Chapter 6: A Closer Look at Machines That Learn

1. TRUE/FALSE - The learning-from-data approach of deep neural networks has generally proved to be more successful than the "good old-fashioned AI" strategy, in which human programmers construct explicit rules for intelligent behavior. However, contrary to what some media have reported, the learning process of ConvNets is not very humanlike.
2. Why does your professor like the previous question?

Despite deep neural networks boasting success amongst various media outlets, it is difficult to say that those machines are true "artificial intelligence," as it does not simulate the human phenomenon of intelligence. Professor Graci likes that distinction notably as it supports the idea that symbolic AI is true "artificial intelligence" and thus is more successful than subsymbolic approaches in that sense.

3. TRUE/FALSE - As we've seen, the most successful ConvNets learn via a *supervised-learning* procedure: they gradually change their weights as they process the examples in the training set again and again, over many epochs (that is, many passes through the training set), learning to classify each input as one of a fixed set of possible output categories.
4. List some significant differences between the way that humans learn about objects and the way that ConvNets learn about objects.

ConvNets rely on training against data training sets and adjusting weights to "learn" about objects. Knowledge only consists of possible categories. Humans are active learners who ask questions and explore. Unlike ConvNets, humans do not require massive training sets to categorize objects (97).

5. Why is it inaccurate to say that today's successfully ConvNets "learn on their own?"

ConvNets do not "learn on their own" since humans maintain control over their configuration ("tuning the hyperparameters") and create their training sets using human-made data (97).

6. In answer to the rhetorical question "Where does all of the data come from to fuel big data applications?" MM answers "You - and probably everyone you know." Please elaborate on the answer.

As a rule of thumb in the Digital Age, if any software is advertised as free, the product is its users. The Terms of Service for these applications that oftentimes go unread by the general public usually include clauses pertaining to mass data collection of everything about its users for engagement and advertising purposes. Even by not signing up to a platform, your data can end up on there by your peers – perhaps, by a friend who posted a photo including you on FaceBook.

7. How do car companies acquire the big data (labeled images of pedestrians, cyclists and other obstacles) needed to train robo-cars?

"Self-driving car companies collect these training examples from countless hours of video taken by cameras mounted on actual cars driving in traffic on highways and city streets. These cars may be self-driving prototypes being tested by companies or, in the case of Tesla, cars driven by customers who, upon purchase of a Tesla vehicle, must agree to a data sharing policy with the company," (100).

8. What is the "long tail" phenomenon, and how does it relate to machines that learn (ConvNets)?

The "long tail" phenomenon accounts for the reason why it is so difficult to train AI systems for self-driving cars. Essentially, there are so many possibilities for what can happen on the road – some common, many uncommon – that graphically the possibilities look like a long tail due to the low probabilities of uncommon events on the graph (101). Mitchell writes, "Most real-world domains for AI exhibit this kind of long-tail phenomenon: events in the real world are usually predictable, but there remains a long tail of low-probability, unexpected occurrences," (102).

9. TRUE/FALSE - A commonly proposed solution to the long tail problem in AI systems is to complement supervised learning with unsupervised learning.
10. What is "unsupervised learning?"

"The term *unsupervised learning* refers to a broad group of methods for learning categories or actions without labeled data," (103).

11. What colorful remark did Yann LeCun make about unsupervised learning?

Yann Lecun said, "'unsupervised learning is the dark matter of AI,'" (103).

12. TRUE/FALSE - For general AI, almost all learning will have to be unsupervised, but no one has yet come up with the kinds of algorithms needed to perform unsupervised learning.

13. TRUE/FALSE - Humans have a fundamental competence lacking in current AI systems: common sense. We have vast background knowledge of the world, both its physical and social aspects. We have a good sense of how objects - both animate and living - are likely to behave, and we use this knowledge extensively in making decisions about how to act in any given situation.
14. TRUE/FALSE - Many people believe that until AI systems have common sense as humans do, we won't be able to trust them to be fully autonomous in complex real-world situations.
15. TRUE/FALSE - Superficial changes to images, such as slightly blurring or speckling an image, changing some colors, or rotating objects in the scene, can cause ConvNets to make significant errors even when these perturbations don't affect humans' recognition of objects. This unexpected fragility of ConvNets - even those that have been said to "surpass humans at object recognition" - indicates that they are overfitting to their training data and learning something different from what we are trying to teach them, a phenomenon that results in various manifestations of unreliability.
16. The unreliability of ConvNets can result in embarrassing - and potentially damaging - errors. Select a particularly embarrassing/damaging example of unreliability in ConvNets, and describe it in just a sentence or two.

    Google's image tagging system labeled a photo of two African Americans as "Gorillas," (106).

17. At the end of the section on biased AI, MM observes that the problem of bias in applications of AI has been getting a lot of attention recently, with many articles, workshops, and even academic research institutes devoted to this topic. What questions does she raise in conjunction with this observation? What do you think are the appropriate answers to these questions?

Mitchell writes, "Should the data sets being used to train AI accurately mirror our own biased society – as they often do now – or should they be tinkered with specifically to achieve social reform aims? And who should be allowed to specify the aims or do the tinkering?" (108). I think that AI should not be in these critical systems to begin with, seeing as it can cause real damage based on existing biases in society. An attempt to create a utopian society using AI as a medium

is not the solution either; I think we are just throwing technology at problems that do not need and, perhaps, are exacerbated by a technological solution.

18. TRUE/FALSE - You can often trust that people know what they are doing if they can explain to you *how* they arrived at an answer or a decision. However, "showing their work" is something that deep neural networks - the bedrock of AI systems - cannot easily do.
19. TRUE/FALSE - Recall that a convolutional neural network decides what object is contained in an input image by performing a sequence of mathematical operations (convolutions) propagated through many layers. For a reasonably sized network, these can amount to billions of arithmetic operations. While it would be easy to program the computer to print out a list of all the additions and multiplications performed by a network for a given input, such a list would give us humans *zero* insight into how the network arrived at its answer. A list of a billion operations is not an explanation that a human can understand.
20. What, according to MIT's *Technology Review* is **the dark secret at the heart of AI?**

Mitchell writes, "Even the humans who train deep networks generally cannot look under the hood and provide explanations for the decisions their networks make. MIT's *Technology Review* magazine called this impenetrability 'the dark secret at the heart of AI,'" (109.

21. What does the phrase "theory of mind" refer to, and how is it related to our interactions with AI systems such as deep networks?

Theory of mind encompasses the idea that humans understand the general way humans think and interact based on their experiences and knowledge. Mitchell writes that theory of mind is "a model of the other person's knowledge and goals in particular situations," (109). With AI, we do not have this understanding of its "thinking" works, which makes them untrustworthy.

22. One of the hottest new areas of AI is variously called "explainable AI," "transparent AI," or "interpretable machine learning." To what do these terms refer?

These terms refer to a field of AI that is trying to create deep learning systems that explain its behavior in a way that humans can decipher.

23. The field of "adversarial learning" has emerged in response to the fact that AI systems can readily be fooled in dramatic fashion, like mixing up a guy in glasses with Milla

Jovovich, or misclassifying a stop sign for a speed-limit sign. Briefly describe the field of **adversarial learning**.

Adversarial learning is a field that is dedicated to taking protective measures against malicious attacks on machine-learning systems by studying weaknesses in the machines and providing solutions for the aforementioned threats (112-113).

24. Jeff Clune, an AI researcher at the University of Wyoming, made a very provocative analogy when he noted that there is "a lot of interest in whether Deep Learning is 'real intelligence' or a 'Clever Hans.'" Explain the essential question that underlies this analogy, being sure to incorporate a few words on the actual Clever Hans.

We do not know whether or not deep learning is another Clever Hans situation. Since deep learning gives no insight into how or why it behaves the way it does, machines could be responding to confounding cues in the data to give its correct answer, as opposed to giving correct answers based on the proper path of thinking and understanding. In the Clever Hans story, the horse was responding to the subconscious cues of the person asking the math questions to get the answer right for the reward. Hans was only clever in the sense that he understood how to trick the trainer and lacked mathematical understanding entirely. The same predicament could be happening with companies training artificial intelligence using deep learning.