Education About AI Is Fundamental to the Ethics Debate

AI is, without question, a transformative technology. A 2026 Google/Ipsos survey across 21 countries found that 66% of respondents had used an AI tool or application during the previous year. Despite this extraordinary level of adoption, it would appear that understanding of how the technology actually works remains far more limited.

There is, however, considerable debate concerning the ethical implications of artificial intelligence and the appropriate direction and limits of AI research.

This book does not attempt to resolve those questions.

It does, however, take a position on how such questions should be considered: important decisions should be based on knowledge rather than hearsay, rumor, or unfounded speculation. One need not be an AI researcher, computer scientist, or mathematician to participate meaningfully in the debate.

Two observations developed throughout this book are particularly relevant. First, AI systems are computational systems. Their operation consists entirely of computation; no additional non-computational mechanism is required to explain how they produce their outputs. Second, the essential ideas underlying generative AI can be understood without mastering the mathematics used to describe or implement them.

The mathematical material in this book is therefore optional. It is included for readers who want a deeper understanding of encoding, learning, representation, and inference, or who wish to examine the implementation in greater technical detail. It is not necessary, however, for understanding the book's central explanation

of how a generative AI system works.

For readers who do wish to engage with the mathematics, the material has deliberately been kept accessible. Most of it should be within reach of a college student with basic mathematical training and some familiarity with probability. Linear algebra is helpful but not essential. The mathematics adds depth and precision, but it is not a prerequisite for understanding the fundamental ideas developed in the book.

Modern AI systems are extraordinarily large and complex. But scale and implementation complexity should not be confused with complexity of the underlying principles. A system containing billions or trillions of parameters may nevertheless operate according to abstract computational and mathematical principles that are comparatively simple and understandable.

The importance of understanding underlying principles is not merely an academic concern. It is the author's personal view that a significant contributing factor to the financial crises of 1987 and 2008 was market practitioners' failure to appreciate the importance of independence among random variables when applying a relatively simple principle of probability known as the Central Limit Theorem.[1] This led to the underestimation of extreme downside risk. Both crises illustrate the consequences of the widespread use of sophisticated models without fully understanding their underlying assumptions.

The lesson is directly relevant to artificial intelligence. As increasingly sophisticated AI systems are developed and deployed, understanding the fundamental principles governing their operation becomes more, rather than less, important. Education is an essential component of using the technology responsibly, evaluating its limitations, recognizing its risks, and making informed decisions about its development and governance.

[1] In 1987, portfolio insurance was blamed for the crash. However, it replicated the payoff of a put option through dynamic trading, using models that assumed normally distributed returns. That assumption rested on the Central Limit Theorem, and in particular on the independence of returns, a condition that financial markets did not necessarily satisfy. The true underlying distribution of returns was always heavy-tailed (non-normal); what changed when the market moved sharply was the market's perception of it, and the replication failed. The disappearance of liquidity was a symptom, not the cause. In 2008, mortgage-backed securities were valued on the assumption that defaults were independent, so those models also relied on the Central Limit Theorem. That financial returns are not normally distributed was well known, but the reliance of many models on the Central Limit Theorem was obscured by model complexity. Other explanations, such as leverage, market structure, and incentives, are commonly emphasized and were also contributing factors. However, the view that a failure to appreciate the assumptions underlying the Central Limit Theorem was a major contributing factor is the author's interpretation, based on professional experience with such models.

The purpose of this book is therefore educational. It develops a complete 9-parameter AI system that can be trained, inspected, modified, and used for inference within a narrow, deliberately bounded domain. It learns a representation of terrain and uses that representation to perform goal-directed tasks. Under the generation-loop characterization developed in this book, which draws on Malach's (2024) analysis of autoregressive generation, TIM qualifies as a generative AI system, yet by parameter count it is many billions of times smaller than the largest contemporary models.

Although TIM is not an LLM, it shares several important functional characteristics with LLMs, including encoding, learning, the use of a learned representation, a generation loop, and goal-directed inference. It therefore provides a useful, greatly simplified system in which to examine principles and issues that also arise in much larger AI systems.

Its small size makes experimentation possible: the reader can observe the entire process, change its components, and see directly how encoding, learning, learned parameters, inference, and generation interact. Much can be learned about the fundamental principles of AI by experimenting with a system small enough to understand in its entirety. The model also provides a simple environment in which broader issues, such as dual use and the unintended reuse of learned representations, can be explored.

Education and understanding does not resolve the ethical questions surrounding AI, but they provide a far better foundation from which to consider them. Questions about safety, regulation, accountability, bias, dual use, employment, and the appropriate limits of AI development are too important to be left to slogans or assumptions about how the technology works.

This book attempts to contribute to that understanding by stripping away much of the scale and complexity of modern AI and replacing it with TIM—a complete generative AI system small enough for its essential processes to be observed, understood, and experimented with directly. By making the underlying ideas accessible without requiring advanced mathematics or specialized knowledge, the hope is that a much wider audience can develop a practical understanding of what AI is, how it works, what it can do, and where its limitations and risks may arise. The broader that understanding becomes, the better equipped society will be to judge both the extraordinary potential and the very real risks of AI, and to make informed decisions about how the technology should be developed, governed, and used.