Podcast

In this Lex Fridman podcast episode, Judea Pearl explains that today's machine-learning systems excel at finding correlations but lack the ability to reason about cause and effect—an ability he argues is essential for achieving true artificial general intelligence. He outlines his "ladder of causation," moving from simple association (seeing) to intervention (doing) and finally to counterfactual reasoning (imagining what would have happened under different circumstances), and shows how this framework—together with Bayesian networks, structural causal models, and his *do-operator*—lets scientists answer not just "what is related to what" but also "what happens if we act" and "why did this outcome occur." Pearl and Fridman discuss the limits of current deep-learning models, the challenge of confounders and hidden variables, the philosophical implications of causality for free will and explanation, and why empowering machines with causal reasoning is key to making them more human-like in understanding and decision-making.

What is Bayesian Network?

A Bayesian network is a probabilistic model that uses a directed acyclic graph to show how variables are related: each node is a variable, arrows indicate dependencies, and each node has a conditional probability table that defines its likelihood given its parent nodes. This structure lets us efficiently compute and update probabilities as new evidence appears and often reflects causal relationships, making it useful for reasoning under uncertainty in fields like medicine, AI, and risk analysis.

Compare Causation with Correlation

Correlation refers to a statistical association between two variables—meaning they tend to rise or fall together—but it does not explain why the relationship exists or whether one variable influences the other. For instance, higher ice cream sales often occur alongside increased cases of sunburn; they are correlated because of a third factor (hot weather) rather than the other.

Causation, by contrast, implies a direct cause-and-effect relationship in which a change in one variable produces a change in the other, such as flipping a switch causing a light to turn on. While every causal link typically shows some correlation, many correlations arise from coincidence, confounding factors, or indirect relationships, so identifying causation requires deeper analysis—such as controlled experiments, interventions, or causal models—to rule out spurious associations.