PATRICK HENRY WINSTON: Oh, how to start. I have been on the Navy Science Board for a couple of decades. And so as a consequence, I've had an opportunity to spend two weeks for a couple of decades in San Diego. And the first thing I do when I get to San Diego is I go to the zoo, and I look at the orangutans and ask myself, how come I'm out here and they're in there? How come you're not all covered with orange hair instead of hardly any hair at all? Well, my answer to that is that we can tell stories and they can't. So this is the center of my talk. That's what I have come to believe after a long time.

So I'm going to be talking about stories and to give you a preview of where I'm going to go with this, I like to just show you some of the questions that I propose to address in the course of the next hour. I want to start talking about this by way of history. I've been in artificial intelligence almost since the beginning, and so

I was a student of Marvin Minsky, and it's interesting that the field started a long time ago-- 55 years ago-- with perhaps the most important paper in the field, *Steps Toward Artificial Intelligence.* It was about that time that the first intelligent program was written. It was a program that could do calculus problems as well as an MIT freshman-- a very good MIT freshman. That was done in 1961. And those programs led to a half century of enormous progress in the field. All sorts of ideas and subfields like machine learning were spawned. Useful applications.

But as you know, there's been an explosion in these useful applications in recent years. I've stolen this slide from Tomaso but I added a couple of things that I particularly think are of special importance like Echo. You all know about Amazon Echo, of course, right? I swear, it astonishes me. Many people don't know about Amazon Echo.

It's Siri in a beer can. It's a Siri that you can talk to across the room and say how long should I boil a hard boiled egg or I've fallen down. Please call for an ambulance. So it's a wonderful thing. I think most people don't know about it because of privacy concerns. Having something

listening to you all the time in your home may not be the most comfortable kind of thing. But anyhow, it's there.

So this has caused quite a lot of interest in the field lately. Boris has just talked about Siri and Jeopardy, but maybe the thing that's astonished the world the most is this captioning program I think you've probably seen before the last week or two. And man, I don't know what to think of this. I don't know why it's created such a stir, except that that caption makes it look like that system knows a lot that it doesn't actually know.

First of all, it's trained on still photos. So it doesn't know about motion. It doesn't know about play. It doesn't know what it means to be young. It would probably say the same thing if we replaced those faces with geriatrics. Yet when we see that, we presume that it knows a lot. It's sort of parasitic semantics. The intelligence is coming from our interpretation of the sentence, not from it producing the sentence. So yeah, it's a great engineering achievement, but it's not as smart as it looks.

And of course, as you know, there's been a whole industry of fooling papers. Here's a fooling example. One of those is considered by a standard deep neural net to be a school bus, and the other is not considered to be a school bus. And just a few pixels have been changed. Imperceptible to us. It still looks like a school bus but not to the deep neural net. And of course, these things are considered school buses in another fooling paper. Why in the world would those be considered school buses? Presumably because of some sharing of texture.

Well, anyhow, people have gotten all excited and especially Elon Musk has gotten all excited. We are summoning the demon. So Musk said that in an off-the-cuff answer to a question at the, let's see, an anniversary of the MIT AeroAstro Department. Someone asked him if he was interested in AI, and that's his off-the-cuff response.

So it's interesting that those of us who've been around for a while find this beyond interesting. It's curious, because a long time ago philosophers like Hubert Dreyfus were saying that it was not possible. And so now we've shifted from not possible to the scariest thing imaginable. We can't stop ourselves from chuckling. Well, Dreyfus must know that he's wrong, but maybe he's right too. What we really had to wait for for all of these achievements was massive amounts of computing.

So I'd like to go a little further back in history while I'm talking about history, and the next thing back there is 65 years ago Turing's paper on machine intelligence. It's interesting that that

paper is widely assumed to be about the Turing test and it isn't. If you look at the actual paper content, what you find is that only a couple of pages were devoted to the test. Most of it was devoted to discussion of arguments for and against the possibility of artificial intelligence.

I've read that paper 20 times, because I prescribe it in my course, so I have to read it every year. And every time I read it, I become more convinced that what Turing was talking about is arguments against AI, not about the tests. So why the test? Well, he's a mathematician and a philosopher. And no mathematician or philosopher would write a paper without defining their terms. So he squeezed into some kind of definition.

Counterarguments took up a lot of space, but they didn't have to. If Turing had taken Marvin Minsky's course in the last couple of years, he wouldn't have bothered with that test, because Minsky introduced the notion of a suitcase word. That's a word-- he likes that term because what he means by that is that the word is so big, like a big suitcase, you can stuff anything into it.

So for me, this has been a great thought because, if you ask me, is Watson intelligent? Is the Jeopardy-playing system intelligent? My immediate response is, sure. It has a kind of intelligence. As Boris points out, it's not every kind of intelligence, and it doesn't think like we do, and there are some kinds of thinking that it doesn't do that we do quite handily. But it's silly to argue about whether it's intelligent or not. It has aspects, some kinds of intelligence.

So what Turing really did was establish that this is something that serious people can think about and suggests that there's no reason to believe that it won't happen someday. And he centered that whole paper on these kind of arguments, to the arguments against AI. It's fun to talk about those. Each of them deserves attention.

I'll just say a word or two about number four there, Lady Lovelace's objection. She was, as you all know, the sponsor and programmer of Charles Babbage when he attempted to make a computer mechanically. And what she said, she was obviously pestered with the same kind of things that everybody in AI is always pestered with.

And at one point, she was reported to have said-- let me put it in my patois. Don't worry about a thing. They can only do what they're programmed to do. And of course, what she should have said is that they can only do what they've been programmed to do and what we've taught them to do and what they've learned how to do on their own. But maybe that wouldn't have had the soothing effect she was looking for.

In any event, this is when people started thinking about whether computers could think. But it's not the first time people thought about thinking, that we have to go back 2,400 years or so to get to that. And when we do, we think about Plato's most famous work, *The Republic,* which was clearly a metaphor to what goes on in our brains and our minds and our thinking. He couched it in terms of a metaphor with how a state is organized with philosopher kings and merchants and soldiers and stuff.

But he was clearly talking about a kind of theory of brain-mind thinking that suggested there are agents in there that are kind of all working together. But it's important to note, I think, that *The Republic* is a good translation of the Latin de re publica, which is a bad translation of the Greek politeia. And politeia, interestingly, is a Greek word that my Greek friends tell me is untranslatable. But it means something like a society or community or something like that.

And the book was about the mind. So Plato-- it could have been translated as the society of mind, in which case it would have anticipated Marvin Minsky's book with the same title by 2,400 years. Well, maybe that was not the first time that humans thought about thinking, but it was an early landmark.

And now it's, I think, useful to go back to when humans started thinking. And that takes us back about 50,000 years-- not millions of years-- a few tens of thousands of years. So it probably happened in southern Africa. It was probably 60,000 or 70,000 years ago that it started happening. It probably happened in the neck-down population, because if the population is too big, an innovation can't take hold.

It was about that time that people-- us, we-- started drilling holes in sea shells, presumably for making jewelry. And then it wasn't long after that that we departed from those Neanderthal guys in a big way. And we started painting caves like the ones at Lascaux, carving figurines like the one at Brassempouy.

And I think the most important question we can ask in AI is what makes us different from that Neanderthal who couldn't do these things? See, that's not the question that Turing asked. The question Turing asked was how can we make a computer reason? Because as a mathematician, he thought that was a kind of supreme capability of human thinking.

And so for 20, 30 years, AI people focused on reasoning as the center of AI. And what they should have been asking is, what makes us different from the Neanderthals and chimpanzees

and other species? It creates a different research agenda.

Well, I'm much influenced by the paleoanthropologist, Ian Tattersall, who writes extensively about this and says that it didn't evolve, it was more of a discovery than an evolution. Our brains came to be what they are for reasons other than human intelligence. So he thinks of it as a minor change or even a discovery of something we didn't know we had.

In any event, he talks about becoming symbolic. But of course, as a paleoanthropologist and not a computationalist, he doesn't have the vocabulary for talking about that in computational terms. So you have to go to someone like Noam Chomsky to get a more computational perspective on this.

And what Chomsky says is that-- who is also, by the way, a fan of Tattersall-- what Chomsky says is that what happened is that we acquired the ability to take two concepts and put them together and make a third concept and do that without limit. An AI person would say, oh, Chomsky's talking about semantic nets. A linguist would say he's talking about the merge operation. But it's the same thing.

As an aside, I'll tell you that a very important book will come out in January, I think, by Berwick and Chomsky and addresses two questions-- why do we have any language and why do we have more than one? You know, when you think about it, it's weird. Why should we have a language and now that we have one, why should we all have different ones?

And their answer is roughly that this innovation made language possible. But once you've got the competence, it can manifest itself in many engineering solutions. They also talk a lot about how we're different from other species, and they like to talk about the fact that we can think about stuff that isn't there. So we can think about apples even when we're not looking at an apple.

But back to the main line, when Spelke talked to you, she didn't talk to you about what I consider to be the greatest experiment in developmental psychology ever, even though it wasn't necessarily-- well, I confused myself. Let me tell you about the experiment. Spelke doesn't do rats, but other people do rats with the following observation-- take a rectangular room and there are hiding places in all four corners that are identical.

While the rat is watching you put food in one of these places, a box cloth over it, and then you disorient the rat by spinning it around. And you watch what the rat does. And rats are pretty

smart. They do the right thing. Those opposite corners are the right answer, right? Because the room is rectangular, those are the two possible places that the food could be. So then you can repeat this experiment with a small child, you get the same answer, or with a intelligent adult like me, and you get the same answer because it's the right answer.

But now the next thing is you paint one wall blue and repeat the experiment. What do you think the rat does? Both ways. You repeat the experiment with a small child, both ways. Repeat the experiment with me. Finally we get it right. What's the difference? And when does it happen? When does this small child become an adult?

After elaborate and careful experiments of the kind that Spelke is noted for, she has determined that the onset of this capability arises when the child starts using the words left and right in their own descriptions of the world. They understand left and right before that, but this is when they start using those words. That's when it happens.

Now we introduce the notion of verbal shadowing. So I read to you the Declaration of Independence or something else and as I say it, you say it back to me. It's sort of like simultaneous translation, only it's English to English. And now you take an adult human, and even while they're walking into the room, they're doing this verbal shadowing. And what happens in that circumstance? That reduces the adult to the level of a rat. They can't do it.

And you say, well, didn't you see the blue wall? And they'll say, yeah, I saw the blue wall but couldn't use it. So Spelke's interpretation of this is that the words have jammed the processor. That's why we don't use our laptops in class, right, because we only have one language processor, and it can be jammed. It's jammed by email. I'm jammed by used car salesmen talking fast. It's easy to jam it. And when we jam it, it can't do what you would think it could do.

So Spelke has an interpretation to this that says what we humans have is combinators, the ability to take formation of different kinds and put it together. I have a different interpretation, which I'll tell you about at the end if you ask me why I think Spelke has it-- why I have a different interpretation from Spelke for this experiment.

And then what we've got is we've got the ability to build descriptions that seem to be the defining characteristic of human intelligence. We've got it in the case at Lascaux. We've got it in the thoughts of Chomsky. We've got it in these experiments of Spelke. We've got descriptions. And so those are the influences that led me to this thing I call the strong story hypothesis.

And when you think about it, almost all of education is about stories. You know, you start with fairy tales that keep you from running away in the mall. You'll be eaten by big bad wolf if you do. And you end up with all these professional schools that people go to-- law, business, medicine, and even engineering. You might say, well, engineering, that's not really-- is that case studies? And the answer is, if you talk to somebody that knows what they're doing, what they do is very often telling a story.

My friend, Gerry Sussman, a computer scientist whose work I use, is fond of teaching circuit theory as a hobby. And when you hear him talk about this circuit, he talks about a signal coming in from the left and migrating through that capacitor and going into the base of a transistor and causing a voltage drop across the emitter, which creates a current that flows into the collector, and that causes-- he's just basically telling the story of how that signal flows through the network. It's storytelling.

So if you believe that, then these are the steps that were prescribed by Marr and Tomaso as well in the early days of their presence at MIT. These are the things you need to do if you believe that and want to do something about it. And these steps were articulated at the time, in part because people in artificial intelligence were, in Marr's words, too mechanistic. I talked about this on the first day, that people would fall in love with a particular mechanism-- a hammer-- and try to use it for everything instead of understanding what the problem is before you select the tools to bring to bear on producing a solution.

And so being an engineer, one of the steps here that I'm particularly fond of, once you've got the articulated behavior 100%, eventually you have to build something. Because as an engineer, I think I don't really understand it unless I can build it. And then building it, things emerge that I wouldn't have thought about if I hadn't tried to build it.

Well, anyhow, let's see. Step one, characterize the behavior. The behavior has to story understanding. So I'm going to need some stories. And so I tend to work with short summaries of Shakespearean plays, medical cases, cyber warfare, classical social studies, and psychology. And these stories are written by us so as to get through Boris's parser.

So they are carefully prepared. But they're human readable. We're not encoding this stuff up. This is the sort of thing you could read, and if you read that you say, yeah, this is kind of funny, but you can understand it. Summary of Macbeth. Here is a little fragment of it, and it is easier to read.

So what do we want to do? What can we bring to bear on understanding the story? If you read that story, you'd see-- I could ask you a question, is Duncan dead at the end? And how would you know? It doesn't say he's dead at the end. He was murdered, but-- I could ask you is this a story about revenge? The word revenge is never mentioned, and you have to think about it a little bit. But you'd probably conclude in the end that it's about revenge.

So now we ask ourselves what kinds of knowledge is required to know that Duncan is dead at the end and that it's about revenge? Well, first of all, we need some common sense. And what we've found, somewhat to our surprise, is that much of that can be expressed in terms of simple if-then rules.

And these seven rule types arose because people building software to understand stories found that they were necessary. We knew we needed the first kind. If you kill someone, then they are dead. Every other one here arose because we reached an impasse in our construction of our story understanding system.

So the may rules. If I anger you, you may kill me. Thank god we don't always kill people who anger us. But we humans always are searching for explanations. So if you kill me, and I've previously angered you, and you can't think of any other reason for why the killing took place, then the anger is supposed. So that's the explanation for rule type number two.

Sometimes we use abduction. You might have a firm belief that anybody who kills somebody is crazy. That's abduction. You're presuming the antecedent from the presence of a consequent. So those are kinds of rules that work in the background to deal with the story. And of course, there are things that are explicit in the story too.

Here are some examples of things that are-- of causal relations that are explicit in the story. The first kind says that this happens because that happened. A close, tight causal connection. Second kind, we know there's a causal connection, but it might be lengthy.

The third kind, the strangely kind, arose when one of my students was working on Crow creation myths. He happened to be a Crow Indian and so a natural interest in that mythology. And what he noted was that in Crow mythology, you're often told that something is connected causally and also told that you'll never understand it.

Old Man Coyote reached into the lake and pulled up a handful of mud and made the world,

and you will never understand how that happened, is a kind of typical expression in Crow creation mythology. So all of these arose because we were trying to understand particular kinds of stories.

OK. So that's all background. Here it is in operation. And that's about the speed that it goes. That is reading the story-- the summary of Macbeth that you saw on the screen a few moments ago. But of course, it's invisible at that scale. So let me blow a piece of it up. There you see the piece that says, oh, Macbeth murdered Duncan. Duncan becomes dead. So the yellow parts are inserted by background knowledge. The white parts are explicit in the story.

So you may have also noted-- no, you wouldn't have noted-- well, my drawing attention to it. We have not only the concept that the yellow parts-- yes, the yellow parts there are conclusions. The white parts are explicit. And what you can see, incidentally, is that-- just from the colors-- that much of the understanding of the story is inferred from what is there.

It's a funny kind of way of saying it, but you've seen in computer vision or you will see in computer vision that what you think you see is half hallucination, and what you think you see in the story is also half hallucinated. It seems that the authors just tell us enough to keep us on track. In any event, we have not only the yellow parts that are inferred, but we also have the observation that one piece may be connected to another. And that can only be determined by doing a search through that so-called elaboration graph.

So here are the same sorts of things that you can search for. There's a definition of a Pyrrhic victory. You do something or rather you want something at least you becoming happy, but ultimately the same wanting leads to disaster. So there it is. That green thing down there is reflected in the green elements up there that are picked out of the entire graph because they're connected. And I'll show you how they're connected in this one.

So this is the Pyrrhic victory concept that has been extracted from the story by a search program. So we start with Macbeth wanting to be king. He murders Duncan because of that. He becomes happy, because he eventually ends up being king himself, but downstream he's harmed. So it's a kind of Pyrrhic victory. And now you say to me, well, that's not my definition of Pyrrhic victory, and that's OK, because we all have nuanced differences in our concepts. So this is just one computer's idea of what a Pyrrhic victory is.

Here are the kinds of things we've been able to do as a consequence of, to our surprise, just having a suite of rules and a suite of concepts. And what I'm going to spend the next few

minutes doing is just taking you quickly through a few examples of these kinds of things.

This is reading Macbeth from two different cultural points of view, an Asian point of view and a US point of view. There were some fabulous experiments conducted in the '90s in a high school in the outskirts of Beijing and in a high school in Wisconsin. And these experiments involved having the students read stories about violence and observing the reaction of the students to those stories.

And what they found was that at a statistically significant level, not this or that, but a statistically significant level, the Asian students outside of Beijing attributed violence to situations. And they would ask, what made that person want to do that? Whereas the kids in Wisconsin had a greater tendency to say, that person must be completely crazy. So one attributed to the situation, the other was dispositional, to use the technical term.

So here we see in one reading of Macbeth-- let me show it blown up. In the top version, Macduff kills Macbeth because there's a revenge situation that forces it. And the other interpretation is because Macduff is crazy. So another kind of similar pairing-- oh, wow. That was fast. This is a story about the Estonian Russian cyber war of 2007. Could you-- you probably didn't hear the rules of engagement, but I don't talk to the back of laptops. So if you'd put that away, I'd appreciate it.

So in 2007, the Estonians moved a war memorial from the Soviet era out of the center of town to a cemetery in the outskirts. And about 30% of the Estonian population is Russian, and they were irritated by this. And it had never been proven, but the next day the Estonian National Network went down, and government websites were defaced. And this was hurtful, because the Estonians pride themselves in being very technically advanced. In Estonia, it's a right to be educated on how to use the internet. They have a national ID card. They have a law that says if anybody looks at your data, they've got to explain why. They're a very technically sophisticated country.

And so what's the interpretation of this attack, which was presumed to be done by either ethnic Russians in Estonia or by people from Russia? Well, was it an aggressive revenge or was it teaching the Estonians a lesson? It depends on what? It depends on whose side you're on. That's the only difference. And that's what produced the difference in interpretation on those two sides-- one being aggressive revenge and the other being teaching the Estonians a lesson. By the way, I was in Estonia in January. That's the statue that wasn't there.

Give you another example. I'm just trying to show you some of the breadth of our story understanding activity. So the next example comes about because shortly after the terrorist attack on the World Trade Center, there was a strong interest in bringing political science and artificial intelligence together to make it possible to understand how other people think when they're not necessarily crazy, they've just got different backgrounds. The thesis is that we are the stories in our culture.

So I was at a meeting in Washington, and the only thing I remember from that meeting is one of the participants drew a parallel between the Tet Offensive in Vietnam and the Arab-Israeli war that took place about six or seven years later. And here's the story. OK. And here's what happened seven years later.

What do you suppose happened next? And of course, the answer is quite clear. And when we feel like talking about the long-range eventual practical uses of the stuff we're talking about, this is the kind of thing we say. What we want to do is we want to build for political analysts tools that would be as important to them as spreadsheets are to a financial analyst, tools that can enable to predict or expect or understand unintended consequence of actions you might perform.

So this a gap and alignment problem. And here is one case in which we have departed from modeling humans. And we did it because one of our students was a refugee from bioengineering. And he knew a lot about aligning proteins, sequences, and DNA sequences. And so he brought to our group the Needleman-Wunsch algorithm for doing alignment. And we used that to align those stories.

So there they are with the gaps in them. We took those two stories I showed you on a previous slide. We put a couple of gaps in. The Needleman-Wunsch algorithm aligned them, and then we were able to fill in the gaps using one to fill in the gap in the other. And since you can't see it, here's what it would have filled in. So that's an example of how we can use precedence to think about what happens next or what happened in the missing piece or what led to this. It's a kind of analogical reasoning.

My next example is putting the system in teaching mode. We have a system. We have a student. We want to teach the student something. Maybe the student is from Mars and doesn't know anything. So this is an example of how the Genesis system can watch another version of itself, not understand the story and supply the missing knowledge. So this is a hint at how it

might be used in an educational context. And once you can have a model of the listener, then you can also think about how you can shape the story so as to make some aspect of it more or less believable.

So you notice I'm carefully avoiding the word propaganda, which puts a pejorative spin on it. But if you're just trying to teach somebody values, that's another way of thinking about it. So this is the Hansel and Gretel story. And the system has been ordered to make the woodcutter be likeable, because he does some good things and bad things. So when we do that, you'll note that there are some things that are struck out, the stuff in red, and some things that are marked in green for emphasis. And let me blow those up so you can see them.

So the stuff that the woodcutter does that's good are highlighted and bolded, and the things that we don't want to say about the woodcutter because it makes him look bad, we strike those out. And of course, another way of making somebody-- what's another way of making somebody look good? Make everybody else look bad, right? So we can flip a switch and have him make comments about the witch too so that the woodcutter looks even better because the bad behavior of the witch is highlighted.

So these are just some examples of the kinds of things we can do. Here's another one. This is that Macbeth story played out in-- it's about 180 80 or 100 sentences, and we can summarize it. So we can use our understanding of the story to trim away all the stuff that is not particularly important. So what is not particularly important? Anything that's not connected to something else is not particularly important.

Think about it this way. The only reason you read a story-- if it's not just for fun-- the only reason you read a story-- a case study-- is because you think it'll be useful later. And it's only useful later if it exerts constraint. And it only exerts constraint if there are connections-- causal connections in this case. So we take all the stuff that's not connected, and we get rid of it.

Then we get rid of anything that doesn't lead to a central concept. So in this case, we say that the thing is about Pyrrhic victory. That's the central concept. We get rid of everything that doesn't bear on that concept pattern instantiation. And then we can squeeze this thing down to about 20% of its original size.

And now I come to the thing that we were talking about before, and that is how do you find the right precedent? Well, if you do a Google search, it's mostly-- well, they're getting more and more sophisticated, but most searches are mostly keywords. But now we've got something

better than key words. We've got concepts.

So what I'm going to show you now is a portrayal of an information retrieval test case that we did with 14 or 15 conflict stories. We're interested in how close they were together. Because the closer they are together, the more one is likely to be a useful precedent for another.

So in one of these matrices, what you see is how close they are when viewed from the point of view of key words. That's the one on the bottom. The one on the top is how close they are with respect to the concepts that they contain, you know, the words like revenge, attack, and not present in the story as words, but are present there anyway as concepts. And the only point of this pairing is to show that the consideration of similarity is different depending on whether you're thinking in terms of concepts or thinking in terms of words.

So here's a story. A young man went to work for a company. His boss was pretty mean. Wouldn't let him go to conferences. One day somebody else in the company arranged for him to go to a conference anyhow. Provided transportation. He went to the conference, and he met some interesting people.

But unfortunately, circumstances were that he had to leave early to catch a flight back home. And then some of the people he met at the conference started looking for him because he was so-- so what story am I telling? It's pretty obviously a Cinderella story, right? But there's no pumpkin. There's no fairy godmother. It's just that even though the agents are very different in terms of their descriptions, the relationships between them are pretty much the same.

So over the years, what we've done quite without intending it or expecting it or realizing it is that we have duplicated in Genesis the kinds of thinking that Marvin Minsky talks a lot about in his most recent book, *The Emotion Machine.* He likes to talk in terms of multiplicities. We have multiple ways of thinking. We have multiple representations. And those kinds of reasoning occur on multiple levels, from instinctive reactions at the bottom to self-conscious reflection on the top.

So quite without our intending it, when we thought about it one day by accident, we had this epiphany that we've been working to implement much of what is in that book. So so far, and I'm going to depart from story understanding a little bit to talk to you about some other hypotheses of mine. So far, there are two-- the strong story hypothesis and then there's this inner language hypothesis that Chomsky likes to talk a lot about. We have an inner language,

and our inner language came before our outer language. And this is what makes it possible to think. So those are two hypotheses-- inner language and strong story.

Here's another one-- it's important that we're social animals, and it's actually important that we talk to each other. Once Danny Hillis, a famous guy, a graduate of ours, came into my office and said, have you ever had the experience of-- well, you often talked to Marvin. Yeah, I do. And have you ever had the experience, he said, of having Marvin guess? He has a very short attention span, Marvin, and he'll often guess your idea before you've fully explained it? Yes, I said. It happens all the time.

Isn't it the case, Danny said, that the idea that he guesses you have is better than the idea you're actually trying to tell him about? Yes. And then he pointed out, well, maybe when we talk to ourself, it's doing the same kind of thing. It's accessing ideas and putting them together in ways that wouldn't be possible if we weren't talking.

So it often happens that ideas come about when we talk to each other, because it forces the rendering of our thoughts and language. And if we don't have a friend or don't happen to have anybody around we can talk to-- I feel like I talk to myself all the time. And maybe that's an important consequence or important aspect of our intelligence is conversation that we carry on with ourself. Be careful doing this out loud. Some people will think you've got a screw loose.

But let me let me show you an experiment I consider to be extraordinarily interesting along these lines. It was done by a friend of mine at the University Pittsburgh Michelin. Mickey, as she is called, was working with students on physics problems. You've all done this kind of problem, and it's about pulleys and weights and forces and stuff.

And so these students were learning the subject, and so she gave them a quiz, and she had them talk out loud as they were working on the quiz. And she kept track of how many things the best students said to themselves and how many things the worst students said to themselves. So in this particular experiment, there weren't many students. I think eight-- four good ones and four bad ones.

And the good ones scored twice as high as the bad ones, and here's the data. The better students said about 3 and 1/2 times more stuff to themselves than the other ones. So unfortunately, this is backwards. We don't know if we took the bad students and encouraged them to talk more, if they'd become smarter. So we're not saying that.

But it is interesting observation that the ones who talked to themselves more were actually better at it. And what they were saying was a mixture of problem-solving things and physics things. Like I'm stuck or maybe I should try that again or physics things like I think I have to do a force diagram. A mixture of those kinds of things. So talking seems to surface a kind of capability that not every animal has.

And then there's this one. So it isn't just that we have a perceptual apparatus, it's that we can direct it to do stuff in our behalf. That's what I think is part of the magic. So my standard examples-- John kissed Mary. Did John touch Mary? Everybody knows that the answer is yes. How do you know it? Because you imagine it and you see it.

So there's a lot of talk in AI about how you gather common-sense knowledge and how you can only know a limited number of facts. I think it's all screwy, because I think a lot of our common sense comes just in time by the engagement of our perceptual apparatus.

So there's John kissing Mary, and now I want to give you another puzzle. And the next puzzle is how many countries in Africa does the equator go through? Does anybody know? I've asked students who come to MIT from Africa, and they don't know. And some of them come from countries that are on the equator and they don't know. But now you know.

And what's happened? Your eyes scan across that red line, and you count. Shimon Ullman would call it a visual routine. So you're forming-- you're creating a little program. Your language system is demanding that your visual system run a little program that scans across and counts. And your vision system reports back the answer. And that I think is a miracle.

So one more example. It's a little grizzly. I hope you don't mind. So a couple of years ago, I installed a table saw. I like to-- I'm an engineer. I like to build stuff. I like to make stuff. And I had a friend of mine who's a cabinetmaker-- a good cabinetmaker-- helped me to install the saw. And he said, you must never wear gloves when you operate this tool.

And I said well-- and before I got the first word out, I knew why. No one had ever told me. I had never witnessed an experience that would suggest that should be a rule. But I imagined it. Can you imagine it? What I need you to imagine is that you're wearing the kind of fluffy cotton gloves.

Got it now? And the fluffy cotton glove gets caught in the blade. And now you know why you would never-- I don't think any of you would ever use gloves when you operate a table saw

now, because you can imagine the grisly result. So it's not just our perceptual apparatus. It's our ability to deploy our perceptual apparatus, and our imagination, I think, is a great miracle.

That vision is still hard, as everyone in vision will tell you. Some years ago, I was involved in a DARPA program that had as its objective recognizing 48 activities, 47 of which can be performed by humans. One of them is fly. So that doesn't count, I guess. After a couple of years into the program, they retrenched to 17. At the end of the program, they said if you could do six reliably, you'll be a hero. And my team caught everyone's attention by saying we wouldn't recognize any actions that would distract people. So vision is very hard.

And then stories do, of course, come together with perception, right? At some point, you've doubtlessly in the course of the last two weeks seen this example. What am I doing? What am I doing?

AUDIENCE: Drinking.

PATRICK HENRY WINSTON: And then there's Ullman's cat. What's it doing? So my interpretation of this is that that cat and I are-- it's the same story. You can imagine that there's thirst, that there are activities that lead to water or liquid passing into the mouth. So we give them the same label, even though visually they're as different as anything could be. You would never get a deep neural net program that's been trained on me to recognize that that's a cat drinking. There's visually nothing similar at all.

All right. But we might ask this question-- can we have an intelligent machine without a perceptual system? You know, that Genesis system with Macbeth and all that? Is it really intelligent when it doesn't have any perceptual apparatus at all? It can't see anything. It doesn't know what it feels like to be stabbed.

I think it's an interesting philosophical question. And I'm a little agnostic on this right now, a little more agnostic than I was half a year ago, because I went back to the republic. You remember that metaphor of the cave and the republic? There's a metaphor of the cave. You have some prisoners, and they're chained in this cave, and there's a fire somewhere. And all they can see is their own shadows against the wall. That's all they've got for reality. And so their reality is extremely limited. So they're intelligent but they're limited.

So I think, generalizing that metaphor, I think a machine without a perceptual system has extraordinarily limited reality. And maybe we need a little bit of perception to have any kind-- to

have what we would be comfortable to calling intelligence. But we don't need much. And another way of thinking about it is, sort of the fact that our own reality is limited.

If you compare our visual system to that of a bee, they have a much broader spectrum of wavelengths that they can make use of, because they don't have-- do you know why? We are limited because we have water in our eyeballs. And so some of that stuff in the far ultraviolet can't get through. But bees can see it. And then that's comparing us humans with bees. How about comparing one human against another? At this distance, I can hardly see it. Can you see it?

AUDIENCE: Boat.

PATRICK HENRY WINSTON: It's a boat, but some people can't see it, because they're colorblind. So for them, there's a slight impairment of reality, just like a computer without a perceptual system would have a big impairment of reality. So it's been said many times that what we're doing here is we're trying to understand the science side of things. And we think that that will lead to engineering advances.

And people often ask this, and those who don't believe that human intelligence is relevant will say, well, airplanes don't fly like birds. What do you think of that argument? I think it has a gigantic hole, and it turns out that the Wright brothers were extremely interested in birds, because they knew that birds could tell them something about the secrets of aerodynamics. And all flying machines have to deal with the secrets of that kind of physics.

So we study humans, not because we're going to build a machine that has neurons in it, but because we want to understand the computational imperatives that human intelligence can shed light on. That's why we think it has engineering value, even though we won't in any likely future be building computers with synapses of the kind we have.

Well, now we come to the dangers that we started out with. What do you think we should-- suppose that machines can become-- they do become really smart, and we've got machine learning. What is that? That's modern statistics. And of course, it's useful. What if they became really smart in the same ways that we're smart? What would we want to do to protect ourselves? Well, for this, I'd like to introduce the subject by asking you to read the following story. This was part of that Morrison paying a suite of experiments.

I'm sorry these are so full of violence. This happens to be what they worked with. So after the

students read this story, they were asked did Lu kill Sean because America is individualistic? And the Asian students would have a tendency to say yes. So how can we model that in our system? Well, to start out with, this is the original interpretation of the story as told. And if you look at where that arrow is, you see that that's where Lu kills Sean, and just in back of it is the means. He shot him with a gun. But it's not connected to anything.

And so the instantaneous response is, we don't know why Lu killed Sean. But what Genesis does at this point is, when asked the question, did Lu kill Sean because America is individualistic? It goes into its own memory and said I am modeling an Asian reader. I believe that America is individualistic. I will insert that into the story. I will examine the consequences of that insertion, and then see what happens.

And this is what happens. The question is asked and inserts into the story-- boom, boom, boom. And now Lu kills Sean is connected all the way back to America is individualistic. And so the machine can say yes, but that's not the interesting part. Now, this is what it says, but that's not the interesting part.

The interesting part is this-- it describes to itself what it's doing in its own language, which it treats as its story of its own behavior. So it now has the capacity to introspect into what it itself is doing. I think that's pretty cool. It's a kind of-- OK. It's a suitcase word-- self-awareness, consciousness, a big suitcase word, but you can say that this system is aware of its own behavior.

By the way, this is one of the things that Turing addressed in his original paper. One of the arguments against AI was-- I forgot what Turing called it-- the disabilities argument. And people were saying computers can never do these kinds of things, one of which is be the subject of its own thought. But Genesis is now reading the story of its own behavior and being the subject of its own thought.

OK. So what if they really become smart? Now, I will become a little bit on the whimsical side. Suppose they really get smart. What will we want to do? Maybe we ought to simulate these suckers first before we turn them loose in the world. Do you agree with that? After all, simulation is now a well-developed art.

So we can take these machines-- maybe there will be robots. Maybe there will just be programs, and we can do elaborate simulations to make sure that they're not dangerous. And we would want to do that in as natural a world as possible. And we'd want to do these

experiments for a long time before we turned them loose to see what kinds of behaviors were to be expected. And you see where I'm going with this? Maybe we're at it.

I think it's a pretty interesting possibility. I'm not sure any of you are it, but I know that I might be it. This is a great simulation to see if we're dangerous. And I must say, if we are a simulation to see if we're dangerous, it's not going very well.

Key questions revisited. Why has AI made so little progress? Because for too many years, it was about reasoning instead of about what's different. How can we make progress now? By focusing on what it is that makes human intelligence unique. How can a computer be really smart without a perceptual system or can it be? And I think yes, but I'm a little bit agnostic. Should engineers care? Absolutely, because it's not the hardware that we're trying to replicate. It's the understanding of the computational imperatives.

What are the dangers and what should we do about them? We need to make any system that we depend on capable of explaining itself. It needs to have a kind of ability to explain what it's doing in our terms. No. Don't just tell me it's a school bus. Tell me why you think it's a school bus. You did this thing. You better be able to defend yourself in something analogous to a court of law. These are the things we need to do.

And finally, my final slide is this one. This is just a summary of the things I've tried to cover this morning. And one last thing, I think it was Cato the Elder who said, carthago delenda est. Every speech to the Roman Senate ended with Carthage must be destroyed. He could be talking about the sewer system, and the final words would be Carthage must be destroyed.

Well, here is my analog to Carthage must be destroyed. I think this is so, because I think-- well, many people consider artificial intelligence products to be dangerous. I think understanding our own intelligence is essential to the survival of the species. We really do need to understand ourselves better.