August 25, 2025

Sharp on Industry, Thin on Science

By: Suvrat Raju

'AI Con' scrutinises the inflated promises of AI and highlights the social risks. Though timely and provocative, some arguments need refinement, and the limited recognition of genuine advances in machine learning yields an imbalanced, occasionally puzzling picture of progress in the field.

The AI Con is a valuable intervention. At a time when it is hard to escape the crescendo of corporate hype about Artificial Intelligence (AI), Emily Bender and Alex Hanna have written a heterodox, thought-provoking book challenging some of the outlandish claims of the industry and pointing out several social pitfalls. However, some of the book's arguments could be sharpened, and the book is marred by a puzzling reluctance to acknowledge some of the genuine advances in the field of machine learning and AI.

When social institutions are under strain, fawning political leaders can team up with unscrupulous corporations and present technology as a universal panacea. The book describes this phenomenon in the United States, where some states have started using statistical algorithms to advise judges on whether those awaiting trial should be released on bail. In some jurisdictions, child welfare services rely on algorithms to identify children 'at risk' in their homes, who could then be separated from their parents and placed in foster care.

These algorithms are opaque and lack accountability. They also imbibe biases that are present in their training data. Given the history of racism in the US, it is hardly surprising that the algorithms are biased against people of colour.

'Artificial intelligence' is an ill-defined term. But the corporations hawking these products slap it on as a marketing label to attract clients. Bender and Hanna are right in branding these models, which claim to predict social behaviour, an "AI con".

Unacknowledged Labour

Most people encounter AI through a different class of applications called 'generative AI', which includes chatbots like ChatGPT. The book dispels the myth that these models are purely technological marvels. Under the hood, they are supported by large amounts of unacknowledged labour.

> Generative AI models owe their existence to the vast amount of digital data that is available on the internet. This includes code that people post on Github [and] high-quality textual data on sites like Wikipedia.

For instance, commercial AI models are trained to provide anodyne responses when asked obviously harmful questions. One way this is achieved is by recruiting human moderators to flag toxic content produced by the model. This is emotionally taxing work that is crucial for producing the polished product that consumers eventually see. But the workers who contribute to it are rarely acknowledged and certainly do not see any significant share of the fortunes that AI companies have accumulated.

Generative AI models acquire their capabilities by studying examples of human output. Image-generating models generate beautiful art after being trained on billions of images produced by humans. However, AI companies do not share any revenue with these creators, even as they seek to supplant them professionally.

Although Bender and Hanna do not discuss it sufficiently, the problem extends beyond copyrighted work. Generative AI models owe their existence to the vast amount of digital data that is available on the internet. This includes code posted on Github, high-quality textual data on sites like Wikipedia, open scientific repositories like the arXiv, and answers to questions provided on community message boards. Much of this content was created by voluntary labour and shared by people in a spirit of openness. Large AI corporations profit from this openly shared data but do not contribute significantly to the creative human ecosystem that underpins their products.

Central Contradiction

There is no doubt that the AI industry is dominated by greedy corporations and rife with unethical social practices. Yet the core paradox is that, despite these serious issues, recent advances in AI still constitute major scientific progress. Bender and Hanna fail to

acknowledge this dichotomy.

In their discussion of AI and science, they explain how the technology is misused by some unscrupulous researchers. But they inexplicably fail to note that the 2024 Nobel Prize in Chemistry was partly given for major advances in the protein-folding problem. This problem stubbornly resisted progress for years before advances were made by researchers using an AI model and drawing on algorithms originally developed for natural language processing. Thousands of scientists now use machine-learning models every day in fields that range from particle physics to biology.

Do Chatbots 'Understand' Language?

The authors reserve their sharpest disdain for Large Language Models (LLMs) like ChatGPT. In 2021, Bender coined the term "stochastic parrots" to describe these models. The word 'parrots' is meant to suggest that these models repeat what they have learnt without 'understanding'.

In this book and elsewhere, Bender argues that LLMs are familiar with the form of language but lack a sense of "meaning." For instance, the word 'cat' has specific grammatical properties, and might appear commonly with other adjectives like 'furry'. But the meaning of 'cat' to a human goes beyond these linguistic properties. A cat is a real-world organism with specific physical traits, and the word can also carry intangible emotional associations. Bender's point is that since models are trained purely on a corpus of text, they can never understand the 'meaning' of the words they produce.

> The ability of large language models to respond meaningfully to long inputs cannot be explained by positing that they have memorised their training data and are repeating parts of it mindlessly.

This critique is not without its merits. It is well known that LLMs are limited in their understanding of the physical dynamics of the real world, including concepts like simultaneity, which humans and, even animals, understand intuitively. Nevertheless, it seems clear that these models have managed to acquire some understanding of the world that they use to respond to queries.

An example from the world of games might help to elucidate this point. A game of chess can be recorded by noting the starting and ending square for every move. A typical game might start with *1. d2d4 d7d5 2. c2c4... .* It is possible to train a model solely on game records in this notation so that, given a sequence of opening moves, it predicts a suitable next move.

To a human, these symbols have meaning—*d2d4* means that the player with white pieces pushed the pawn in front of their Queen twice and *d7d5* means that the player with black pieces advanced their pawn similarly. A model whose vocabulary only comprises pairs of starting and ending squares cannot possibly express concepts like 'pawn' or 'Queen'. Nevertheless, if the model succeeds in playing legitimate moves, one would conclude that it has gained some understanding of chess, even if it is limited and different from that of humans.

This conclusion holds irrespective of whether the model needs billions of games as training data. A billion is a large number, but the number of possible positions is much larger after even a few moves. Therefore a successful model cannot simply memorise all possible combinations of moves and must acquire some effective understanding of the game.

This example is not purely hypothetical. Researchers have successfully performed this experiment, on a small scale with chess and then in more detail with Othello (a version of 'Reversi'). In the latter case, they were able to probe the internal workings of the model and show that it had developed an internal representation of the game purely after training on games recorded in notation.

A similar argument applies to LLMs. Modern models regularly have a vocabulary in excess of 100,000 tokens, where each token is a word or a part of a word. So there are more than a nonillion (one followed by 30 zeros) distinct sequences of six tokens. This vastly exceeds the number of sentences that have ever been written in all languages combined. Therefore, the ability of models to respond meaningfully to long inputs cannot be explained by positing that they have memorised their training data and are repeating parts of it mindlessly. Instead it indicates that they have developed some efficient internal representation of language and the world.

LLMs are trained on an absurdly simple algorithm where they are taught to predict the next token on being given a sequence of input tokens. So Bender and Hanna are right that one might naively have expected these models to only produce plausible-sounding responses that were most often inaccurate or meaningless. The fact that LLMs can do significantly better is one of the most startling discoveries of the past few years.

To be clear: LLMs are still severely limited. They often make mistakes, are less reliable than humans at many tasks, and have surprising deficiencies in their abilities. But it is untenable to trivialise the progress in the field, as Bender and Hanna appear to do.

AI, Capital, and Empire

Their dismissal of the power of AI also leads them to underestimate its social impacts. They insist that "AI is not going to replace your job. But it will make your job a lot shittier". But jobs in at least some sectors of the economy are threatened by AI.

The prowess of AI models in coding poses a serious medium-term threat to some information technology (IT) workers. Skills that they have acquired over many years might become redundant and they might be forced into worse paying and less secure jobs.

> The concentration of AI companies in a small geographical region of the US is a serious concern for the rest of the world.

We have recent historical evidence of the social impacts of such technological transitions. When textile mills in cities like Mumbai and Ahmedabad shut down, about 40 years ago, a relatively privileged class of mill workers were suddenly 'pauperised'. This contributed to the rise of parochial and communal politics. Cities like Bengaluru now face a similar threat. Needless to say, the captains of the IT industry, having made their fortunes on the backs of IT workers, are perfectly happy to make them redundant and move on to their next AI-powered profit-seeking ventures.

Bender and Hanna also overlook the impact of AI on the international order. The concentration of AI companies in a small geographical region of the US is a serious concern for the rest of the world. In recent times, the US has seen a steady erosion of its hegemonic position. This is undoubtedly a positive development for most of the world's people, contrary to US liberal discourse that paints the US empire as somehow benign or superior to straw-man alternatives.

The US government perceives AI as a route to regaining its dominant position. Its AI action plan published in July 2025 asserts that it is a "national security imperative for the United States to achieve and maintain unquestioned and unchallenged global technological dominance". The challenge, especially for those of us who live in India and other developing countries, is to prevent the US from consolidating its stranglehold on the field.

AI is a powerful technology. Its potential could be used for social benefit. Or it could be used to enrich a few giant corporations and prolong the US empire. Ultimately, the choice is ours. We must educate and organise ourselves, and use strong public action to ensure that AI serves the common good.

Suvrat Raju, a theoretical physicist, is a professor at the International Centre for Theoretical Sciences, Tata Institute of Fundamental Research, Bengaluru. The views expressed are personal.