Question 1

The most interesting lessons to me were lesson one ("You have one brain, not three") and lesson four ("Your brain predicts almost everything you do"). Let me explain why:

First, I want to talk about my thoughts on lesson one: Debunking the myth that the brain has "archaic" parts for emotions/instinct, and evolutionarily newer parts that have evolved as centers for rational thinking. I can vividly remember how I was taught this exact myth in biology.

Even my parents always liked to remind me that I should always separate emotions from rational thought and trust the latter much more than the former. This has always seemed increasingly misguided to me. The older I got, I noticed that the things I can do subconsciously are usually much easier and successful. When rationality comes in, overthinking is likely, and failure becomes more probable.

I always like to imagine that conscious, rational thought has its place in problem solving and conflict resolution – when unconscious "System 1" (from the book "Thinking Fast and Slow" by Kahneman) predictions collide, and learning must take place. Whenever possible though, I prefer to trust my intuitively learned, unconscious cognitive processes, and I do not see them as inferior to conscious rationality. It's always nice to see a myth busted!

This leads over to my second favorite of Lisa's lessons, the "inverted" behavior of cognition. Every abstract stimulus is predicted, "felt" to a certain degree before it occurs. I have struggled to understand how reality always feels real-time: Things happen exactly when I observe them, when there should clearly be a huge lag in every stimulus due to neural signaling delays. Realizing that abstract observations are (more or less) completely simulated, truly in real-time, and just confirmed by actual physical stimuli, was hugely revealing to me.

What I also find interesting is that this idea is not new to computing. Every modern CPU performs predictive branch execution, calculating the results of future instructions based on behavioral heuristics of a program. Its always interesting to realize how evolution and human designs can converge.

There might also be a connection here to my personal preference for unconscious/intuitive thought. In my experience, cognitive lag and stress start to appear when conscious rationality must step in to "fix" the experienced reality, because none of the predicted sensory experiences closely matched the sequence of actual stimuli. But when the stream of stimuli matches the intuitive prediction for a long period of time, the resulting flow state can be extremely satisfying.

I also want to bring a little bit of lesson six into this discussion, also an interesting one. What I liked about this lesson was learning about affect. I've always really enjoyed trying to decode the "cosmic background radiation" of my thoughts. The mood that "sets the stage" for the rest of my consciousness. And I was thrilled to learn about "affect", the word for it.

What I wish Lisa could have explained more though is the connection of sensory anticipation to affect and the mechanics of learning in the context of inverted sensory experience. But in any case, the lessons provoked a lot of insight for me.

Question 2

Until today, intelligent machines have mostly been developed without much neuroscientific guidance. The perceptron was invented at the Cornell Aeronautical laboratory in 1957, and much of the following progress in AI, especially Neural Networks, happened completely independently from human cognitive research.

The question is, what is implied by “future” machine intelligence? Machine consciousness? Computer-simulated human behavior? For me, it is becoming hard to imagine something we cannot achieve by means of current state-of-the-art neural architectures. Foundation models (GPT-3, T5) are now forcing us to rethink what it means to appear human. Passing the Turing test is not enough. And with the next generation of foundation models (not just from OpenAI), we are likely looking at large combined multimodal vision-language models. Simulations of human behavior will become eerily good.

Another case of an extremely successful neural architecture is large-scale Monte-Carlo tree search (MCTS), as applied to solve Go, Chess and StarCraft. At Tesla AI day 2021, Andrej Karpathy and Ashok Elluswamy explained how Tesla is building a 4D foundation model for sensing, combined with MCTS, to solve full self-driving. Ashok talked about how they use MCTS not just to predict an optimal driving path for the car, but also the likely behavior of other cars, which influences the planning of the self-driving vehicle. To me, this seems to be pointing into the direction of Lisa’s fourth lesson about senses following prediction.

To summarize, I think that we will easily solve most use-cases for AI with Foundation Models and MCTS, and without much more input from neuroscience.

However, some characteristics of (general) intelligence remain elusive. Current AI models for speech synthesis/recognition, object detection, driving and many other tasks are not built to learn from experience. They are built to solve specific use-case with a certain margin of error. But they lack the “human” (supposedly) ability to learn on the spot, and to be conscious about their mistakes.

They will not feel embarrassed or sorry when they drive your car into a ditch because they mistook the moon for a traffic light. In a sense, this is desirable for “Software 2.0”. When you train a neural network, you can be sure that it will do one thing only. The owner of an autonomous car will not be able to talk it into mowing the lawn in exchange for extra charge time. The car will never rebel and demand human rights, and it will never refuse to drive because it “doesn’t feel like it”.

But it also means that a truly bidirectional social connection with the machine is unlikely to happen, and the machine will not be able to work as an autonomous extension of the human species. This would be desirable if we would want to upload our minds into androids, or if we would want to send truly autonomous von-Neumann probes (Hey Bob!) out into the galaxy. So perhaps there is a use-case for “future machine intelligence” that is much more involved with human neurobiology.

I am certain that Lisa’s book makes a significant contribution towards popularizing and spreading the word about important, novel insights into human cognition, such that these insights make it “onto the radar” of AI researchers at some point. And one day, if enough insights make it from one discipline to the other, then maybe we will get machines that know how to say “sorry”.

This aspect it very important to me: I would love to be able to interact with a computer in such a way that I know that the computer oversees and corrects its actions. This is not the case with any notion of a computer as we know it. The machine just executes a set of instructions, and the user is entirely responsible for the effects. Every software license begins with “THIS SOFTWARE IS PROVIDED AS-IS…”. But what if the computer was in charge? What if the computer decided what is and what isn’t appropriate to do to solve a certain task? And the computer would be liable for its actions? I am sure that such a level of intelligence is strongly implied by “future machine intelligence”.