manage and oversee the 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 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 variety of tasks in different contexts and environments. AGI still 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, where 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 is a systematic error or deviation from the true value of a prediction that originates from a model's assumptions or the input data itself. 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. 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 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.

TERM	DEFINITION
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 other media. Because chatbots are often used for customer service and other personal help applications, chatbots often ingest users' personal information.
Classification model	A type of model (see 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 where 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	Refers to the processing resources that are available to a computer system. This includes the hardware components such as the central processing unit or graphics processing unit. Computing is essential for memory, storage, processing data, running applications, rendering graphics for visual media, powering cloud computing, among others.
Computer vision	A field of AI that enables computers to process and analyze images, videos and other visual inputs.
Conformity assessment	An analysis, often performed by a third-party body, 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. Often referred to as audit.
Contestability	The principle of ensuring that AI systems and their decision-making processes can be questioned or challenged. This ability to contest or challenge the outcomes, outputs and/or actions of AI systems can help promote transparency and 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.
Decision tree	A type of supervised learning model used in machine learning (see 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	Audiovisual content that has been altered or manipulated using AI techniques. Deepfakes can be used to spread misinformation and disinformation.

TERM	DEFINITION
Discriminative model	A type of model (see 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. Examples are traditional neural networks, decision trees and random forests.
Disinformation	Audiovisual content, information and synthetic data that is intentionally manipulated or created to cause harm. Disinformation can spread through deepfakes by those with 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 AI that draws inferences from a knowledge base to replicate the decision-making abilities of a human expert within a specific field, like a medical diagnosis.
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 to a predetermined addressee. XAI is important in maintaining transparency and trust in AI. → Acronym: XAI
Exploratory data analysis	Data discovery process techniques that take place before training a machine learning model in order to gain preliminary insights into a dataset, such as identifying patterns, outliers, and anomalies and finding relationships among variables.
Fairness[1]	An attribute of an AI system that prioritizes relatively equal treatment of individuals or groups in its decisions and actions in a consistent, accurate 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 machine learning model) to be trained on the local data of multiple edge devices or servers. Only the updates of the local model, not the training data itself, are sent to a central location where they get aggregated into a 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.
Foundation model	A large-scale, pretrained model with AI capabilities, such as language (see large language model), vision, robotics, reasoning, search or human interaction, that can function as the base for use-specific applications. The model is trained on extensive and diverse datasets.
Generalization	The ability of a 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.

TERM	DEFINITION
Generative AI	A field of AI that uses deep learning trained on large datasets to create new content, such as written text, code, images, music, simulations and videos. Unlike discriminative models, Generative AI makes predictions on existing data rather than new data. These models are capable of generating novel outputs based on input data or user prompts.
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 longer-term optimal solution.
Hallucinations	Instances where a generative AI model creates content that either contradicts the source or creates factually incorrect output under the appearance of fact.
Inference	A type of machine learning process where a trained model (see 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 machine learning model for the purpose of producing an output. It forms the basis upon which the machine learning model will learn, make predictions and/or carry out tasks.
Large language model	A form of AI that utilizes deep learning algorithms to create models (see machine learning model) pre-trained on massive text datasets for the general purpose of language learning to analyze and learn patterns and relationships among characters, words and phrases. 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 generative AI) and discriminative models that make classification predictions based on probabilities of data features and weights learned from its training data (see discriminative model). The term "large" generally refers to the model's capacity measured by the number of parameters and to the enormous datasets that it is trained on. → Acronym: LLM
Machine learning	A subfield of AI involving algorithms that enable computer systems to iteratively learn from and then make decisions, 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 cleansing, feature selection, training, testing and validation. Companies and government agencies deploy machine learning algorithms for tasks such as fraud detection, recommender systems, customer inquiries, health care, or transport 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 audiovisual content, information or synthetic data that is unintentionally misleading. It can be spread through deepfakes by those who lack intent to cause harm.

TERM	DEFINITION
Multimodal models	A type of model used in machine learning (see 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 manipulate human language by transforming information into content. It enables machines to 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 machine learning model) used in machine learning that mimics the way neurons in the brain interact with multiple processing layers, including at least one hidden layer. This layered approach enables 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 diagnosis.
Overfitting	A concept in machine learning in which a model (see machine learning model) becomes too specific to the training data and cannot 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 assessments and regulatory authorities responsible for enforcement.
Parameters	The internal variables that an algorithmic model learns from the training data. They are values that the model adjusts to during the training process so it can make predictions on new data. Parameters are specific to the architecture of the model. For example, in neural networks, parameters are the weights and biases of each neuron in the network.
Post processing	Steps performed after a machine learning model has been run to adjust the output of that model. This can include adjusting a model's outputs and/or using a holdout dataset — data not used in the training of the model — to create a function that is run on the model's predictions to improve fairness or meet business requirements.
Preprocessing	Steps taken to prepare data for a machine learning model, which can include cleaning the data, handling missing values, normalization, 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.
Random forest	A supervised machine learning (see supervised learning) algorithm that builds multiple decision trees and merges them together to get a more accurate and stable prediction. Each decision tree is built with a random subset of the training data (see bootstrap aggregating), hence the name "random forest." Random forests are helpful to use with datasets that are missing values or are very complex.

TERM	DEFINITION
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.
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 ensures a resilient system that maintains its functionality and performs accurately in a variety of environments and circumstances, even when faced with changed inputs or adversarial attacks.
Safety	The development of AI systems that are designed to minimize potential harm, including physical harm, to individuals, society, property and the environment.
Semi-supervised learning	A subset of machine learning that combines both supervised 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.
Supervised learning	A subset of machine learning where the model (see 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 that can mimic and resemble the structure and statistical properties of real data. It is often used for testing or training machine learning models, particularly in cases where real-world data is limited, unavailable or too sensitive to use.
Testing data	A subset of the dataset used to test and evaluate a trained model. It is used to test the performance of the machine learning 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	A subset of the dataset that is used to train a machine learning model until it can accurately predict outcomes, find patterns or identify structures within the training data.
Transfer learning model	A type of model (see 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.

TERM	DEFINITION
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. it focuses 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 is able to comprehend the meaning of a word in the context of the whole sentence.
Transparency[1]	The extent to which information regarding an AI system is made available to stakeholders, including disclosing whether AI is used and explaining how the model works. It implies openness, comprehensibility and accountability in the way AI algorithms function and make decisions.
Trustworthy AI	In most cases used interchangeably with the terms responsible AI and ethical AI, which all refer to principle-based AI governance and development, including the principles of security, safety, transparency, explainability, accountability, privacy, nondiscrimination/ nonbias (see 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.
Underfitting	A concept in machine learning in which a model (see machine learning model) fails to fully capture the complexity of the training data. This may result in poor predictive ability and/or inaccurate outputs. Factors leading to underfitting may include too few model parameters, too high a regularization rate, or an inappropriate or insufficient set of features in the training data.
Unsupervised learning	A subset of machine learning where 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, etc.) and dimensionality reduction (feature learning, principal component analysis, etc.).
Validation data	A subset of the dataset used to assess the performance of the machine learning model during the training phase. Validation data is used to fine-tune the parameters of a model and prevent overfitting before the final evaluation using the test dataset.
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. → Sometimes referred to as features.

TERM	DEFINITION
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 that 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.

Notes

1 Different definition than IAPP privacy training glossary.

More resources

→ AI Governance Center

→ AI Body of Knowledge

→ AI topic page

Key terms

Updated October 2023.

Find the latest version at iapp.org/resources