2.4 Machines and Logic

[00:00] One of the more pointed comparisons between the ways that people think and the way that machines think has to do with debates and discussions around the theme of logic. Logic is a very broad term. Logic in general refers to a tradition of thinking that goes back to at least as far as Aristotle who wrote about logic. Aristotle's portrait of logic centered around syllogisms, patterns of reasoning like “All men are mortal. Socrates is a man, therefore Socrates is mortal.” You've probably heard these kinds of syllogisms before. Syllogisms were the mainstay of logic for a very long time, but logic over the past couple of centuries has expanded and changed and grown more expressive in a lot of ways.

[01:11] The discussion having to do with machines and minds often centers on this idea that machines are particularly logical or the ways in which computers can think is especially logical. That's a complex question. We can program computers in ways that are less rigidly logical than we often associate with computational information processing or thinking. That is to say, we can program computers so that they can behave a little less rigidly, a little less formally, a little less strictly according to the rules of established logic. Nonetheless, logic fits well. It lends itself to computational implementation, at least in a lot of cases.

[02:17] So let me begin not quite at the beginning, but with the beginning of modern logic. So I mentioned discussions of logic go back to Aristotle. For our purposes, modern logic starts with a book by the English mathematician George Boole called *The Laws of Thought*. It's a very difficult and challenging book, in part because Boole was thinking out all these ideas on the page. So it's not the best introduction to ideas of logic. In fact, in Boole's book, he mixes different formalisms from things that today would be referred to as Boolean logic, that is, logic having to do with ones and zeros, propositional logic, that is logic, based on the idea of manipulating true or false sentences, and finally, set theory. All of those things are mixed in Boole's book and that makes it rather difficult to read. The explanations often shift from one domain to another. All those topics are in fact closely related, but Boole's treatment is difficult to read. So I wouldn't advise picking up Boole's book to learn about propositional logic or Boolean logic or set theory. I wouldn't advise reading it for that purpose, but it's an interesting book to read once you have learned something about the modern formalisms of logic to go back and see where those ideas began.

[04:01] In any event, we'll focus for the time being on propositional logic. So, in propositional logic, the idea is that letters or symbols stand for entire propositions that can be true or false. Like “five is a prime” - that happens to be true. “Paris is the capital of Italy” - that happens to be false. Think of P and Q and R and other symbols here as representing declarative sentences, the easiest way to think about them, that could either be true or false, but not both. Now, propositional logic allows you to reason with sentences of that kind by combining them with what are called connectives, like “and” and “or”, “if then” and then doing certain fairly straightforward reasoning using the propositions and the sentences that you've asserted.

[05:07] I've just got one teeny example here. I'm not going to go into propositional logic in depth. If you haven't seen it, you don't need to know it in depth for the purposes of this discussion, But this is the kind of reasoning that gets done in propositional logic. So you have a bunch of assertions. You have a bunch of sentences that you say are true - we're going to treat as true. Then from those sentences, we will see what else we can deduce that should also be true. So in this case, we've asserted three things. Number one, we've asserted IF P OR (NOT Q). That is to say, if it is the case that P OR (NOT Q) THEN R" Again, what the meaning of P and Q and R should be depends on the use of propositional logic. When you're using it, you might substitute for P and Q and R things that make sense in this particular form of reasoning. For our purposes, we're just leaving these as uninterpreted symbols. We're just stating as a given that if P OR (NOT Q) is true, then R is true. We also are given, we are told, that Q is true and that P is true. From those three sentences, we can deduce still other true statements. For example, sentence four tells us since we know that P is true, we also know that P OR (NOT Q) is true. We didn't even use sentence two in this case. But from sentence three, we can deduce that since P is true, it must be the case that P OR (NOT Q).

[07:08] Now, one thing I should mention is that in standard propositional logic OR is interpreted as what we would call inclusive OR. P OR (NOT Q) is a true statement if P is true, if (NOT Q) is true, or if *both* are true. So the OR here is inclusive as opposed to the exclusive OR, which is usually written XOR and which is only true if one or the other, but not both, are true. In this case, it happens in the case of sentence four, it happens to be the case. We happen to know that P is true and (NOT Q) is false. Still P OR (NOT Q) is true. So this would be true regardless of whether we had, in this case, we could interpret OR as inclusive or exclusive OR, but for sentence four, we've written it with an inclusive OR. Sentence five then follows from sentence four and one. We know that P OR (NOT Q) is true from sentence four. Therefore, plugging that into sentence one, we can deduce that R is true. Because if P OR(NOT Q) is true then R is true.

[08:27] You may notice that this feels like a very formal and roundabout and effortful way to deduce things, and indeed, it is in practice. As I also mentioned though, computers are really good at this. So propositional logic does lend itself well to programmed implementation. There are many complications in doing that, but yes - computers are quite good at reasoning with propositional logic. So in this sense, this is a kind of logic that machines do well with. But even as we're talking about this, think of the title of Boole's book which we saw in the previous slide. Boole's book was called *The Laws of Thought*. He was writing in the 1850s before there were any computers. The way that he regarded logic was that this is the way, not only that thinking should be, but in a sense good thinking is, that people reason with propositional logic and that they should reason with propositional logic. So the discussions around logic as applied to people are often an uneasy mix of descriptive and normative. Sometimes people want to argue that we do think logically, and sometimes people want to argue that even if we don't, we should: this is a good way of thinking. It certainly lends itself well to machine reasoning, in many cases, not in all. But it often lends itself well to machine reasoning because it follows sets of formal rules that for people can often be stressful to follow in ways that are free of mistakes. We often make mistakes. Machines when suitably programmed can be very effective at doing this.

[10:45] You may have noticed that even in this reasoning, I made use of certain kinds of deductive steps. In propositional logic, a number of these standard deductive steps are called modus ponens. It's a very old rule in logic. Basically it says that if you know that “if A, then B” is true, if you know that that's true, and you know that A is true, then you know that B is true. Sounds pretty straightforward. So the example on the slide: “if there is fire there is smoke.” We know there is fire, therefore, there is smoke. An equivalent rule, that is to say, equivalent to modus ponens (means really the very same thing), but it goes by a different name - modus tollens. It has a different syntactic structure where we can say, “if there is fire, there is smoke”, and we know “there is not smoke”, therefore “there is not fire”. Because if there were fire, there would be smoke.

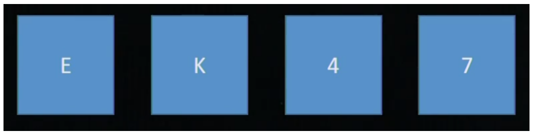
[11:55] Now, to a machine, these two rules are essentially identical. By the way, I've mentioned the difference between inclusive and exclusive or, I should mention that in propositional logic, “if A then B” means something very specific. In English, we often use if-then sentences in a looser informal way. But in propositional logic, when you say “if A then B”, what you mean is, that will be true, if A is true and B is true. It is also true if not A is true, and B is either true or false. In other words, let's take this first sentence: “if there is fire, there is smoke”. If there is fire and there is smoke, that’s a good step for modus ponens. If there is not fire, we don't really know whether there is smoke or not. Or to put it another way, the only way that the sentence if there is fire there is smoke could be wrong, the only way that sentence could be false, is if it is the case that there is fire and there is not smoke. That would prove the sentence false. All other three possibilities - fire and smoke, not fire and smoke, not fire and not smoke - would allow the sentence to be true.

[13:31] So in propositional logic classes, people are very careful about interpreting if-then sentences. For example, the sentence “if Paris is the capital of Italy, then two equals five”. That is a true sentence. The first part of the if-then sentence is false, and from that point, it doesn't really matter whether the second part is true or false. “If Paris is the capital of Italy, then five is a prime” - that's also true because the first part is false. The only version of an if-then statement that can be false is if I say something like “if Paris is the capital of France, then two equals five”. That's a false statement because the first part is true, but the second part is false. But we have to be that long-winded to explain the nature of if-then in propositional logic as opposed to colloquial English.

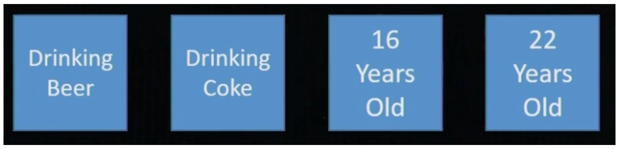
[14:36] Modus ponens and modus tollens are both totally perfect deductions from two earlier sentences. They only have a slightly different syntactic structure. For a machine, these two things are essentially the same. For a person, they're not quite the same. People find it much easier to reason using modus ponens than they do using modus tollens. So that already should give us a hint that the ways in which typical computer programs deal with propositional logic and the ways in which people often informally deal with logic already have some differences to them.

[15:25] People also are prone to make mistakes in this kind of logic. So these are two examples of fallacies in logic, deductions that are not true, but that are tempting often because of the way we think of an if-then statement. So if we have a sentence like “if there is fire, there is smoke” and then we're told “there is not fire”, it is not correct to deduce that there is not smoke because after all there might be smoke for other reasons than fire. So this is called denial of the antecedent fallacy, and this is a form of faulty reasoning. So is the second example: “If there is fire, there is smoke”, and we know “there is smoke”, “therefore, there is fire”. Incorrect - again, there could be smoke for other reasons than fire. So these are both examples of problematic or incorrect uses of logic that people are prone to. There are other logical mistakes that people are not prone to make. But the fact that we are prone to make certain logical mistakes, again, tells us that there's more to human reasoning than is expressed in the formal rules of logic. Even though Boole wanted to call his book *The Laws of Thought*, propositional logic doesn't seem to be the laws of thought. It seems to be a formal representation of certain kinds of effortful thought, but we don't follow the rules of propositional logic terribly faithfully.

[17:14] There's a classic experiment around these lines that was done back in the 1960s. I will give this to you as an example. Here is the experiment. People are given four cards. I've represented the four cards here in the slide. They are told that on one side of the card is a letter and on the other side of the card is a number. Then you are given a rule, which your job is to test. You want to test this rule to see whether it's true or false. So the rule that you're going to be given to test is “If a card has a vowel on one side, then it has an even number on the other side”. And you are asked, given this set of four cards, to turn over exactly and only those cards that you would need to test whether this rule is true. So this takes some thinking when people do it, and they often make errors. I don't quite recall the statistics, but many many people make errors in this task. As it turns out, you need to turn over exactly two cards. You need to turn over the E. If the card has a vowel on one side, then it has an even number on the other side. If you turn over the E and you find an odd number, you know the rule is false. The other card that you have to turn over is the seven. If you turn over the seven and you find a vowel like an A, then you know the rule is false. The other two cards, you don't need to turn over because no matter what you see on the other side, you won't learn anything about the truth or falseness of this rule. You won't learn anything about whether the rule happens to be true or false. If you turn over a K, then regardless of whether there's an even or odd number on the other side, it doesn't tell you anything about this rule. If you turn over the four, maybe you might see a vowel that would be consistent with the rule, but you may see a consonant that would also be consistent with the rule. So the only cards that you need to turn over are ones that could disprove the rule. In a sense, this is a tiny model of some philosophical treatments of scientific reasoning in which the purpose for experiments is to disprove a theory. Most scientists don't feel that that's an accurate representation of scientific pursuit. But nonetheless, some people argue that when you do an experiment, it should be with an eye toward disproving a theory. In this case, turning over the E or turning over the seven could disprove this rule. Turning over the other two could do nothing to disprove it.



[20:33] As I say, people have a lot of difficulty with this problem. But one interesting thing is that they have far less difficulty with a, what in some sense, is an identical problem - an identically structured problem. But in this case, people have a much easier time solving the problem. The distinction between people's performance in these two cases is interesting to speculate about. So here's the new task. You're given four cards. Usually, people say, imagine yourself as a bartender or something like that. You're given four cards. On one side of the card



is the drink that a person is having. On the other side of the card is the person's age. Your job is to see whether this rule is true, or if you want to phrase it this way, whether this law is being upheld. If a person is drinking beer, then the person must be over 19 years of age. Your job is to turn over only and exactly those cards that will show whether the rule is being held to. Here, people have a much easier time. You turn over the drinking beer card. If the age on the other side is 16, then the rule is being violated. You turn over the 16 year old card. If the person is drinking beer, the rule is being violated. The other two cards, you don't need to turn over. Because, for example, turning over the drinking Coke card can tell you nothing about whether the rule is being violated, similarly with the 22 year old person.

[22:31] So why is it that this task seems to be so much easier than the previous task involving letters and numbers? There are different explanations for this. There's not a unique explanation to it. The original experimenters who presented this version of the card task would make an argument that goes roughly as follows: we are very good reasoners when it comes to situations that are ecologically or evolutionarily realistic for us. Now, you might say being a bartender is not evolutionarily realistic, but seeing whether laws are being followed, seeing whether rules are being obeyed or violated, it's a venerable human activity. In communities, we often care quite a bit about whether people in a community are obeying the group laws or whether they're not. So in this case, we might say that we're exercising not so much a kind of abstract talent for logic, but a rather specific talent or rather specific human ability to detect cheaters in legal situations. That's not the only explanation for this distinction, but it's one explanation. In any event, just to leave you with this sort of reflection, what we've seen is that the ways in which people reason in situations that could be modeled logically doesn't seem to be quite the same as the ways in which the formal rules of logic dictate or pure mathematics dictates or, in fact, the ways in which it is relatively easy to program machines. Machines can be made to follow logic reliably. For us, it seems to be more of a problem.

2.5 Judgment and Decision Making

[00:00] We've spent some time talking about the computational view of problem-solving that is in contrast and in comparison to the way that humans solve problems. We've looked at how understanding problem-solving in machines can lend new insights into how we understand problem-solving in people. Many of the problems that we looked at in that earlier discussion had to do with things like puzzles or fairly abstract mathematical problems or problems in the physical sciences. Those are very important types of problems that we often have to solve in a professional way. But now, we're moving to a discussion of other kinds, you might call them problem-solving, but in the literature, they usually go by the names of tasks in judgment and decision-making.

[01:11] Now off hand, there are some gray areas between what we might call judgment and decision-making on the one hand, and problem-solving on the other hand. Let's take some archetypal situations which were thought of as judgment or decision-making. You are in the market to buy a new car. So which kind of car are you going to get? That's a decision-making problem. You have a number of distinct models of car. You could get a used car, you get a new car, you could get cars of different makes and so forth. Which car are you going to get? That's a decision-making problem. You might also think of it as a problem. You're going to search the space of choices that you could make and make one of them. But often, we don't think of these, the Rubik's Cube or puzzle problems, in the same way that we think of decision-making problems. For one thing, decision-making in the literature usually has to do with situations that are broad, that are perhaps a little less well-defined than chess problems or Rubik's Cube, situations where we may have to make a decision with incomplete information or fairly quickly, situations where maybe the information is, you might think of it as incomplete or you might think of it as just so enormous that we can't take into account all of the potential information that we might have. We have to make some kind of decision based on all kinds of different factors.

[03:02] These are situations, if we were being uncharitable, we might say that problem-solving situations are the ones that are less realistic in our lives and decision-making or judgment problems are the ones that often we care about in our lives. Should we take this course or that course? What should we major in? Having dated the person once, should we date them again? Is it worthwhile to invest in this house? All kinds of large scale important decisions in life get cast under this notion of judgment and decision-making. What that means also, because these situations are often more important to us, is that we don't have the genial point of view about getting them wrong. Let me put it this way. If you're doing a Rubik's cube and you fail to solve the problem, you may be frustrated, and it may matter to you. But you may, in talking to your friends, just say, "I'm not very good at Rubik's cube. That's life. What are you going to do?" On the other hand, people don't feel that way about things like what they're going to major in or who they're going to marry or what city they're going to live in. People take those decisions pretty seriously, and they don't take a light view of themselves if they habitually get these things wrong. So it's like, “well, I keep choosing the wrong profession”. That has a more important ring to it than “I can't solve Rubik's cube”.

[05:00] So when people talk about judgment and decision-making, and the ways in which people make judgments, there's a kind of, I refer to it here as a normative stance as opposed to a descriptive stance. We're not only interested in how people solve problems or, in this case, how people make decisions, we’re also interested in the question: are there continual mistakes or bad habits that come into our decision-making? And are those linked? Can those be elucidated by a computational viewpoint? By looking at these decisions and judgments from the computational standpoint, can we understand not only the descriptive reasons that we make certain decisions, but can we improve the way that we make certain decisions? You could see the normative stances being important in classic problem solving as well. We'd like to learn how to solve Rubik's cube and things like that. But it takes on a more urgent note when we're talking about things like judging what subject to major in or a potential spouse or a potential city to live in or things like that. These are large-scale decisions, often with lots and lots of factors, and we'd like to avoid making easy mistakes anyway. So some of the judgment and decision-making literature has this flavor of, what are the mistakes we make, and can we do better?

[06:56] Similarly, there's a distinction between problem-solving, and judgment and decision-making. And again, that's not a hard and fast distinction, but a loose distinction between how we think of this evolutionarily. Think puzzles, math problems, science problems - the staples of the problem-solving literature - those feel to us like evolutionarily recent situations. It's not like we evolved having to solve problems in logic or abstract mathematics. These may be still very important to us in our lives, but things of this sort don't seem to have deep evolutionary roots.

[07:48] On the other hand, there are situations in judgment that you could argue have deep evolutionary roots. If you are encountering a new person, a person that you haven't seen before, are they a friend or a foe? What cues might you use to determine whether to be nervous around this new person or whether to be fairly accepting of this new person? If you meet up with an animal, is it a dangerous animal or is it an animal that does not pose a threat? Perhaps, it's an animal that could be prey for us. These are judgments that we make based on lots and lots of factors. And the idea of judgment and decision-making is that it's something that we as a species, we have evolved with certain ways of making these kinds of judgments, and often these techniques that we've evolved with serve us very well. They've evolved for a reason. We survived because we had these techniques of making judgments. However, the techniques that we evolved with may not be perfect, particularly for our situation now, that is, the situation of our species now in, say, urban or developed environments. So we'll talk about some of these things.

[09:26] Let me begin by giving an example that illustrates some of the themes that come up in the judgment and decision-making literature. So this is an example of what can be called problem-framing. It's also an example of certain economic decisions, and you may think of it as still being mathematical, but it illustrates - you might call them difficulties that we have in making certain judgments. So imagine this is an offer that's being made to you. So you're being offered, first, a bonus of $300. I'm going to give you $300. Then you are asked to choose between the two following alternatives (one or the other alternatives, one or the other possibilities is what you're going to choose): either to receive another $100 for sure, or to flip a coin. If you win the toss, you get $200. If you lose the toss, you get nothing. So you think to yourself what choice would you make? I've given you $300, and now I'm going to give you a choice between just taking another $100, or flipping a coin and receiving $200 if you win and zero if you lose. In either case, you keep the original $300. Now, if you know anything about probability, you know that the expected value of both choices A and B is $400 at the end of the whole situation. In other words, in Situation 1, you definitely have $400. In Situation 2, you have a 50 percent chance of having $500 and 50 percent chance of having $300. So you might think offhand, there's no reason to choose between A and B, but that's not correct. You may have a very good reason for choosing A or B in this situation. The expected value of A and B is the same, but they're not identical situations. So think to yourself what you would choose in this situation.

[11:59] Now, here's a different situation. Let me pose this for you. I give you $500, and now I offer you two alternative possibilities. I'm going to take $100 back from you, we'll call that possibility C, and possibility D: you'll flip a coin. If you win the toss, you lose nothing. If you lose the toss, you have to pay $200. Now, think to yourself which of these two alternatives you would choose. Now, here's the interesting point: these two situations are, in fact, identical. The one that I showed you before, which I'll go back to here, that's really the very same situation as this one. In choice C here, you end up with $400, for sure, and in choice D, you have a 50 percent chance of having $500, and a 50 percent chance of having $300. That was the same case for choices A and B. They feel different though for a lot of people, and I confess this is true for me too. When you look at this choice, your mileage may vary, but I tend to look at this choice and think I'll go for A. I'll take the additional $100 and end up with $400. In this case, I think you just gave me $500, I don't want to just give you back $100 - I'll gamble to keep what I've got. Now, again, there may be a very good reason for choosing A over B, or B over A. What there isn't is any rational explanation offhand for why you would choose A in this situation and D in this situation. They are identical in the long-term. At the end of the day, they're identical situations, and we shouldn't switch our decision merely because of the way the problem has been framed.

[14:33] So this is an example of problem framing. For a lot of people, when they see an example of this kind, they get a little bit disturbed by it, like thinking, "Well, what's wrong with me? Maybe there's something wrong with the way I'm thinking about this." This is taking advantage, by the way, of a well-documented tendency in people's economic judgments, that they tend to value more the things that they already have as opposed to the things that they don't yet have, but can bargain for. So once you own something, once I've given you that $500, there's a greater threshold to giving up some of it. In the first situation, I gave you only $300, and the idea of getting an additional 100 seemed attractive. This is one example of the interesting phenomena that turn up in judgment and decision-making.

[15:46] Here's a related one. You could call it a problem framing example. It's from a very fun book by cognitive scientist Dan Ariely. I think he describes himself as a behavioral communist, but in any event, here's the situation. It's a kind of problem framing, but this one is taken, I think, from real-life. He describes it in the book, and this is an ad that he saw about subscriptions to the magazine, *The Economist*. So notice the three choices that are being given here. *The Economist,* you could get a subscription to the website to just economist.com for $59. You get a print subscription to *The Economist* for $125. You could get a print plus web subscription for $125. Now, you look at those three choices. If you're like me anyway, the first thing you think is what kind of lunatic would choose choice number two? Seeing as choice number three is the very same cost, but they're throwing in $59 value for free. You get the print plus web subscription for the exact same cost that you would have for the print subscription. Why in heaven's name would anyone choose point number two? Indeed, that's quite reasonable. There is a reason that *The Economist* gives you these three choices, however. They're intending for you to focus on the last two choices. Since it's an obvious decision between those last two choices, you would choose number three, print plus web subscription. That's what they want you to choose. They would prefer you to choose print plus web subscription for $125 rather than just a web subscription. That's their preference. They're making this sub-choice easy for you. So, in fact, people tend to disproportionately, when given three choices, two of which are easily comparable and one of those two is obviously better, people tend to choose the better of those two and ignore the third choice altogether. Is that rational? Well, there are reasons for it. There may be good reasons for it in human thinking. But in another respect, it's an example where our judgment is being pushed in one direction or another.

[18:51] Ariely describes a very interesting experiment in his same discussion of this, and I'll see if I can describe it for you. It takes a little setting up. He wanted to see if he could design a situation where people would be prompted to make the choice that he wanted them to make by giving them a similar three-way choice that *The Economist* does here. So here's what he did. First, he took lots of photos. You could do this, by the way, as the mathematicians say, without loss of generality. You could do this with photos of men and women looking at photos of men or you could do it with photos of women and men looking at photos of women. So apparently, the phenomenon works the same so we'll just say for description that we're dealing with photos of males, and there are women looking at photos of males. And in this case, they are rating the photos by attractiveness. So in the first phase of the experiment, you show many, many photos of males to female subjects and have them rate these photos according to attractiveness. Having done that, let's just say that you have two photos of two particular males. These are not famous people or anything; they're just photos of random folks. But we have two photos of males, call them A and B, who in the first phase of the experiment have been found to be equally attractive. So without any other information, if you were to take a new female subject and show her photos of A and B, you might expect about a 50-50 chance that she would find A more attractive than B, or vice versa.

[21:04] Now, here's what Ariely did. He took some of these photos. Let's take the photo A and the photo B, and remember these are equally attractive folks, and he used a little bit of photoshop on one or another of these photos. So let's say that we take A, and we do a photoshop on photo A to produce a new photo - we'll call it A’. Now, this being a psychological experiment, the photo of A is morphed a little bit to be a little asymmetric or the face, the head shape is a little odd or something like that. In other words, A’ looks kind of like A, but less attractive. Now, if you show these three photos to a new female subject, someone who hasn't been involved in the first part of the experiment, you show her A, A’, and B, and you ask which of these three people is most attractive? If the choice A’ had not been presented, you would expect about a 50-50 chance that she would choose A or B. However, the photo A’ here acts like choice two in *The Economist* situation. So she looks at these three photos and is much more likely to choose A. A’ is like that print subscription to *The Economist*. No one's going to choose A’, but the choice between A and A’ is easy, and so the focus is driven to make a choice between A and A’. Similarly, if you do not show the photo A’, but do a similar photoshop job, and now show A, B, and B’ to a new female subject, she'll be much more likely to choose B. The choice between B and B’ is easy, and the focus is then made on this easy choice, and it is much more likely that she'll choose B than A. So what would have been a fairly difficult or delicate decision between A and B, can be made to be an almost automatic decision, by giving an irrelevant third choice that somehow drives attention toward one choice or the other. That's pretty interesting.

[24:12] Now you might think again that's irrational. By the way, Ariely says there's a lesson in this. Whether you're male or female, if you're going out looking for a date, maybe going to a bar or a party or something, he says it's a good idea to bring along somebody who looks just like you, but a little less attractive. That may work. Of course, you don't want to tell the other person that that's why you're inviting them. But anyway, that would fall into the category of this problem framing. Again, when some people look at examples like this, like *The Economist* example over this photo experiment, they get a little disturbed about it. What it means is that their judgment is perhaps not as foolproof as they felt it was going to be, and this is where this normative side of judgment and decision-making comes in. We'd like our judgment to be, if we want to put it this way, as rational as possible. We'd like to be able to explain to ourselves why we made certain judgments, and we'd like those explanations to feel as though they're in accord with what we think of as reason.

[25:50] This is often rather difficult. Reason is a slippery concept in situations like this. We may begin by asking are these really mistakes altogether? Maybe we don't want to think of them as mistakes. Maybe there's a certain sense in that first situation, if you have a certain amount of money or if you have an object, maybe there are good reasons why you should work harder to keep what you have and be less willing to give it up than in situations where you don't have something and you might gamble to get a bit more. In other words, maybe there are good reasons that you should value what you own already, place a greater value on those objects than on objects that you don't have. That may not be a mistake. But if they are mistakes, if we want to cast them as mistakes, what are the mechanisms that cause us to make them? Those are really interesting questions. Could we mimic those processes in a machine or could we make machines that make judgments which are less prone to human type errors? Those are twin questions about this thing.

[27:19] This last example gives rise to yet another question, which is should we be worried that others, perhaps others whose interests are not aligned with ours, are trying to do things to manipulate our decision making? *The Economist*, in Ariely's example, was trying to manipulate the situation so that you would make a particular choice. Now, you may end up being happy with that choice, but it's clear that *The Economist's* point of view on your decision, and yours, are not identical. So now we deal with the situation where other people or institutions can exploit the properties of our decision making. Those are questions that we need to deal with as we discuss the subject of judgment and decision making in general.

2.6 Heuristics and Biases in Judgment

[00:00] While we're on the subject of judgment and decision-making, we began talking about some of the issues involved in human judgment and decision-making, and we contrasted those with some of the ways in which we discussed this from a computational viewpoint when we were talking about more abstract problem-solving. What I want to do right now, in this particular session, is describe a couple of more examples of what have been called heuristics and biases in judgment. Much of the pioneering literature on this, a lot of the original work, was done by Daniel Kahneman and Amos Tversky.

[01:04] Kahneman, a few years ago, wrote a very interesting book called *Thinking Fast and Slow*, which is a compendium of a lot of the work that he and Tversky, who died some years ago, did together. Much of the research that they did on certain kinds of heuristics and biases that people use in making decisions or rendering judgments. Now, the word heuristic, again, is a description of a rule of thumb, a computational technique that we can use that will help us solve problems, but is not guaranteed to solve problems. So heuristics are things that we can try, that are often good ideas, not always, but ways in which we can approach a particular problem. Our association with words like bias is more negative. The idea being that we have certain ways of thinking about judgments, and these cloud our judgment. In any event, you'll get a sense of some of the phenomena.

[02:37] What I'm going to try and describe are just a couple of the standard experiments that are used to demonstrate these phenomena in judgment. So here's an example. This is actually a much discussed and somewhat controversial example. When Kahneman wrote about it in his book, he mentioned that this particular example and the ones along the same lines, has garnered the most debate about its importance, what it means, whether it's a problem. Let me show it to you first. It goes by a general name called the Conjunction Effect. The idea is here you first see a description of a person. So you're given a little story about this person Bill: “Bill is 34 years old. He is intelligent, but unimaginative, compulsive, and generally lifeless. In school, he was strong in mathematics but weak in social studies and humanities.” That's the description you're given about Bill. Now you're given a collection of statements about Bill, like these, and your job is to rank order these statements from most to least probable. That's the idea. So you take these statements about Bill, and you order them from most to least probable. Now, when people think about that description of Bill and they look at these statements about Bill, stereotypes being what they are, people often tend to assume that perhaps “Bill is an accountant”. Fairly reasonable. They don't think that it's especially likely, for example, that “Bill plays jazz for a hobby”. The three statements that we're interested in here, the others are additional choices, but the three statements for us to focus on here are: “Bill is an accountant”, “Bill plays jazz for a hobby”, and “Bill is an accountant who plays jazz for a hobby”.

[04:58] Now, here's the thing. When people answer this question, they rank “Bill is an accountant” as most probable, “Bill is an accountant who plays jazz for a hobby” as next most probable, and “Bill plays jazz for a hobby” as least probable. In other words, they view this statement as more likely than this statement. Now, mathematically, that's impossible. Actually, if you think about it from the standpoint of set theory, imagine all the situations in which Bill is an accountant, and Bill plays jazz for a hobby, and Bill is an accountant who plays jazz for a hobby. If he is an accountant who plays jazz for a hobby, that's a subset. A smaller number of situations, a much smaller number of situations, than all the situations in which Bill plays jazz for a hobby. In other words, it simply can't be more probable that Bill is an accountant who plays jazz for a hobby than that he simply plays jazz for a hobby. It can't be more probable. The number of situations in which this is true is a small subset of the situations in which this is true. However, people view this as more probable.

[06:26] Now, when people see this example, me included, I read this example, thought about it, yeah, of course once you think about it, it has to be more likely that Bill plays jazz for a hobby than that he plays jazz for a hobby and he has a particular profession. Sometimes people look at this and they think, "Well, the difficulty is kind of the word probable. What do you mean by probable?" People might not, for example, be interpreting this question as given 10 million distinct situations in the world, rank order the frequency in which they occur. What they may be responding to, at least what many students when they see this problem, what people may be responding to is not so much what's probable mathematically as where do you think this story is going? What's a probable ending to this story? That's a fuzzier question, but you might reason that why are you telling me these things about Bill unless you're telling me something that is informative or archetypical about the kind of profession that he has. So maybe you could argue back and forth about how a word like probability is interpreted for something like this. And indeed, questions about probability are pervasive in the judgment and decision-making literature, and how we interpret probability.

[08:18] Probability is actually a rather remarkable and complex and many mathematicians would say beautiful area of mathematics, but it's one that has controversies associated with it - about how to interpret probability. So not everyone interprets probability in the same way. It should also be mentioned that the notations, the formal discussion of probability, is relatively young in human affairs. That is to say, you could trace it back to people like say Pascal, who wrote about probability in the 1600s. Now, the 1600s may feel to us like "Oh, that's a long time ago." It's not all that long ago, 300, 400 years ago. So we're dealing with ideas about probability that we did not evolve with. Therefore, when we look at questions that involve some probability, often we're dealing with concepts that are relatively new in human experience.

[09:30] Let me give you another example of a similar problem that is cited in the judgment and decision-making literature. So here's another description of somebody, not Bill, a different person, Steve. Steve has some similarities with Bill, but let's ignore those. “Steve is very shy and withdrawn, invariably helpful, but with little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.” Now you're asked to rank order the probability that Steve has one of these professions. Is Steve an architect, farmer, librarian, biologist, or taxi driver? Now again, the issue here is somewhat different than it was in the previous example. When people look at these choices, and again thinking in terms of stereotypes, they often hone in on the idea that Steve is a librarian. Seems like the kind of person who might be a librarian. So they decide that librarian is the most probable choice for Steve's profession.

[10:53] Now, one way of looking at this question though is to say, really and truly, given what we know, let's assume that we're dealing with the population of adult men in America. What's the most probable profession for Steve given this description? Well, let's just put it this way. Agriculture is not the prominent profession that it used to be in America, but it's still a pretty important profession. Let us ignore the description of Steve for the moment, and just look at these five professions, and say without any information at all, where are the numbers? I mean, what are the most likely of these professions? Certainly, it is much more likely for a randomly chosen adult man in America to be a farmer than to be a librarian. There may not be nearly as many farmers as there used to be, but there's still lots more farmers than librarians around. So one way of thinking about this, going back to the original question, is to say well, Steve could be a typical librarian, or he could be a fairly shy orderly farmer. And since there are, let's say 500 times more farmers than there are librarians, it may really be much more likely that Steve is a farmer, even a somewhat unusual farmer, than that he is a fairly typical librarian.

[12:42] The idea being introduced here is sometimes called base rates in probability reasoning. Without any other information, given no information, we would find it much more likely if we were choosing a random person out of a huge urn, and we were just taking descriptions of people, adult males in America, out of a huge urn, it's much more likely that we'll choose a farmer than a librarian. If this happens to be a description of a farmer, that may still be much more likely than that we have chosen a description of a librarian. So again, this is not something that people generally take into account when dealing with a situation like this. Kahneman and Tversky argue that we have a blind spot when it comes to these base rates and that this can cost us in certain decision-making situations.

[13:46] We'll get back to discussions of that idea, but let me show you a couple of other phenomena. One is called, it's usually put forward in situations that involve numbers or quantities (needn't always), but it's called anchoring. This is an example of the kind of question that displays anchoring when people try to answer it. So you take a population of subjects and you ask them a question like “Was Ronald Reagan 120 years old when he died?” or “Was Ronald Reagan 50 years old when he died?” Now, both of those answers are wrong, and the answer is no in both cases. When people see these questions they know well Reagan lived past 50 years old and just about nobody lives to 120 years old. So when they look at these two questions they answer no. But then in both cases, if you ask the person a subsequent question, “how old was he when he died?” - the people who are asked the first question will reliably give you a larger number, an older number than the people who were asked the second question. In other words, when people are asked the first question and then they say, "No, Reagan was not 120", you ask them how old was he when he died, and it appears that what they're doing more or less, to a good approximation, is taking the original number of 120 and reasoning downward from it, moving downward until they find a number where they think that might be the age at which Reagan died. In the second case, they're taking the number 50 and moving upward until they think they've got a reasonable number. In both cases, what people do is they end up with an age guess that is closer to the number, if you want to put it this way, the number that they were primed with. So you give people this question, and you can move their numeric judgment in one way or another by getting them to start from an original position, an anchoring position.

So this is an example of what could be called a bias in judgment, and one that might be exploited in other situations.

[16:19] Here's another example. It's an example that draws on research from memory, from the studies of recall of words and facts and so forth. So people are asked this question: “Estimate the proportion of English words that begin with the letter K versus words that have a K in the third position.” Now, as it turns out, there are more words in English with K in the third position than there are words with K in the first position. However, the structure of our memory for words seems to be indexed by first letter, maybe first sound. In any event, when people are asked to come up with words that either have K in the third position or K in the first position, they find it easier to come up with words with K in the first position. That's an easier job. They can think of the words that they know that begin with K. It takes more effort to think of the words that we know with K in the third position. So in this case, even though the correct answer is there are more words with K in the third position, the words that begin with K are more available to us, more available to our memories,and that leads us to make a mistake in answering this question. You could view it as a mistake in judgment.

[17:58] These are just a few of the examples of the literature and judgment and decision-making. You notice that many of these examples have this flavor that you can present people with situations where it then appears that people are making foolish decisions. Now, the question of whether these are really bad decisions, whether this is really a problem for us, that itself is a matter of some debate. There are researchers who feel that these are not especially important or dire problems in judgment for us. Maybe in some sense, we don't even want to call them mistakes ,but if we do want to call them mistakes, are they terribly important? If they are important, why do we make them and can we train ourselves not to make them? Some people in the literature describe these as illusions, analogous to optical illusions, which means that they could be fairly pessimistic about the idea that we can train ourselves not to make these errors. When we see certain kinds of optical illusions, even if we know that there's an illusion, take illusions where you see two lines which happen to be the same length. Here, I'll draw one for you. The famous Müller-Lyer illusion. I don't know if I'm doing it perfectly, but here's the idea. You're given these two lines, one of which has arrowheads pointing inward on the ends and the other has, if you want to call them that, arrowheads pointing outward. And the lines are the same length, but often when we look at these two lines we view this one as being smaller and this one as being longer.

[20:12] It's hard to train yourself out of that illusion. It's not a question of, you may even know intellectually that the two lines are the same length, and yet when you look at the picture, you keep saying to yourself “they look like they're different lengths - I can't help myself”. Some people describe the phenomena involving judgment and decision making in a similar way, that we might even know that this isn't an error, but we can't quite help ourselves. So that leads to other questions about whether even if we know that we're making these errors, can we fix it? A related question is can we design machines? Can we design programs whose heuristics or biases are different than those that we have or that are perhaps more mathematically rigorous or use different internal models of probability than the ones that we typically use? Can we design programs that are far less likely to make these kinds of errors than human beings are?

Those are interesting questions, and as we continue to talk about the judgment and decision-making literature, we'll touch on some of these debates.