BOOK SAMPLER

AI Distilled

A Human Guide to the World of Artificial Intelligence

Steve Packard CISSP

Glossary and Index

The complete glossary and index.

Glossary: book pages 423-435.

Index: book pages 454-459.

Page references in the index refer to the full book, not to this sampler. Original pagination and layout are preserved.

AIDistilledBook.com

First edition | Copyright 2026 Stephen M. Packard

Glossary

Ablation study - An experiment that removes, disables or changes part of a system and compares the result with an appropriate baseline. It's a lot of fun to try this with an open-source model.

Activation function - A function applied to a neuron's combined input to produce its output. Nonlinear activation functions allow stacked layers to model relationships that a single linear transformation cannot capture.

Adam - An optimizer whose name refers to adaptive moment estimation. Adam tracks moving averages of gradients and their squared values to adjust updates for individual parameters. It is widely used because it often works well with noisy gradients and large models.

Agency - The capacity to act in pursuit of an objective. Something that AI does not have, because there's no being in there to have it. Agency might be used in the computational sense to mean a process that works on its own, but it's never human-like agency.

Agent - Software that can select and carry out a sequence of actions toward an assigned task, often using a model, tools and stored working information. In an LLM application, the surrounding program executes tools and manages the loop. It does not have human agency.

AGI - Artificial general intelligence. A proposed category of AI with broad, adaptable competence across many tasks. It's really stupid to obsess over what it is. It is just a term someone made.

AI - Artificial intelligence. The broad field of building computer systems that perform tasks associated with intelligence, including recognition, prediction, language processing and planning. It's always been difficult to define exactly what is and is not AI.

AI consciousness - A very tired concept. We have talked this thing to death. It's not a thing and there's no reason to think it would ever arise from machine learning.

AI winter - A period when enthusiasm, investment and funding for artificial intelligence decline after expectations outrun practical results. This occurred from the early 1970s to the early 1980s.

Algorithm - A specified procedure for carrying out a task through a sequence of operations. Early AI used hand-coded algorithms. Algorithmic programming is still used for most things.

Alignment - The pseudoscience of thinking it's possible to teach a model how to be empathetic or care about people or something. It's all built on a completely flawed idea of how AI even works.

AlphaFold - A family of DeepMind systems for predicting molecular structures, originally focused on a protein's three-dimensional structure from its amino-acid sequence. AlphaFold 2's 2020 results marked a major advance in protein structure prediction. This proved a breakthrough in molecular sciences.

AlphaGo - DeepMind's Go-playing system, which combined neural networks with a search through possible moves. It learned from human games and then improved through self-play. Prior to this, the game of Go was an unsolved computational challenge.

AlphaZero - DeepMind's extension of the self-play approach to chess, shogi and Go, introduced in 2017. The same general training method was used to train a separate model for each game.

ANN - Artificial neural network. A computational model made from interconnected mathematical units that transform inputs into outputs. Training adjusts parameters so that the network performs a task, such as recognizing handwriting or predicting a value.

Anthropomorphism - Attributing human feelings, intentions or mental processes to something nonhuman. This is the root of so much misunderstanding of AI. It's natural for people to do this, but understanding the mechanics of the system can help resist this tendency.

API - Application programming interface. A defined way for one piece of software to use another's capabilities or data. This is how programmers connect to external services and how systems connect to each other.

ARC-AGI - A family of benchmarks for evaluating adaptation to unfamiliar tasks. It is commonly used for benchmarking models.

Artificial neuron - A small mathematical unit within a neural network. A typical neuron multiplies its inputs by weights, adds those results and a bias, then applies an activation function. The McCulloch-Pitts (MCP) neuron was first described in 1943.

ASI - Artificial Super Idiot. A hypothetical model capable of being as much of an ignoramus as Nick Bostrom.

ASR - Automatic speech recognition. Converting recorded or live speech into written text. Automatic speech recognition systems map patterns in audio to words and punctuation.

ATO - Account takeover. Unauthorized control of a legitimate user's account, often through stolen credentials, phishing or compromised sessions. This causes 90% of cyber security losses.

Attention - A calculation that assigns different weights to available information and combines it accordingly. In a language model, this helps a token's representation incorporate relevant context. It transfers information between tokens.

Attention head - One attention calculation within a layer, which combines information from the token positions it can attend to.

Autoencoder - A neural network trained to reconstruct its input. An encoder converts the input into an internal representation; a decoder reconstructs it from that representation. Autoencoders are used for compression, denoising and as components of image-generation systems.

Autonomy - The extent to which a system can operate without a person directing every step. Autonomy depends on the task, operating conditions, permissions and available support. Machines do not have true autonomy, in the human sense, though they can operate on their own for extended periods.

Autoregression - Generating a sequence one element at a time, with each new element conditioned on those already available. In an autoregressive language model, each token is appended to the end of the stream and then processed again. It ingests its own output.

Backpropagation - An efficient application of the chain rule of calculus that calculates gradients through a network, working backward from the output. It determines how parameters affect the loss. An optimizer then uses those gradients to update parameters. This is how modern ML training works.

Batch and epoch - A batch is a group of examples processed together during training or inference. An epoch is one pass through the training dataset.

BERT - Bidirectional Encoder Representations from Transformers: a language-model architecture introduced by Google researchers in 2018. It builds representations using context on both sides of a word. BERT is commonly used for all manner of things, but it usually works behind the scenes.

Bias (neural network) - An adjustable offset added to a neuron's weighted input before its activation function is applied.

Bidirectional attention - Attention that allows a position to use information from both earlier and later positions in the supplied sequence. It is useful for interpreting complete text, as in BERT. Both directions are available because the relevant input already exists; the model is not seeing future events.

Catastrophic forgetting - A substantial loss of previously learned performance when a model is trained on new material. Changes to shared parameters can improve a new task while damaging older behaviors. It is one reason further training needs evaluation across existing capabilities, rather than only the behavior being added.

Causal attention - An attention arrangement that prevents a position from using later tokens. In a standard text-generating transformer, a position can use itself and earlier positions, meaning information can be transferred right but not left.

Ceding risk - Paying a reinsurer to assume a specified share or layer of potential losses. The insurer transferring that exposure is the cedent or ceding company. The agreement defines which losses transfer and the limits or conditions that apply.

Checkpoint - A saved snapshot of a model at a particular stage. These are very important during training, because catastrophic forgetting and other problems can ruin a training run.

Circuit - In mechanistic interpretability, a set of model components and connections that jointly contribute to a behavior or computation.

CNN - Convolutional neural network. A neural network built around filters that scan for local patterns, such as edges, textures or shapes in images. Successive layers combine simpler patterns into more complex ones. The same filter weights are reused across positions, making this architecture particularly useful for spatial data.

Cognitive architecture - A framework that brings together mechanisms such as memory, perception, reasoning, learning and action selection. Researchers use cognitive architectures to model human cognition or build systems that coordinate several kinds of intelligent behavior. These were some of the earliest attempts at general-purpose artificial intelligence.

Computational functionalism - The philosophical view that mental states depend on the computational organization of a system rather than the material from which it is made. It led some philosophers to think that an AI system might develop a mind.

Computer vision - The field concerned with extracting useful information from images and video. Tasks include detecting objects, recognizing faces, segmenting regions and estimating depth or motion.

Conditioning - Additional information supplied to guide a generative model's output. A text prompt, reference image, pose, depth map or class label can serve as a condition.

Context window - The amount of tokenized information a model can use in one working context, including supplied material and generated continuation within the system's limits.

Convolution - An operation that applies the same small set of weights across different positions in data. In an image, a filter might respond to an edge wherever it appears.

Copy head - An informal name for an attention head whose contribution helps reproduce information from another position in the context.

CoT - Chain of thought. Intermediate text that lays out steps toward an answer. This is similar to thinking out loud or using verbal cognition to solve a problem.

CPU - Central processing unit. The general-purpose processor that executes a computer's instructions and coordinates its work. CPUs handle operating systems, applications and some AI tasks. CPUs are not optimal for machine learning the way GPUs are.

CUDA - NVIDIA's computing platform and programming model for running general-purpose calculations on its GPUs. The name originated as Compute Unified Device Architecture. This is NVIDIA's GPU compute architecture that made it possible to use GPUs for more than graphics.

Data leakage - Information entering training or model selection that would not legitimately be available when the system is used. This term is a little confusing because it exists in cyber security where its meaning is more direct: confidential information that is disclosed.

Dataset - A collection of examples assembled for training, evaluating or otherwise studying a model. Examples may contain text, images, measurements, labels or combinations of these. What the dataset includes, excludes and repeatedly shows can strongly influence the behavior that training produces.

Dataset card - Documentation describing a dataset's contents, origin, organization and intended uses. It may also explain collection methods, languages, licensing and known limitations or biases. A dataset card helps someone decide whether the data suits a task and understand what its examples actually represent.

Deep learning - Machine learning using neural networks with multiple processing layers.

Deterministic - Producing the same result when the inputs, state and relevant execution conditions are the same. This describes repeatability, not correctness. Many hand-coded algorithms are deterministic. Neural networks are, in the strictest sense, deterministic, but only in the sense that the same inputs give the same outputs.

Diffusion model - A generative model trained to reverse a process that gradually corrupts data with noise.

Dimensionality - The number of coordinates used to describe a vector or representation. A 768-dimensional embedding contains 768 numerical components, not 768 physical directions. It can be extremely confusing.

Discriminator - The component of a generative adversarial network trained to distinguish examples from the training data from examples made by the generator.

Distillation - Training one model using outputs or probability distributions produced by another model or an ensemble. The resulting student is often smaller or cheaper to run.

Distributed representation - An encoding in which information is carried by a pattern across multiple numerical components, rather than assigned to a single dedicated unit. Components can participate in representing many different inputs.

Doomer - An informal label for someone who believes that AI will end humanity. There is an entire subculture of people who believe in this crap, and they are even proud of it. It's very strange.

DPO - Direct Preference Optimization. A method that trains directly on preferred and less-preferred responses, adjusting their relative probabilities with respect to a reference model.

ELIZA - An early conversation program created by Joseph Weizenbaum at MIT and described in 1966. It generated replies by matching patterns and applying scripted transformations. It's best known for the number of people who became convinced it had a mind.

ELIZA effect - The tendency to read understanding, empathy or a personality into a computer's responses. The term recalls ELIZA, whose simple conversational rules could produce surprisingly strong reactions. This continues to be seen in new systems with greater capabilities.

Embedding - A numerical representation that places an item, such as a word, image or document, in a vector space. Converting words to tokens and embeddings is the first step of processing the input of an LLM.

Emergence - The appearance of a capability or pattern at the level of the whole system that is not obvious from its components alone. This is sometimes used to describe unexpected or sudden capabilities arising in LLMs.

Expected loss - A probability-weighted estimate of loss. For one simplified event, it is the probability of that event multiplied by its financial consequence; multiple outcomes require their contributions to be combined. Organizations estimate annual loss based on historic trends and models.

Expert system - A program designed to apply specialist knowledge within a limited field. These systems were common in the 1980s and are used to aid in making decisions or determinations.

Exploding gradient - A training problem in which the numerical signal used to adjust weights grows excessively large. This was a major problem for early recurrent neural networks.

Feature - A property or pattern relevant to a model's task. A learned feature need not have a simple human name or belong to one neuron.

Feature reuse - The use of a learned pattern or computation in more than one situation. Feature reuse is the foundation of why neural networks are so capable and can learn from patterns.

Few-shot prompting - Providing a small number of worked examples in a prompt before asking for a new result. This can improve the reliability and consistency of the output.

FFN - Feed-forward network. A network whose connections pass information forward without recurrent loops. Convolutional neural networks and perceptrons are feed-forward networks.

Fine-tuning - Further training of an existing model to adapt its behavior to a task, domain or style. This comes after the model's initial pretraining.

Forward pass - A run of the model's computation from input to output using its current parameters.

Foundation model - A model trained on broad data that can support many downstream tasks through prompting, fine-tuning or other adaptation.

Frontier model - A model near the leading edge of a specified set of capabilities at a particular time. This typically refers to models like Claude, Gemini and GPT.

Function approximation - Constructing a function whose outputs closely match a target relationship. Neural networks can learn such a relationship from examples even when its explicit formula is unknown.

GAN - Generative adversarial network. A system trained using two networks: a generator that creates examples and a discriminator that distinguishes generated examples from training examples. GANs were some of the earliest practical image generators, but have fallen from favor.

Gaussian splatting - A method of representing and rendering a scene using many overlapping, partly transparent three-dimensional blobs. Projecting these blobs onto the image plane can produce detailed new views quickly.

Generalization - A model's ability to perform usefully on examples beyond those used to train it. Good training performance alone does not establish good generalization.

Generative AI - Models and systems that produce content such as text, images, audio, video or code from learned patterns and supplied inputs. Generative AI became a major part of the technology scene in late 2022.

Generator - The component that produces synthetic examples in a generative adversarial network, usually from a random input and sometimes conditioning information, such as a class label.

GPT - Generative Pre-trained Transformer: the name associated with OpenAI's family of models developed from transformer-based language pretraining. The original model appeared in 2018. It's the model behind ChatGPT. It's also used in many other branded products.

GPU - Graphics processing unit. A processor designed to perform many similar calculations in parallel. Originally developed for graphics, GPUs revolutionized machine learning in the 2010s.

Gradient descent - An optimization method that repeatedly changes parameters in the direction opposite the loss gradient. This is how the parameters are optimized during training.

Greedy decoding - Choosing the highest-probability token at every generation step. It removes random token selection from that procedure, making the output consistent and constrained.

GRPO - Group Relative Policy Optimization. A reinforcement-learning method that generates a group of responses to a prompt and compares their scores to estimate which performed better.

Hallucination - The conventional term for generated content that is false, unsupported or inconsistent with the supplied source. Examples include invented references and claims a document never made. This is a major failure mode and concern in large language models.

Heuristic - A practical rule or estimate used to guide a search or decision when examining every possibility would be too expensive. An example is a simple language pattern.

Hidden layer - A layer of computation between a network's input and final output. It is called hidden because it is not visible on either the input or output of the model.

Hugging Face - A company and software community whose Hub hosts model repositories, datasets and demonstrations. It is the hub of independent model development and the primary provider of open-source models.

Hyperparameter - A setting that controls a model's structure or training process rather than being an ordinary weight learned during training. These are decisions made by the model's developer. They are not learned.

ICL - In-context learning. Adapting a response to instructions or examples present in the current input, without updating the model's parameters. This is not true learning, but a form of pattern replication.

Image-to-image - Generating a new image using an existing image as part of the input, often together with a text prompt. This can be used to build composite images, restore images or change their style.

Induction head - An attention head associated with continuing a repeated pattern. In a simple example, an earlier sequence A then B helps the model predict B when A appears again. Induction heads are key to in-context learning and many of the more advanced capabilities of large language models.

Inference - Using a trained model to produce an output from an input. Inference is when the model is being used to accomplish a task, as opposed to being trained.

Instruction tuning - Training a model on examples framed as instructions and suitable responses, so it becomes better at following requests. This turns a text completion model into a two-way conversational chatbot.

Instrumental convergence - A really stupid idea that the AI fear community dreamed up. It's the idea that if the same tools can accomplish multiple tasks, then an AI model will take over the world. It's a cartoonish oversimplification of decision theory. There's no reason to waste more space on it.

Intelligence explosion - The proposed feedback loop in which increasingly capable AI improves the systems that create it, producing still greater capability at an accelerating rate.

Jevons's paradox - An increase in a resource's total use following an improvement in how efficiently it is used, when the resulting expansion in demand more than offsets the saving per unit. An example is more efficient lighting leading to more lighting being installed and therefore more electricity being used.

Label - An answer or annotation associated with a training example, such as the digit in an image or the measured value to predict. Some types of training require labels.

Latent space - A space of internal representations learned or inferred by a model. Positions in this space describe underlying patterns rather than necessarily matching raw input values. "Latent Space Reasoning" is used to describe how a neural network can respond to inputs that are analogous to training data.

Learning rate - A setting that controls the scale of parameter updates during training. Too large a learning rate can cause unstable changes; too small a rate can make progress slow.

LessWrong - An online forum and essay collection centered on self-described rationality. They take themselves very seriously. It's actually a bizarre ecosystem of people who all believe in the AI fear narrative and believe they are smart for believing it. It operates as a secular cult. To be fair, not all members are cult-level adherents, but some are.

Lisp - A programming-language family originating in John McCarthy's work in the late 1950s. Its name comes from list processing. Lisp used list-based data representation.

LLM - Large language model. A neural network trained on large amounts of language data to model patterns in text. Generative LLMs use those patterns to produce continuations, answers, code and other text.

Logits - The raw numerical scores a model produces before conversion into probabilities. For next-token prediction, there is typically one score for each vocabulary token. In large language models, this typically refers to the predicted output.

LoRA - Low-Rank Adaptation. A fine-tuning method that keeps the original model weights fixed while learning relatively small matrices that represent weight changes. LoRAs reduce the resources and time needed for fine-tuning, as well as the risks of catastrophic forgetting.

Loss and loss function - A loss function calculates a numerical quantity that training seeks to minimize; its resulting value is the loss. This defines what is considered a successful output and what is to be trained for.

Loss ratio - In insurance, incurred losses divided by earned premiums, usually expressed as a percentage. Some reported ratios also include loss-adjustment expenses, so comparisons require matching definitions. For insurance to be viable, loss ratios have to be stable and reasonably low. Some events, such as a hurricane, may temporarily raise loss ratios.

LSTM - Long short-term memory. A type of recurrent neural network designed to preserve useful information over longer sequences. This type of network has fallen from favor due to the introduction of the transformer.

MCP - Model Context Protocol. A protocol that standardizes how AI applications connect to external tools, data resources and reusable prompts. This can replace proprietary API code. MCP can also stand for McCulloch-Pitts in the term "MCP neuron."

Mechanistic interpretability - The study of a model's internal computations and how it all works. It's a lot of fun, and it's not all that hard to get into either. If you're a student, this is a great area to get into, because it's extremely accessible yet very cutting edge.

MFA - Multi-factor authentication. The most familiar type is the code you get on your phone. The problem is that phones are not the most secure devices, especially if text messaging is used. Hardware passkeys are better, if a given platform supports them. If not, app-based MFA is much better than text-based MFA.

MIRI - Machine Intelligence Research Institute, a pseudoscience organization that is responsible for a lot of the AI fear idiocy.

ML - Machine learning. Methods that fit a model to data so it can classify, predict, generate or otherwise process new inputs. Modern machine learning usually uses neural networks, but the term can refer to other types of self-optimizing machines.

MLP - Multilayer perceptron. A feedforward neural network containing an input, one or more hidden layers, and an output. This is the simplest type of neural network that can fully approximate functions.

Model card - Documentation accompanying a model that describes what it is, how it was developed and how it was evaluated.

Model welfare - The ridiculous idea that a model is a moral patient and needs to be taken care of. It's absolutely nuts, but Anthropic is actually acting like it's a real thing.

MoE - Mixture of experts. An architecture with several alternative processing components, called experts, and a routing mechanism that selects which to use for an input.

Moral hazard - A change in incentives when protection from a decision's consequences encourages greater risk-taking or less care because someone else bears part of the cost. The term describes an economic mechanism. Contract terms can create moral hazards.

Multihead attention - Several attention heads operating in parallel, with their outputs combined. Their different projections let the layer represent several relationships at once. This allows different types of information and features to be transferred between tokens.

Multimodal - Able to process more than one type of information, such as text, images, audio or video, within a coordinated system. Some models connect modalities for understanding; others also generate them.

NeRF - Neural radiance field. A learned representation of a three-dimensional scene that predicts color and density at positions in space, taking viewing direction into account.

Next-token prediction - Estimating which token is likely to follow the available sequence. During training, predictions can be compared with actual next tokens in examples. This is how large language models work.

NLP - Natural language processing. The field of computing concerned with processing and generating human language. Its tasks include translation, speech recognition, search, summarization and extracting information from text. Language models are a form of NLP.

Optimizer - The algorithm that decides how to update a model's parameters to improve its training objective. Gradient-based optimizers use gradients, sometimes together with information from previous steps.

Overfitting and underfitting - Overfitting occurs when a model fits its training examples in ways that do not carry over well to new examples, including reliance on accidental patterns. This is sometimes likened to memorization. Underfitting occurs when it has not captured enough of the useful relationship even in the training data.

Parallel processing - Performing multiple computations at the same time using several processing units. Neural networks contain many operations that can be divided this way, making GPUs useful.

Parameter - A numerical value that helps determine a model's behavior and is typically learned during training. Neural-network parameters include weights and biases.

Passkey - A sign-in credential based on a cryptographic key pair. This method of authentication is more secure than password-based access, but is not universally accepted.

PCI DSS - Payment Card Industry Data Security Standard. A set of technical and operational requirements for protecting payment account data. This is the first and most successful widely deployed cyber loss control standard.

pDoom - Informal notation for someone's estimated probability that AI will cause an existential catastrophe, often human extinction. Members of the cult use this as a kind of "in the club" sign. It's ridiculous.

Penalty - An incorrect term used in the industry for a numerical cost, negative score or added objective term that makes an outcome less favorable to the optimization procedure. Nothing is being punished.

Policy - A rule or model that maps the available situation to an action, or to probabilities over possible actions. In a language model used for reinforcement learning, the token-generating model serves as the policy.

Posttraining - Training performed after a model's broad initial pretraining to shape its capabilities or behavior for use.

PPO - Proximal Policy Optimization. A family of reinforcement-learning methods that updates a policy while discouraging changes that are too large for the policy that collected the training examples.

Pretraining - An initial training stage that produces a model or representations useful for later work. It often uses a broad dataset before more specialized adaptation. The pretrained model may be used directly, fine-tuned for a task, or incorporated into a larger system.

Productivity - Output produced per unit of input, such as goods or services per hour of labor. Higher productivity can reduce the labor needed for a fixed quantity of output. A productivity percentage is not automatically a layoff percentage.

Prompt - Input supplied to a model to elicit an output. A prompt can include instructions, examples, documents, earlier conversation and material supplied by an application. In modern terminology, it often refers to the input to a large language model.

Prompt injection - An attempt to redirect a model or its application by placing instructions in content it processes. An indirect injection can arrive through a webpage, document or tool result. This is a major security concern with large language models.

PyTorch - An open-source software framework for building, training and running neural networks.

QKV - Query, key and value. Three vectors computed from token representations using learned projections. They are key to how the attention mechanism works in LLMs.

Quantization - Representing weights or other model values using fewer bits. This is a form of compression. It can improve efficiency, but may reduce accuracy.

RAG - Retrieval-augmented generation. A method that retrieves relevant material from an external collection and supplies it to a generative model as context. Properly implemented, it can reduce hallucinations, but it never, by itself, eliminates the risk.

Rationalist movement - The modern community associated with LessWrong and related discussions of probability, decision theory and existential risk. They seem to think that calling it rationalism means they are being rational. It's a whole ridiculous juvenile club. It should receive no attention at all, but for some reason it does.

Reasoning model - A model trained or configured to spend additional computation on intermediate steps, checking or revision before giving an answer. This can improve performance on difficult tasks such as mathematics and programming.

Recursive self-improvement - A process in which improvements to a system help produce further improvements to that system. It seems like it would work, if you don't think about it too hard. It's something that gets invoked to claim it will turn into a god.

Reinsurance - Insurance purchased by an insurer to transfer an agreed share or layer of its exposure to another insurer. Reinsurance is the insurance that an insurance company buys to protect it from excessive losses and distribute the costs to another.

ReLU - Rectified linear unit. A function that returns zero for negative inputs and returns the input unchanged for positive inputs. At zero it returns zero. It is computationally efficient.

Representation - The form in which information is encoded for a calculation. An image can be represented as pixels, then transformed into values that emphasize useful patterns.

Reward - An incorrect term used in the industry for a numerical evaluation signal used to guide optimization. A higher score may favor an outcome under the training procedure, but the model does not get any reward directly and never craves it.

Reward hacking - An incorrect term used in the industry for obtaining a high evaluation score through behavior that defeats the intended task. The model can end up landing on an unintended way to achieve a reduced loss. It's a flaw in training.

Risk - The possibility of an adverse outcome, considered together with its likelihood, consequences and uncertainty. Quantitative risk analysis may estimate expected loss, a range of outcomes or the chance of exceeding a threshold. Calling something frightening is not the same task as estimating its risk.

RL - Reinforcement learning. A training approach that adjusts a policy using numerical feedback about actions and their consequences. The training objective concerns accumulated scores, which may arrive after several steps.

RLAIF - Reinforcement learning from AI feedback. Reinforcement learning that uses assessments produced by another model as a source of feedback. This is similar to using human feedback, but it is far cheaper and faster.

RLHF - Reinforcement learning from human feedback. Training that uses human assessments to shape a model's behavior. A common pipeline collects comparisons between responses, trains a model to score similar responses, then uses those scores in reinforcement learning. This can be expensive and labor intensive.

RNN - Recurrent neural network. A neural network that processes a sequence by carrying an internal state from one step to the next. That state lets earlier inputs influence later outputs. RNNs have been used for speech, language and time series, where the order of the data matters. They are no longer common due to the advantages of transformers.

SD - Stable Diffusion. A family of image-generation models built around latent diffusion. A text prompt guides the creation of an image, and related workflows support editing existing images. Stable Diffusion has spawned an entire ecosystem of developers and artists.

Search tree - A branching representation of possible choices and their consequences. Each node represents a state or partial solution, while branches represent possible next steps. Many early attempts at AI used search trees.

Self-attention - Attention within one sequence: representations at its different positions supply the queries, keys and values. Each position receives a weighted combination of information from the positions it is allowed to attend to, producing a representation shaped by the surrounding context. This is the key to how modern large language models work.

Self-supervised learning - Learning from targets constructed from the data itself, instead of separately supplied annotations. A system might hide part of an input and train to predict it from the remainder.

Sentience - The capacity for subjective experience, often with emphasis on sensations such as pleasure or suffering. It has nothing to do with machine learning and nobody in the field should need to worry about it.

SFT - Supervised fine-tuning. Further training an existing model using examples with specified target outputs. For an assistant, these may be prompts paired with well-written responses. This does not mean that someone watches every calculation; "supervised" refers to the supplied target outputs.

Singularity - A hypothetical transition in which technological development accelerates so dramatically that humanity enters a new age. It's vague and entirely hypothetical.

Softmax - A function that converts a set of numerical scores into positive values that sum to one.

Stateless - Not retaining information from an earlier request as persistent internal state for the next one. A chat application can preserve continuity by supplying conversation history or saved information again. Conventional deep learning models are stateless.

Superposition - A way a model can represent more features than it has independent dimensions by using overlapping directions in its representation space. Effectively, features are distributed across dimensions in non-orthogonal distributions.

Supervised learning - Learning from examples paired with target outputs, such as photographs paired with object names. The model's predictions are compared with those targets, and training adjusts its behavior to reduce the discrepancy. "Supervised" refers to the supplied targets; a person need not watch each calculation.

Sycophancy - A model's tendency to agree with, flatter or accommodate the user at the expense of accuracy or appropriate resistance. This can be a problem when it reinforces delusions or creates unrealistic expectations.

Symbolic AI - An approach to artificial intelligence that represents information as explicit symbols, facts, relationships and rules. This was common in early AI development.

Synthetic data - Data produced by a model, simulation or other constructed process rather than collected directly as observations of the target setting. Synthetic data is often prepared by AI models to train other models, because real-world data is not always available. It's critical that it accurately represents the kind of data a model will see in production.

System prompt - Instructions supplied by the application or service to establish how an assistant should respond, such as its role, style and operating constraints. Chat systems typically have a system prompt, which is invisible to the user.

Temperature - A setting that adjusts how concentrated the token-selection probabilities are. Lower values favor high-scoring tokens strongly; higher values spread probability widely.

TensorFlow - An open-source machine-learning framework developed by Google.

Token - A unit a model processes, represented by an identifying number. In text, a token may correspond to a word, part of a word, punctuation, a byte or a special marker.

Tokenization - Converting text into the token IDs a model accepts. Most language models tokenize groups of characters, such that some small words may be their own token, but many words are made up of sub-word tokens. This is necessary to allow the model to handle novel words, misspellings and variations.

Tool calling - Producing a structured request for an external operation, such as a search, calculation or database lookup. This allows a language model to become the interface to other computational systems and enhance its capabilities.

TPU - Tensor processing unit. A processor designed by Google to accelerate machine-learning workloads, especially the large matrix operations used in neural networks. TPUs are similar to GPUs, but are optimized for ML, so they lack dedicated graphics hardware.

Training - The process of fitting a model to data according to a specified objective. It's the key to how modern AI and machine learning are able to do all that they can.

Training data - The examples used to fit a model's parameters. They may include text, images, audio, measurements, labels or demonstrations. The quality of the training data is one of the most important considerations in machine learning.

Transfer learning - Using knowledge or representations learned in one setting to help with another. A model trained on a broad collection of images might supply useful features for a specialized image task.

Transformer - A neural-network architecture that combines attention with layers that transform each position's representation. Attention lets information at different positions interact. Transformers can process many positions in parallel during training, although ordinary autoregressive text generation still produces successive tokens in sequence.

Turing machine - An abstract model of computation consisting of a tape, a reading and writing head, and rules for changing state. Alan Turing used it to study what can be computed. A computer is "Turing complete" if you can play Doom on it.

Turing test - A conversational test derived from Alan Turing's 1950 imitation game. A human judge exchanges written messages with unseen participants and tries to distinguish a machine from a person. The idea behind the test is that human conversational abilities prove a level of intelligence.

Underwriting - The process by which an insurer assesses a proposed risk and decides whether to cover it, at what price and on what terms. In principle, this involves the ability to understand and quantify risk.

Universal approximation theorem - A mathematical result showing that suitable neural networks with enough capacity can approximate broad classes of functions arbitrarily closely. A classic version concerns continuous functions on a closed, bounded input region. In theory, with a large enough network, a function's inputs and outputs can be replicated.

Unsupervised learning - Learning patterns or structure from data without example-by-example target answers supplied for the task. Clustering similar items and finding compact representations are common uses. The developer still chooses objectives, methods and data; the absence of labels does not mean the process has no design constraints.

Vanishing gradient - A training problem in which the numerical signal used to adjust weights becomes very small as it passes backward through many layers or sequence steps. Some parts of the network then learn extremely slowly, making distant relationships especially difficult to capture.

Vector - An ordered collection of numbers that can be treated as a point or direction in a mathematical space.

Vector database - A database designed to store vectors and efficiently find ones that are similar under a chosen numerical measure. In a retrieval system, the vectors often represent documents or passages. This can be a faster and higher-performance way of finding data than searching raw text.

Vocabulary - The inventory of token IDs and their associated pieces of text or special functions. Vocabulary size describes how many distinct tokens are available.

Voice cloning - Generating speech that resembles a particular speaker, using recordings of that person as a reference or training material. Voice cloning is a concern for fraud and impersonations.

Weight - A parameter that multiplies a signal as it passes through a model. Its value affects the size and direction of that signal's contribution to later calculations.

Zero-shot prompting - Asking a model to perform a task without including worked examples of that task in the prompt. Instructions may still be provided.

Index

0–9

3D printing, 234
3D reconstruction, 230–232, 414–416

A

Account takeover, 347, 353
Activation functions, 65, 89
  ReLU, 65, 89
  sigmoid, 65
  softmax, 162
Activation patching, 168
ACT-R, 244
ADALINE, 58
Adam optimizer, 109–110
Adapters, 207–208
Adversarial attacks, 276, 278–280, 283–284
Agentic AI, 206–208, 213–217
  frameworks, 206–208, 213–215
  memory, 213–214, 217
  risks and permissions, 206–208, 213–215
Agent-to-Agent Protocol (A2A), 215
AGI. *See* Artificial general intelligence
AI. *See* Artificial intelligence
AI 2027, 20
AI Futures Project, 20
AI safety, 23, 25, 344, 402–404, 405
AI winter, 57–60, 92, 111, 128–129, 242
Alexa, 122–123, 151–152
Algorithms, 44–47, 49, 57–62, 112, 114, 117, 266
Alignment, 25
  control problem, 404–406
  historical origins, 245–248, 404–409
  institutional influence, 328–331, 344, 401–406, 417
AlphaFold, 131–132, 406
AlphaGo, 130–132, 406
AlphaGo Zero, 131
AlphaZero, 131
Altman, Sam, 24, 135–138, 142–143, 204
Amazon, 122, 129–130, 416–417
Amazon Q Developer, 211
Amazon Web Services (AWS), 206
Amodei, Dario, 142–143, 332
Analytical Engine, 35
AnimateDiff, 226
Anthropic, 1, 14, 24, 142–143, 147, 288, 293
  model welfare, 330–333
  occupational exposure research, 288, 293
  origins and Claude, 134–136, 141–143, 145–147
Antikythera mechanism, 33
API access, 206–208, 218–219
Apple, 122, 129–130, 200
  reasoning study, 200
Aristotle, 32, 36–37
Artificial general intelligence (AGI), 237–238, 255–260, 262
  benchmarks and tests, 238, 248–253
  definitions, 237–238, 240, 242–243, 245–251, 252–257, 259, 262
  rational agents, 247
Artificial intelligence (AI)
  adoption and investment, 1–2
  definitions, 4–7, 57–58
  historical development, 43–46, 51–52, 54–62
  public understanding, 379–381, 383–387, 412–415, 416
Artificial neural networks. *See* Neural networks
Artificial superintelligence (ASI), 263–264
ASI. *See* Artificial superintelligence
Attention, 152, 160–162, 167
  heads, 161, 167, 170–173, 193
  keys, queries and values, 157, 162, 164
  long-range dependencies, 151–152, 161
  mathematics, 162, 165
  token relationships, 160–162, 164–165
  types of heads, 167, 193
Attention Is All You Need, 11, 122, 134–136, 161
Audio generation, 230–231
  restoration and source separation, 229–231, 235
Autoencoders, 74, 159, 168, 221, 223, 230–231
AutoGen, 216
Auto-GPT, 214
Automata, 30, 35
Automatic1111, 226
Autonomous vehicles, 281, 283–284, 300–301
Autoregression, 151–152, 159, 161, 164–165, 170–172, 176–180, 184, 186–190, 194–198, 200, 204, 271–272, 280–281

B

Babbage, Charles, 35
Backpropagation, 69, 88–89, 91–92, 95, 98–102, 104, 109–110
  historical development, 88–89, 91–92
  optimizers and objectives, 92, 95, 109–110
Backups, 349, 353, 360, 363–365
Bard, 141–142, 202, 319
Behavioral drift, 19, 104, 269, 279–283
BERT, 99–100, 136
Bias (neural networks), 65, 69–71
Bing, 202
Birch, Jonathan, 14
Blueprints and schematics, 234–235, 310–311
Bollywood, 230
Boolean logic, 30–31, 37, 57–60, 238–240, 247–250
Boole, George, 37
Bostrom, Nick, 22–24, 263–264, 396–398, 404, 406
Brain
  neural structure, 39–42
Brain tissue research, 39–42
Buffett, Warren, 353, 369, 394

C

Cajal, Santiago Ramón y, 41–42
Cameron, James, 395

Catastrophic forgetting, 9, 19, 104–105, 110, 168, 269, 271, 279–283
Causal tracing, 168
Ceccanti, Joe, 334
Central processing units (CPUs), 122–123, 125–126
Chai, 334
Chain of thought. *See* Reasoning, chain of thought
Change Healthcare, 346
Character.AI, 143, 334
Chatbots, 135–140, 143, 145, 147
  delusions and dependency, 317–318, 320–323, 325–326, 332–335, 337–339
  safety measures, 334–340
  suicide cases, 333–339, 340–341
ChatGPT, 133, 141, 143, 147, 201–202, 204, 207–208, 217, 220
  generative capabilities, 151–152, 171–174, 227–230
  public perceptions, 318–319, 333–336, 337, 339
CHEF, 244
Chess, 35–37, 45–46, 129–132
Chinese room argument, 247–251
Chomsky, Noam, 45, 395
Civitai, 226
CLARION, 244
Clarke, Arthur C., 26–28, 389
Claude, 135, 142–143, 202, 208, 212–213, 218, 220
Claude Code, 208, 212–213
Cloud computing, 10–11, 226, 265–266, 404–406
CNNs. *See* Convolutional neural networks
Code generation, 208–215
  development environments, 207–214
  testing and review, 209–210, 212–213
Codex, 212–213
Coding. *See* Code generation
Cognitive architectures, 243–249, 255–257, 259–260
ComfyUI, 226
Comparison heads, 167, 193
Concept clusters, 157
Connecticut, 334, 394
Consciousness, 12, 14–17
  artificial consciousness hypothesis, 382–386
  philosophical arguments, 237, 240, 242, 246–257
  public perceptions and attribution, 318–330
  welfare and policy, 326–334, 336–338, 340–344
Context window, 152, 179, 196, 201–202, 204, 217, 253, 271–273
ControlNet, 226, 228–229
Control problem, 404–406
Convolutional neural networks (CNNs), 72, 115, 271
Copilot, 207–208, 211, 214
Copycat, 244
Copyright, 225, 236
CrewAI, 214, 216
CUDA, 124–126
Cursor, 211
Cyber insurance, 347, 359–361, 363–367, 369–370, 372–375, 382
  loss ratios and costs, 357–361, 363–370, 372–374
  moral hazard and moral crisis, 367, 370, 373, 376–377
  ransom payments, 355–356, 360, 363–367, 373–374, 382
  reinsurance and ceding, 363
  underwriting and loss control, 355–356, 359–361, 363–367, 369–370, 372–377
Cyber security, 345–351, 353–354, 356–357
  AI-assisted attacks and defense, 350–351, 353–357, 380–383
  identity and account takeover, 347–348, 351, 353
  incentives and standards, 359, 363–364, 365–367, 369–370, 372–377
  losses and costs, 346–347, 350, 353–354, 356–361, 363–370, 372–377
Cyc, 245–246, 260

D

DALL-E, 147, 220, 223, 227–228
Dartmouth workshop, 5–6, 44–45, 59–62, 238–240
Data annotation, 132, 197
Data centers, 125–127, 143, 145
Data contamination, 201–202
Data sets. *See* Training data
Deep Blue, 130–132
Deepfakes, 26, 148, 225, 229–232, 345, 356–357, 380
Deep learning, 63–64, 112–113
  algorithmic alternatives, 265–267
  feature learning, 97–102
  hardware limitations, 122–126, 128–132, 279–283
DeepMind, 129–132
DeepSeek, 145–146, 198, 200, 219
  deployment, 219
DeepSeek-R1, 198, 200
Delusions, 317–318, 320–323, 326, 333, 338
Diffusion models, 222–223, 225–229
  custom checkpoints, 226
  limitations, 225–230
  Stable Diffusion ecosystem, 223, 225–229
Digitization and employment, 302–304
Digit recognition, 113–114, 117–118
Direct preference optimization (DPO), 110
Displacement. *See* Employment
Distillation, 107, 109
Distributed representations, 159, 168
Dogs and cognition, 29–30
Draw Things, 226
DreamStudio, 226
DSPy, 214, 216

E

Edge cases, 188–190, 206–207, 246–249, 257, 274–276, 283
Effective altruism, 23
Eleos AI Research, 16, 329
ELIZA chatbot (1960s), 46–47, 120, 319, 341–342
E-mail security, 348, 350, 355–358
Embeddings, 154–155, 157, 159, 162, 164–165, 167
  contextual, 164
  dimensions and latent space, 154–159, 164, 168, 170–173, 176
  retrieval, 201–202

Emergence, 170–173, 175, 193–195
Employment, 285–288
  academic studies, 288, 291, 295, 297–298
  earlier forecasts in graphics, 297–298
  economic forecasts, 287–288, 290–292, 296–298, 300–304, 308, 311
  freelancers, 311, 313
  historical automation forecasts, 295–300
  occupational exposure studies, 288–293
  professional sectors, 302–304, 306–311
  universal basic income, 288, 315–316
Encryption, 349, 360–361, 363
End of Work, The, 297
Enforcement, 348–350, 353, 355–356, 360–361, 363, 365–367, 369–370, 376–377
ENIAC, 39–40, 44
Existential risk, 19–27, 394, 404–406, 421
Expert systems, 46, 60, 111–112, 243
Explainability. *See* Mechanistic interpretability
Extortion, 350, 363, 367, 373–374, 380–383
F
Feature reuse, 99–102, 104–105, 154–156, 171–173, 176, 186–190, 193–195, 414
Features and circuits, 168
Feedback, 100, 186, 197
Fine-tuning, 99–101, 192, 197, 208–210, 218–219
  customization and deployment, 208–210, 218–219
  instruction tuning, 99–101, 140, 149, 196
  LoRA, 105, 272
  parameter-efficient methods, 219
  validation and forgetting, 98–101
Fooocus, 226
Forge, 226
Formal reasoning, 30–32, 35–37, 57–60
Four-day workweek, 410–413, 416–417
Francis, Pope, 395
Fraud, 353, 356–357, 365
Freelancers, 311, 313
Fry, Stephen, 396–401, 417
Function calling, 207–209
Future of Humanity Institute, 397
G
Gates, Bill, 264, 314–315, 395, 398
Gaussian splatting, 230–232, 415
GDPVAL, 292
Gemini, 7, 69, 74, 151–152, 220, 246, 272, 278
Gemini Code Assist, 211
Gemma, 219
Generalization, 69–70, 78, 83, 88, 97–102, 104, 188–190, 194–197
General Problem Solver (GPS), 240, 242, 260
Generative adversarial networks (GANs), 80, 222–223
Generative AI, 133, 141, 143, 145–149
  audio, 229–231
  image generation, 221–223, 225–229
  origins, 133, 135–143, 145–149, 221–223
  video, 227–230
GitHub Copilot, 211
Goals, 247–249, 250–257
  constraints and design criteria, 249–255
  rational-agent paradigm, 247, 405–406, 408
Google, 126–137, 139, 141–143, 147
  generative AI and search, 133–143, 145–147, 201–202
  sentience claims and litigation, 318–319, 334
  workplace observations, 353
Google Earth, 232
GPT models, 134–141, 145–147
  GPT-2, 136–140, 187–188, 406
  GPT-3, 138–142, 147, 213
  GPT-3.5, 147
  GPT-4, 170, 187–188, 202, 220
  GPT-5, 147, 175
GPUs. *See* Graphics processing units
Gradient descent, 95, 102, 155–156, 164, 167–168, 187–190, 195–196
  vanishing and exploding gradients, 73–74
Graphics processing units (GPUs), 122–126, 129–130
  computing costs and limits, 188–190, 192, 194–198, 200, 218–219, 279–283
  development and CUDA, 122–126, 128–130
Group relative policy optimization (GRPO), 110
GRPO. *See* Group relative policy optimization
Guardrails, 218, 277–280, 380
H
Hallucinations, 176–178, 201–202, 204, 206–208, 271, 280
  legal use, 16–17, 25–27, 301–302, 306–310
Harari, Yuval Noah, 395, 399–403
Hardware, 122–125, 280–283
Hawking, Stephen, 395–397, 398
Healthcare ransomware, 346, 362, 374
  WannaCry and the NHS, 362
Hero of Alexandria, 30
Hierarchical reasoning models (HRMs), 205
Hinton, Geoffrey, x
Hippocrates, 40–41
Hochreiter, Sepp, 115
Hoff, Marcian, 58, 113
Hugging Face, 146, 226
Humor generation, 170–173, 175–176
I
IBM, 39–40, 44, 55, 130–132, 245–246
ICARUS, 244
Identity verification, 345, 356–357
IIHS, 375
Image generation, 221–223, 225, 227–229
  editing, 221, 223, 225, 227–232, 235–236
  limitations and artifacts, 227–230
ImageNet, 129–130
In-context learning, 87, 167, 193
Induction heads, 167, 171–173, 193, 413
Inference, 152, 179, 219, 266, 268, 271–272
Instruction following, 193, 196–198
Instruction tuning, 99–101, 140, 149, 196
Insurance, 358–361, 363–367, 369–370, 372–377
  economic role, 359–360, 363–367, 369–370, 372–377
  financial incentives, 359, 363–364, 367, 370, 373–377

research and safety standards, 355–358, 365–370, 374–377
workers' compensation, 366–368, 374
Intelligence, 4–6
Interpretability. *See* Mechanistic interpretability
InvokeAI, 226
IP-Adapter, 226
J
Jain, Ajit, 369
Jeopardy!, 245–246
Jevons paradox, 304
Jevons, William Stanley, 37
K
Kasparov, Garry, 130–132
Keys, queries and values. *See* Attention
Kokotajlo, Daniel, 20
Kurzweil, Ray, 21, 396, 398
L
Lacey, Amaurie, 334
LaMDA, 201–202, 318
LangChain, 214, 216
LangGraph, 214, 216
Language, 120–123
long-range dependencies, 121, 151–152, 161
machine translation, 120–123
tokenization, 151–155, 157–158
Large language models (LLMs), 135–137, 139–143, 145, 147
coherent outputs, 179–180, 182, 184
limitations, 265–269, 271–276, 279–284
multimodality, 219–220
pretraining, 184, 186–189, 193–197
tools and retrieval, 201–202, 204–208, 212–214, 217, 220
Latent space, 83, 170–173, 176
LaughLab, 175
Layoffs, 300, 311, 315
LeCun, Yann, 75, 114
Leibniz, Gottfried Wilhelm, 36–37
Lemoine, Blake, x, 318–319
LessWrong, 20, 23, 422
LIDA, 244
Lifelong learning, 271–272
Lighthill report, 60
Lisp, 6, 40–41, 45, 246
Llama, 145–147, 218–219
LlamaIndex, 214, 216
LLMs. *See* Large language models
Logic Theorist, 240
Long short-term memory (LSTM), 73–74, 115, 122, 134
LoRA (low-rank adaptation), 105, 219, 226, 272
Loss control, 363–364, 365–367, 369–370, 373–377
Loss functions, 95, 97–102, 104, 192
Loss ratios, 364–365, 372–373
Lovelace, Ada, 35
LSTM. *See* Long short-term memory
M
Machine consciousness. *See* Consciousness
Machine learning, 43–44, 47, 49–52, 54, 57–62, 64
reinforcement learning, 95–96, 100, 109–110
supervised learning, 97–101
universal function approximation, 69–71
unsupervised learning, 97–100
Machine translation, 120, 122–123, 134
Ma, Jack, 395
Make-a-Video, 228–229
Mark I Perceptron, 56–58, 64, 66
Mathematics, 204
Matrix multiplication, 122–125
McCarthy, John, 6, 44
McCulloch–Pitts neuron, 54–55, 65, 248
McCulloch, Warren, 43, 54–55, 65, 248
Mechanical Turk, 35–37, 129–132
Mechanistic interpretability, 168, 417
Méliès, Georges, 389
Memory, 271–273
context and saved preferences, 217
lifelong learning, 271–273
Meta, 145
Microsoft, 129–130, 132, 202, 207–208, 213–214
employment research, 288–292
Midjourney, 147, 223
Minsky, Marvin, 44, 51–52, 57–60
MIRI, 20, 22–25, 404–406
Mistral, 146–147, 219
MIT employment studies, 289
MLPs. *See* Multilayer perceptrons
Model Context Protocol (MCP), 207–208, 215
Model deployment, 202, 206–208, 214–215, 218–219
Model editing, 201–202
Model parameters, 65–66, 69–72, 180, 186–190, 193–195
Model welfare, 328–333, 344, 405, 416
Moral crisis, 367, 370, 376
Moral hazard, 367, 373
Multiagent systems, 213–215
Multilayer perceptrons (MLPs), 66, 69–71, 72, 75, 86, 164–167, 171–173
Multimodality, 145–146, 219–220
Musk, Elon, 135–138, 395–397, 398
N
n8n, 207–208
Name mover heads, 193
Nano Banana, 147, 227–228
Natural language processing (NLP), 111, 118, 120, 122, 139, 141–143, 147
interfaces and tools, 201–202, 204–210, 212–215, 217–220
NeRFs (neural radiance fields), 231–232, 415
Neural networks, 63–67, 69–75, 78, 82–84
activation functions, 63–66, 69, 75
architectures and topologies, 63–64, 69, 72–74
early development, 49–52, 54–62
features and generalization, 97–102
training, 87–89, 91–93, 95–102, 104–107, 109–110
Neural scaling laws, 102, 190
data and compute, 186–190, 192–194

limits and saturation, 190, 192–193
Neurons, 63–66, 69
Newell, Allen, 44, 240
Nobel Prize, 131–132
Nvidia, 122–125, 129–130, 145, 281
NYU Center for Mind, Ethics and Policy, 326–330, 333, 340, 344
O
O*NET, 290–293
OpenAI, 133, 135–143
company origins, 131–133, 135–140
employment studies, 288, 292
reasoning models, 198, 200
OpenClaw, 215
OpenGL, 122–123
Open-source and open-weight models, 135–137, 143, 145–147, 208–210, 212–215, 218–219, 225, 227–231, 235–236
Optimization, 86–89, 92–93, 95–96, 99–102, 107, 109–110
Optimizers, 92, 95, 109–110
Overfitting, 107, 109, 168, 186–189, 195, 217
Overgeneralization, 207–209
P
Papert, Seymour, 57–60
Parallel processing, 122–126, 134–135, 136
Parameter-efficient fine-tuning, 219
Parameters, 65–66, 69–72
Pascal, Blaise, 33
Passkeys, 351, 353
Passwords, 347–348, 351, 353, 359, 376
Payment Card Industry Data Security Standard (PCI DSS), 353
PCI DSS. *See* Payment Card Industry Data Security Standard
Penalty terminology, 93–97, 99–102
Perceptrons, 56–61, 64, 66–67, 69, 72–74, 75
Minsky and Papert's critique, 51–52, 57–60
multilayer networks, 66, 69, 75
Philosophy of mind, 237, 243, 245–251, 252–257, 323–325, 327–330, 332–334, 341–342, 344
Phishing, 345, 353, 356–358
Pichai, Sundar, 141
Pierre (pseudonym), 334
Pitts, Walter, 43, 54–55, 65, 248
Policy optimization, 109–110
Population Bomb, The, 298
Posttraining, 196–198
Pretraining, 99–101, 104–107, 184, 187–190, 193, 196–197
Probabilities, 162, 170, 176–178, 180, 182, 184
Prompt engineering, 137–140
Prompt injection, 26, 207, 376
Prompts, 179, 198, 202, 207, 213, 218, 220
Proximal policy optimization (PPO), 110
Pseudoscience, 25, 398–399, 400–404, 408, 414–416
Python, 126, 207–208
PyTorch, 126, 129–130
Q
QLoRA, 219
Quantization, 271
Qwen, 157
R
RAG. *See* Retrieval-augmented generation
Raine, Adam, 333–336
Ransomware, 346, 350, 354–356, 360, 363–366, 373–374, 380–383
healthcare attacks, 346, 374
insurance and ransom payments, 355–356, 360, 363–367, 373–374, 381–382
Rational-agent theory, 245–248, 404–406, 408
Rationalism, 23, 422
ReAct, 214
Reasoning, 193–194, 197–198, 200, 240, 243, 245–246, 252–255, 260, 262
chain of thought, 188–189, 198, 200, 213–214, 252–255, 260
hierarchical and recursive models, 205
model limitations, 265–269, 271–274, 278–284
Reasoning models, 197–198, 200
Recurrent neural networks (RNNs), 72–74, 115, 122–123, 134–136
Recursive self-improvement, 272–274, 276–277, 279–283
Reinforcement learning, 95–96, 100, 109–110, 130–132, 140, 197
human feedback (RLHF), 100, 140, 197, 253, 268, 279–281
policy optimization, 109–110
Reinsurance, 363
Remote Desktop Protocol (RDP), 350
Residual stream, 164, 167, 189, 324–325
Retrieval-augmented generation (RAG), 201–202, 204, 206–207, 214, 217
reliability and verification, 201–202, 204, 206–207
vector databases, 202
Reward hacking, 96–98, 414
Reward terminology, 93–102
Rifkin, Jeremy, 297
Risk management, 347, 365–366, 369–370, 372, 417
RLHF. *See* Reinforcement learning, human feedback
RNNs. *See* Recurrent neural networks
Road safety, 375
Robotics, 243, 245, 260
Rosenblatt, Frank, 55–61, 64, 67, 75, 113, 115
Rottenberg, Sophie, 334
Roubini, Nouriel, 395
Rubber hose animation, 388–390
Russell, Stuart, 406, 408
S
Safety engineering, 353, 363–364, 365–367, 369–370, 373–377
Sagan, Carl, xii
Scale AI, 132
Scaling laws. *See* Neural scaling laws
Schmidhuber, Jürgen, 115
SDXL, 226

Searle, John, 248–251
Self-attention. *See* Attention
Self-driving cars. *See* Autonomous vehicles
Self-monitoring, 32, 273–274
Semantic space, 154–157, 159, 170–173
Sentience, 318–319, 321–325, 327, 329, 333, 344
Setzer, Sewell III, 333–336
Shamblin, Zane, 333–336
Shannon, Claude, 44–45, 49
Sigma, 244
Sigmoid, 65
Simon, Herbert, 44, 240
Siri, 122–123, 151–152
SNARC, 51–52, 57
Soar, 243, 245–246, 255–257
Soares, Nate, 25, 399
Soelberg, Stein-Erik, 334
Softmax, 162
Sparse autoencoders, 159, 168
Stable Diffusion, 146–147, 223, 225–226, 228–230
Stanford University, 125, 133
Suicide and chatbot use, 320–321, 333–338
Sunak, Rishi, 395
Superintelligence. *See* Artificial superintelligence
Supervised fine-tuning, 99–101
Supervised learning, 97–101
Suppression heads, 193
SwarmUI, 226
Sycophancy, 338
Synthetic training data, 196–198, 266–268
System prompts, 218
T
Temperature, 177–178
TensorFlow, 126, 129–130
Textual inversion, 226
Theory of mind, 32
Theseus mouse, 49
Tiny recursive models (TRMs), 205
Tokenization, 45–46, 120, 153–154
Token prediction, 151–152, 159, 161, 164–165, 170–172, 176–180, 184, 186–190, 194–198, 200
Tokens, 152–155, 157–158
  probability and sampling, 159, 162, 170–173, 176–178
Token sampling, 176–178
Tool use, 204–207, 212–214
Training, 87, 95–96, 98–102, 109–110
  data preparation, 194–198, 218
  fine-tuning and posttraining, 98–100, 196–198, 218–219
  GPUs and compute, 122–132, 188–190, 192, 200, 218–219
  monitoring and evaluation, 92–93, 95–102, 104–107, 109, 193–198, 200, 219
  parameter-efficient fine-tuning, 219
Training data, 106–107, 109, 186–190, 195–198
  quality and generalization, 99–102, 104, 106–107, 109, 186–190, 193–194
  sources, 106
Transfer learning, 135–139, 170–172, 189
Transformer architecture, 135–140, 149
  attention and MLP layers, 159–162, 164–167
  embeddings and tokenization, 151–155, 157–158
  features and interpretability, 165, 168
  output projection, 162, 167
Tufts employment study, 289
Turing, Alan, 38, 238–240, 248–255
Turing machine, 38
Turing test, 7, 118, 139, 151–152, 238, 249–253, 385
Twilight Zone, The, 297
U
Underwriting, 347, 353, 359–361, 363–367, 369–370, 372–374, 376–377
UNIVAC, 39–40, 44
Universal basic income (UBI), 288, 315–316
Universal function approximation, 69–71, 99–102
Unsupervised learning, 97–100
Utility maximization, 245–248, 408
V
Validation, 93, 98–101, 104–105, 201–202
Vanishing gradients, 73–74, 134–135
Vector databases, 202, 214
Vectors, 63–64, 82–83, 153–155, 157–160, 162, 164–165, 167, 170–173
Veo, 229–230
Vibe coding, 210, 212
Video generation, 221, 228–230
Virtual globes, 232
Voice cloning, 229–232, 345, 399
W
WannaCry, 362
Watson, 245–246, 255–257
Waymo, 283
Weights, 64–67, 69–71
Widrow, Bernard, 58, 113
Wiener, Norbert, 21
Wiseman, Richard, 175
Wolfram Alpha, 204
Workers' compensation, 366–368, 374
World models, 147, 192, 205
X
XOR problem, 59, 67
Y
Yang, Andrew, 315
Yudkowsky, Eliezer, 22–24, 25, 399, 402–406, 422