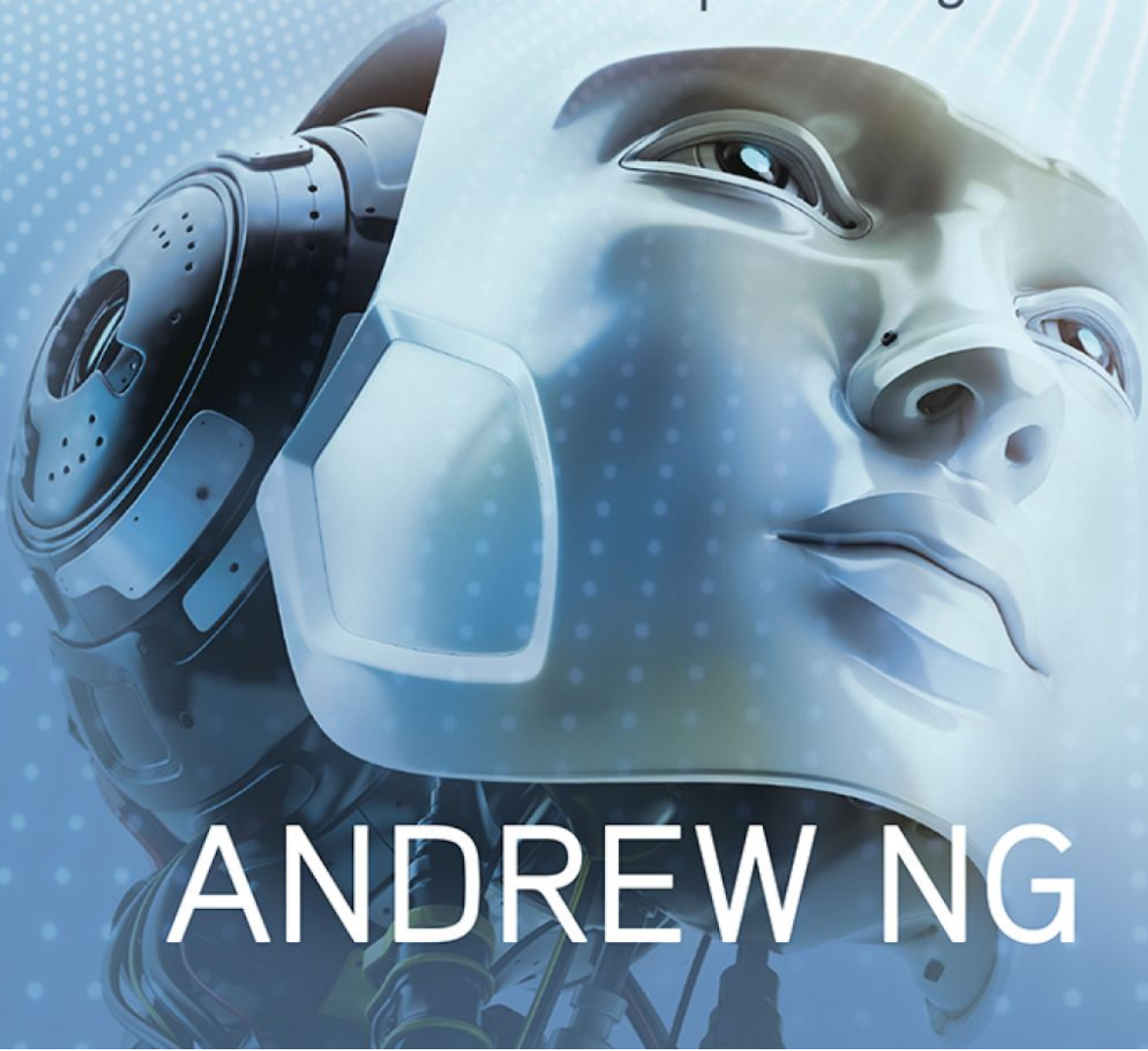
Draft Version

MACHINE LEARNING YEARNING

Technical Strategy for AI Engineers, In the Era of Deep Learning



ANDREW NG

deeplearning.ai

Machine Learning Yearning is a
deeplearning.ai project.

Comparing to human-level performance

33 Why we compare to human-level performance

Many machine learning systems aim to automate things that humans do well. Examples include image recognition, speech recognition, and email spam classification. Learning algorithms have also improved so much that we are now surpassing human-level performance on more and more of these tasks.

Further, there are several reasons building an ML system is easier if you are trying to do a task that people can do well:

1. Ease of obtaining data from human labelers. For example, since people recognize cat images well, it is straightforward for people to provide high accuracy labels for your learning algorithm.

2. Error analysis can draw on human intuition. Suppose a speech recognition algorithm is doing worse than human-level recognition. Say it incorrectly transcribes an audio clip as "This recipe calls for a *pear* of apples," mistaking "pair" for "pear." You can draw on human intuition and try to understand what information a person uses to get the correct transcription, and use this knowledge to modify the learning algorithm.

3. Use human-level performance to estimate the optimal error rate and also set a "desired error rate." Suppose your algorithm achieves 10% error on a task, but a person achieves 2% error. Then we know that the optimal error rate is 2% or lower and the avoidable bias is at least 8%. Thus, you should try bias-reducing techniques.

Even though item #3 might not sound important, I find that having a reasonable and achievable target error rate helps accelerate a team's progress. Knowing your algorithm has high avoidable bias is incredibly valuable and opens up a menu of options to try.

There are some tasks that even humans aren't good at. For example, picking a book to recommend to you; or picking an ad to show a user on a website; or predicting the stock market. Computers already surpass the performance of most people on these tasks. With these applications, we run into the following problems:

- **It is harder to obtain labels.** For example, it's hard for human labelers to annotate a database of users with the "optimal" book recommendation. If you operate a website or app that sells books, you can obtain data by showing books to users and seeing what they buy. If you do not operate such a site, you need to find more creative ways to get data.

- **Human intuition is harder to count on.** For example, pretty much no one can predict the stock market. So if our stock prediction algorithm does no better than random guessing, it is hard to figure out how to improve it.
- **It is hard to know what the optimal error rate and reasonable desired error rate is.** Suppose you already have a book recommendation system that is doing quite well. How do you know how much more it can improve without a human baseline?

34 How to define human-level performance

Suppose you are working on a medical imaging application that automatically makes diagnoses from x-ray images. A typical person with no previous medical background besides some basic training achieves 15% error on this task. A junior doctor achieves 10% error. An experienced doctor achieves 5% error. And a small team of doctors that discuss and debate each image achieves 2% error. Which one of these error rates defines "human-level performance"?

In this case, I would use 2% as the human-level performance proxy for our optimal error rate. You can also set 2% as the desired performance level because all three reasons from the previous chapter for comparing to human-level performance apply:

- **Ease of obtaining labeled data from human labelers.** You can get a team of doctors to provide labels to you with a 2% error rate.
- **Error analysis can draw on human intuition.** By discussing images with a team of doctors, you can draw on their intuitions.
- **Use human-level performance to estimate the optimal error rate and also set achievable "desired error rate."** It is reasonable to use 2% error as our estimate of the optimal error rate. The optimal error rate could be even lower than 2%, but it cannot be higher, since it is possible for a team of doctors to achieve 2% error. In contrast, it is not reasonable to use 5% or 10% as an estimate of the optimal error rate, since we know these estimates are necessarily too high.

When it comes to obtaining labeled data, you might not want to discuss every image with an entire team of doctors since their time is expensive. Perhaps you can have a single junior doctor label the vast majority of cases and bring only the harder cases to more experienced doctors or to the team of doctors.

If your system is currently at 40% error, then it doesn't matter much whether you use a junior doctor (10% error) or an experienced doctor (5% error) to label your data and provide intuitions. But if your system is already at 10% error, then defining the human-level reference as 2% gives you better tools to keep improving your system.

35 Surpassing human-level performance

You are working on speech recognition and have a dataset of audio clips. Suppose your dataset has many noisy audio clips so that even humans have 10% error. Suppose your system already achieves 8% error. Can you use any of the three techniques described in Chapter 33 to continue making rapid progress?

If you can identify a subset of data in which humans significantly surpass your system, then you can still use those techniques to drive rapid progress. For example, suppose your system is much better than people at recognizing speech in noisy audio, but humans are still better at transcribing very rapidly spoken speech.

For the subset of data with rapidly spoken speech:

1. You can still obtain transcripts from humans that are higher quality than your algorithm's output.
2. You can draw on human intuition to understand why they correctly heard a rapidly spoken utterance when your system didn't.
3. You can use human-level performance on rapidly spoken speech as a desired performance target.

More generally, so long as there are dev set examples where humans are right and your algorithm is wrong, then many of the techniques described earlier will apply. This is true even if, averaged over the entire dev/test set, your performance is already surpassing human-level performance.

There are many important machine learning applications where machines surpass human level performance. For example, machines are better at predicting movie ratings, how long it takes for a delivery car to drive somewhere, or whether to approve loan applications. Only a subset of techniques apply once humans have a hard time identifying examples that the algorithm is clearly getting wrong. Consequently, progress is usually slower on problems where machines already surpass human-level performance, while progress is faster when machines are still trying to catch up to humans.