AI | Human Ethics

AI History Companion Guide

INTRO: AI has been multiplying the human mind for generations

Artificial intelligence has evolved over approximately 75 years, built by generations of researchers, engineers, and visionaries. Despite dramatic advances in its capabilities, the fundamental concept has remained consistent: AI systems excel at predicting the next best word, label, or action based on patterns learned from data and the instructions we provide.

Modern systems like ChatGPT are vastly more powerful than early chatbots such as ELIZA from the 1960s, yet they share a common foundation. Both require input, reference patterns from their training, and predict what should come next. The difference lies in scale, sophistication, and the enormous amount of data modern AI can process—not in the underlying principle.

AI has always been a tool and continues to be a tool to accelerate human thinking, not to replace it. By understanding how AI has evolved while fundamentally staying the same, you will realize the importance of your role in working with AI to get the best possible output. By looking back through history, you will have a better understanding of the AI tools you are using today.

To work effectively with AI, you will need to learn skillful prompting and ethical application. But first, it is important to understand the history, mechanics, and implications of AI development to help you become a more informed and capable user. Then we will move on to learn prompting and safe, ethical use guidelines.

The Foundations of AI: 1950–1960s

Alan Turing and the Question of Machine Intelligence (1950)

In 1950, British mathematician Alan Turing published "*Computing Machinery and Intelligence*" in the journal *Mind*, posing the deceptively simple question: "*Can machines think?*" Rather than debate abstract definitions of "thinking," Turing proposed a practical test: *if a human evaluator cannot reliably distinguish a machine's text responses from those of a human, the machine can be considered functionally intelligent*.

This concept, now known as the Turing Test, shifted the conversation from philosophy to measurable behavior and led to the CAPTCHA tests that we must pass when using the Internet.

Turing worked on code-breaking machines to decrypt Nazi ciphers during World War II. His approach to using machines as accelerators of human thinking and analysis laid crucial groundwork for modern computing.

Key insight: Turing established that machine intelligence should be evaluated by results and observable behavior, not by its ability to replicate human biology or consciousness.

Arthur Samuel and Machine Learning (1952)

In 1952, Arthur Samuel at IBM developed a checkers-playing program that marked a watershed moment in computing history. Unlike previous systems, such as Turing's "*Colossus*" that relied on rigid, hand-coded rules, Samuel's program learned from experience. The system played thousands of games against itself, analyzed the outcomes, and adjusted its strategy to improve over time.

Samuel coined the term "machine learning" to describe this approach: systems that improve performance through exposure to data rather than just through explicit programming for every scenario. These early iterations of machine learning were either *supervised learning*, where the machine is shown patterns and categories, or *unsupervised learning,* where the machine must determine patterns and categories on its own.

Key insight: Arthur Samuel demonstrated that machines could learn patterns and improve autonomously, evolving from impressive machines that followed static rules to early AI systems that could adapt dynamically.

The Birth of "Artificial Intelligence" (1956)

In the summer of 1956, John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon organized the Dartmouth Conference, a gathering of researchers interested in "thinking machines". McCarthy proposed the term "Artificial Intelligence" in the funding proposal to the Rockefeller Foundation, which is the first documented use of that phrase, which would define the field.

The conference brought together pioneers who believed substantial progress could be made if experts collaborated intensively on the problem of machine intelligence. McCarthy later created LISP, a programming language that became foundational for AI research. Minsky built early learning machines and co-founded the MIT AI Lab, exploring how intelligence could be decomposed into processes that machines might replicate.

Key insight: The Dartmouth Conference formalized AI as a distinct research discipline and established a community committed to systematically pursuing machine intelligence.

ELIZA: The First Chatbot (1964–1967)

Between 1964 and 1967, MIT researcher Joseph Weizenbaum created ELIZA, an early natural language processing program designed to simulate conversation with a therapist. ELIZA used simple pattern-matching and substitution to identify keywords in user input and reflected them back using scripted templates, creating the illusion of understanding without genuine comprehension.

Despite its simplicity, ELIZA was remarkably convincing. Users frequently became emotionally engaged, sometimes forgetting they were interacting with a computer. Weizenbaum himself was disturbed by how easily people attributed human-like understanding to the program, raising early ethical questions about AI and human perception.

Key insight: ELIZA demonstrated that even rudimentary AI could create powerful illusions of intelligence, highlighting the importance of understanding what AI systems truly do versus what they appear to do.

AI Winter (1970s)

Early AI systems from the 1950s and 1960s were predominantly rule-based: developers manually encoded logic (if X, then do Y) to guide behavior. These systems were clever and useful for specific tasks, but couldn't adapt to situations outside their programmed rules.

The years of optimistic claims about machine learning and neural networks did show promise, but researchers encountered major limitations in computing power and data handling, leading to a crisis of confidence among funders such as DARPA and the UK government.

This decline was heavily influenced by the 1973 Lighthill Report, which famously criticized the "combinatorial explosion" of AI algorithms and their inability to tackle complex, real-world problems beyond "toy" demonstrations. This report led to a massive funding retraction, causing many researchers to abandon the field of AI or shift their focus.

Key insight: The promise of any technology is always ultimately met by the reality of its true, real-world capabilities. The first AI Winter shows how hype and hope overshadowed the reality of what AI could deliver at the time.

The Transition to Modern AI: 1980s–1990s

Neural Networks and Connectionism (1982)

In 1982, physicist John Hopfield popularized a type of neural network that treated AI systems as interconnected webs of simple units (called 'nodes') that could settle into stable patterns, similar to how memories might form in biological brains. Hopfield networks, as they are known, demonstrated that [artificial] intelligence can emerge from many simple components working together rather than from complex centralized processing.

Neural networks consist of layers of nodes (artificial neurons), each performing simple calculations using weights (learned parameters) and activations (outputs). Information flows forward through the network, with each layer extracting information from a previous layer to build upon. The layers of nodes work together collectively to solve the problem and generate the next best prediction.

Key insight: Neural networks introduced a biological metaphor for AI, showing that intelligence could emerge from distributed, parallel processing rather than sequential rule-following.

The Rise of more advanced Machine Learning (1990s)

By the 1990s, machine learning had matured beyond its early approaches of supervised and unsupervised learning. However, each remained suited to solving different types of problems:

Supervised Learning
Models learn from labeled training examples (identified input + correct output). For example, when given many examples of emails marked "spam" or "not spam," a supervised system learns to classify new emails. Other applications include medical diagnosis systems, image classification, and voice recognition.

Unsupervised Learning
Models discover patterns in unlabeled data. These systems cluster similar items, compress data efficiently, or detect anomalies. For example, businesses use unsupervised learning to segment customers or identify unusual transactions that might be fraud.

Reinforcement Learning
Models learn through trial and error while navigating an environment autonomously. They receive rewards for successful actions and penalties for failures. For example, TD-Gammon (early 1990s), a backgammon program, learned to play at an expert level by playing millions of games against itself and adjusting strategy based on wins and losses. This was a significant leap from the checkers-playing program of the 1950s, due to the increased complexity and strategy inherent in backgammon compared to checkers.

Key insight: Reinforcement learning allowed AI to handle far more complex and varied tasks using ever-greater neural networks.

Geoffrey Hinton and Backpropagation

Geoffrey Hinton, David Rumelhart, and Ronald Williams published groundbreaking work on backpropagation, a mathematical technique for efficiently training neural networks. Backpropagation allows a network to adjust its internal weights systematically, "learning" which patterns lead to correct predictions by working backward from errors.

This technique became the backbone of deep learning. Neural networks were now constructed with significantly more layers that could learn hierarchical representations as well as use backpropagation to identify and work backwards from errors before continuing forward in the neural net.

The ability to move both forward and backward through the neural network was a critical, fundamental evolution for AI. Hinton's work in the 2000s and 2010s on these deeper neural networks led to breakthroughs in image recognition and natural language processing (NLP).

Key insight: Backpropagation made it practical to train large, complex neural networks, enabling the deep learning revolution that powers today's AI.

The Modern Era of AI: 2010s to Present

Deep Learning Revolution

In the 2010s, deep learning achieved by neural networks with many layers trained on massive datasets led to results that stunned the research community. Systems surpassed human performance on image classification benchmarks, translated languages with unprecedented fluency, and recognized speech with near-human accuracy.

Deep learning's success relied on three factors converging simultaneously:

- **Advances in AI Algorithms:** Improved training techniques like backpropagation, dropout, and batch normalization
- **Massive datasets:** The internet provided billions of labeled images, text documents, and videos to train AI
- **Computational power:** GPUs and cloud computing made it feasible (from both a time and cost standpoint) to train networks with billions of parameters

The Transformer Architecture (2017)

In 2017, Ashish Vaswani and colleagues at Google introduced the Transformer architecture in the paper "*Attention Is All You Need*". Transformers revolutionized AI by replacing sequential processing with a mechanism called self-attention. Self-attention enables models to process multiple parts of a sequence in parallel rather than word-by-word.

How Transformers (and self-attention) Work:

1. Text is broken into **tokens** (words or sub-word units)
2. Each token is converted into a numerical representation (embedding)
3. The self-attention mechanism allows each token to "attend to" (focus on) other relevant tokens in the sequence
4. Multiple layers of attention and processing extract increasingly complex patterns
5. The model predicts the next token based on learned patterns from training data

This parallel processing makes Transformers vastly more efficient to train than previous sequential models (like RNNs and LSTMs). Crucially, Transformers scale effectively: as you add more data, more layers, and more parameters, performance continues to improve.

“Transformer” is the "T" in GPT: Generative Pre-trained Transformer. Models like GPT-4, Claude, BERT, and countless others rely on this architecture.

Key insight: The Transformer's attention mechanism powers modern conversational AI, code generation, and multimodal systems. *However, transformers are still fundamentally the same prediction engines as early AI models, just capable of greater speed and scale.*

AI as a Multiplier of Human Ingenuity

Despite dramatic improvements in capability, the fundamental approach remains the same: AI systems predict the next best output based on patterns learned from training data and direction provided through prompts. Modern systems do this at enormous scale and speed, processing billions of parameters and terabytes of text, images, and code in just seconds. This scale can make predictions look like understanding or reasoning (much like ELIZA made patients feel empathy), but the underlying mechanism is still pattern recognition and statistical prediction. *AI does not and cannot reason, feel, or make a moral judgment as any human can.*

The story of AI history shows that the value of AI systems is directly a result of what the human user brings to the machine. Sharp, intentional, and smartly structured interactions with AI can augment human ideas and work. However, humans cannot, and should not, expect AI to divine breakthrough ideas on its own, nor should AI be expected to make judgment calls as a human can.

AI is a collaborator and an accelerator. Together with the true magic of human imagination and ingenuity, AI is indeed capable of helping create the next great and truly world-changing idea.

- **AI excels at pattern recognition:** identifying trends, generating fluent text, classifying images, and translating languages
- **AI struggles with true understanding:** common sense reasoning, causation, and context outside of training data are critical weaknesses that must be anticipated
- **AI amplifies human capability:** it accelerates research, automates repetitive tasks, and provides instant access to information
- **AI depends on human guidance:** the quality of prompts, ethical oversight, and critical evaluation determine outcomes. AI has no autonomy to act without an explicit command

References

[1] Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433-460. https://academic.oup.com/mind/article-abstract/LIX/236/433/986238

[2] Wikipedia contributors. (2026). Computing Machinery and Intelligence. *Wikipedia*. https://en.wikipedia.org/wiki/Computing_Machinery_and_Intelligence

[3] Wikipedia contributors. (2026). ELIZA. *Wikipedia*. https://en.wikipedia.org/wiki/ELIZA

[4] Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36-45.

[5] Gaber, M. M. (2023). The roots of AI: Checkers (1952). *Substack*. https://mohamedmedhatgaber.substack.com/p/the-roots-of-ai-checkers-1952

[6] Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3), 210-229.

[7] Computer History Museum. (2019). The 1956 Dartmouth Workshop and its immediate consequences. https://computerhistory.org/events/1956-dartmouth-workshop-its-immediate/

[8] History of Data Science. (2021). Dartmouth Summer Research Project: The birth of artificial intelligence. https://www.historyofdatascience.com/dartmouth-summer-research-project-the-birth-of-artificial-intelligence/

[9] Wikipedia contributors. (2026). ELIZA. *Wikipedia*. https://en.wikipedia.org/wiki/ELIZA

[10] Weizenbaum, J. (1976). *Computer Power and Human Reason: From Judgment to Calculation*. W. H. Freeman and Company.

[11] Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8), 2554-2558.

[12] Tesauro, G. (1995). Temporal difference learning and TD-Gammon. *Communications of the ACM*, 38(3), 58-68.

[13] Radical VC. (2023). Geoffrey Hinton on the algorithm powering modern AI. https://radical.vc/geoffrey-hinton-on-the-algorithm-powering-modern-ai/

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

[15] Singh, H. (2025). Review of "Attention is all you need" (Vaswani et al., 2017). *WordPress*. https://harbisingh.wordpress.com/2025/08/12/review-of-attention-is-all-you-need-vaswani-et-al-2017/

[16] an AI winter is a period of reduced funding and interest in AI research.https://en.wikipedia.org/wiki/AI_winter