The Coming Collapse of the Machine Learning Economy

Supervised Machine Learning requires labeled training data. For example, the label might be a class or value for a particular training example and the prediction task might be to predict that class or value for an unseen example.

Enjoy the Golden Era of Machine Learning While It Lasts. Up until now, the machine learning economy has been based on free or low-cost training data with little or no restrictions on its use. This is what the large language models such as those of ChatGPT have been built on. But this golden era of machine learning is not going to last. Its demise will be based on increasing costs of labeled training data and increasing legal restrictions.

Increasing Costs of Labeled Training Data. When the owners of labeled training data realize that their labels have value, they will raise prices commensurate with the profit that can be made from predictions based on that data. When the cost of that labeled training data becomes high enough to shrink the profit margin for companies making use of that data, such companies will collapse.

Legal Restrictions on Labeled Training Data. The owner of data, including images, music, and writing is clearly covered by copyright laws, corporate contracts, and other usage contracts. For example, a musician can't just perform or play most modern music without paying a royalty. Even Paul McCartney has to pay royalties when he performs a song he himself co-wrote as a Beatle, but no longer owns. Worse still, even a sample that is a fraction of a second in length can infringe on a copyright.

Derivatives. But who owns the rights to a predictive system built on data that is covered by such laws? For example, it might be possible for a machine learning system, trained on the entire Beatles catalog, to generate a new song in the style of the Beatles. Should the owner of the Beatles' songs get royalties in that case?

Does Anyone Even Remember Buffy Saint Marie? Suppose a student of music with nothing better to do studies Beatles songs intensely for 10,000 hours to become an expert and then generates new songs in the style of the Beatles? Should the owner of the Beatles' songs get royalties in that case? What if it can be proved that the student of music was actually influenced more by Canadian singer Buffy Saint Marie?

What if the aforementioned machine learning system was also trained on the songs of Tiny Tim, Loretta Lynn, and the Everly Brothers?

Who Cares About Jimmy Stewart Anymore? What about items that are in the public domain, such as an actor's voice. A few decades ago, Jimmy Stewart voice-imitators were popular on the talk shows. They did not get sued by the estate of Jimmy Stewart for every such imitation. Should an ML system that is trained on Jimmy Stewart's voice from the public domain pay royalties for imitating Jimmy Stewart's voice?

The Legal Problem. Defining style "theft' with legal precision is going to be difficult. Mathematically, it might be possible to say that a musical style is represented by a particular set of 100,000 parameters in a neural net. Such a definition might be extremely difficult for a non-technical judge, jury, and attorneys to understand. Summaries of that many parameters might prove to be useless. Worse still, proving what data such a system was and wasn't trained on could be daunting: tracing a particular set of parameters to a particular data set is nearly impossible.

Fear vs. Understanding of Technology. How is a legal system that doesn't really understand Artificial Intelligence and Machine Learning likely to respond? Based on the current fear-mongering of AI in politics, it seems likely that overly restrictive laws will soon be passed by legislators who fear more than understand the limits of current Artificial Intelligence and Machine Learning. For example, laws might be passed to legally limit predictive systems built from unauthorized data, even if that data turned out to be useless for prediction.

Don't Forget about Hemingway! In contrast, it might still be possible for a human to copy a style with impunity because such fear-mongering does not exist in the human arena. For example, Hemingway never had to pay royalties to Sherwood Anderson, though he adopted many of his stylistic characteristics. Would a machine be required to pay such royalties if it were trained on the very same writings of Sherwood Anderson?

Twaining Data. But wait, you say, Hemingway was also influenced by Mark Twain, The Bible, and various wars! Except for the first, the latter two might be completely open-source and unrestricted for any system--whether human or machine--to freely access. If you throw Mark Twain into the mix for training, could you prove which part of your system was Twained and which part was not?

Get Thee to a Punnery. Bad puns aside, the next decade of Artificial Intelligence and Machine Learning might see little in terms of technological progress but a notoriously large amount in terms of potentially unnecessary legal restriction that could squelch further technological progress.