[SQUEAKING]

[RUSTLING]

[CLICKING]

JOSH MCDERMOTT: So the next topic on our agenda is attention. And so attention is a-- it's a big word that can mean a lot of different things. Colloquially, you're familiar with the notion of paying attention to something versus maybe ignoring it.

And we're talking about attention in a class on perception because attention often determines what you see and hear and in general perceive about the world. So in vision, what you're attending to often is a function of what you look at. So attention oftentimes moves with your eyes, but it doesn't have to. So if you're being really sneaky, you could be looking at me, but maybe paying attention to something that's out here.

So you can pay attention to things if you're not fixating them. So I guess, naively and intuitively, when you pay attention to something, that seems to improve, in some way, the processing of whatever's attended. That's like presumably why we do it, why we use attention. Like when you pay attention to something, you see it better, or you hear it better in some sense. And we're, of course, not specifying what it means for it to be better, so that's an important issue.

Again, intuitively, people often think of this maybe as a way to focus "resources." Again, it's in scare quotes because we don't really know what the resources are, but it's perhaps if resources do exist, and you need them for perception, and they're limited, this could be a way to focus them on whatever seems important. So in this lecture, we're going to talk about how we can measure and study these phenomena.

And so attention has been studied with a dizzying array of paradigms. So this lecture will provide an overview of some of them, some of the best known, that are very relevant to perception. But it's going to be a little bit of a grab bag. So right now, we lack, I would say, a unifying framework for thinking about attention. And it's almost surely not one thing. So there's lots of different things that often get called attention, and they're not necessarily the same thing.

And at the moment, we mostly, I would say, lack rigorous computational understanding. So that's actually a super exciting direction for the future. And I think we're at a point now where we could actually really start to develop rigorous computational theories of attention. But there are nonetheless many interesting phenomena that place constraints on future theoretical understanding. So we're going to go through some of them and talk about what they show us and what they mean.

So here's an initial example. So very popular paradigm for studying attention involves cuing people. So Posner was a cognitive psychologist who did a lot of work on cuing. So the essential idea is to focus visual attention to an area of space using what's called a cue. So that's like a stimulus that will occur at some part of space, or that will indicate some part of space where something's going to happen.

And then in the classical paradigms for studying this, they would measure the time that it would take someone to identify a target stimulus when either the observer does not know where the item will appear versus where they do know where the item will appear by virtue of the cue. And so in this initial example that I'll show, the cue might just be a briefly presented dot at the location of the target. So look in the middle here at the center of the cross.

And so your task is going to be to say what letter is going to appear. So you can all do this task, so do it. You're really slow. What's the letter?

AUDIENCE: A

JOSH MCDERMOTT: A. Wow. That was the longest reaction time in the history of attention experiments. You got it right, though. It is an A.

So this is the uncued condition, and it is supposed to be slower than the cued condition. Here's the cued condition.

AUDIENCE: N.

JOSH MCDERMOTT: N. That was much faster. So that's the cuing benefit. all

So this is a schematic of the kind of results. So what is measured here is reaction time. So again, you assume that most of the time-- and you verify that most of the time people get the letter correct. So you measure how long it takes them to identify the letter. And there's a reduction in reaction time when you're cued to the location. And so in this sense, advanced knowledge of the location improves performance as measured by reaction time.

So the other thing that is associated with this phenomena is that there is-- it's spatially localized. So this is, again, kind of a fake graph schematized the results of lots and lots of experiments. And the graph is plotting the reduction in reaction time that you get from the cue as a function of the cued location relative to the target. And so the point is that when the cue is exactly at the location of the target, that's where you get the biggest benefit. And then if the target is off of the cued location by some amount, there's a reduction in the benefit.

So this type of finding led to a very influential and popular metaphor of attention as a spotlight. So the idea is that there's this thing that's spatially localized. It gets cued to a location, it gets moved to that location, does something that is, as of now, unspecified to the processing that makes you faster or more accurate or whatever.

So the spotlight metaphor of attention. So the idea is that whatever's in the spotlight is attended. The more it's attended, the better it's processed. And the size and the shape of the spotlight can be controlled to some extent.

So now another really important distinction that came out of this sort of study is the idea that there can be two types of cues. So the one that we just saw is what would be normally referred to as exogenous. So that means outside generating.

So the dot flashes up, without really thinking about it, like your attention is kind of drawn to the location of this thing that pops up on the screen. It's almost like a reflex. So typically, exogenous cues would be sudden changes, so flashes or movement. They draw attention automatically.

So by comparison, the other kind of cue that you can get in general is what's called endogenous. So that stands for inside generating So this typically we think requires some kind of high-level control, and typically involves an instruction. So some kind of visual sign or pattern that will be a symbol of where the target is going to appear.

So here's an example of exogenous cuing. Sorry, this is the exogenous one that we just saw, just to repeat it. So there's the dot, and then the letter So you don't have to interpret anything. Your attention just does its thing.

Here's an example of endogenous cuing. So we've got an arrow, and then the letter appears. So the point is that in order for you to benefit from the cue, you have to understand what arrows mean. And willfully move your attention over in that direction, and then you get a benefit.

So both types of cues seem to control an attentional mechanism-- maybe not the same one. But they reflect different strategies. So we think of exogenous cues as tapping into some kind of bottom-up control of attention that's based on physical transients in the environment, and endogenous cues as involving what's often called top-down control of attention based on what an observer believes.

So how do these things actually cash out in terms of the results? Well, one of the clear differences that you see is what's shown here. So this is a graph that is plotting the same thing that we looked at before. So the y-axis is the benefit in reaction time from a valid cue. So it's how much faster you are if you're cued to the location.

But now what's being manipulated here is what's called the stimulus onset asynchrony. And so that's indicated here. So here's what would happen on a trial of an experiment where you're measuring this stuff. This is a case where the cue is exogenous.

So you're fixating that little star, and then the cue pops up. In this case, it's this red box. So something changes, your attention moves over to the left. And then after some amount of time, the target appears. And that some amount of time is the SOA, the Stimulus Onset Asynchrony. So essentially, you can vary the amount of time from when the cue pops up until when that thing that you have to detect pops up.

And so what does this graph show? Well, it shows that initially-- so if the stimulus onset asynchrony is really, really short, then you don't get the benefit. So it's like the cue does something that takes some amount of time to happen. And then in the case of what's called the peripheral cue here, that's the exogenous one, the red box. By the time you-- there's about 100 milliseconds, you're close to getting the maximal benefit from the cue.

But when you have a symbolic cue, what we're calling endogenous, where you have to interpret it, that's like the arrow. It takes you longer to get the benefit because something else has to happen in your head. You see this symbol, and there's some process of interpretation, and that takes a bit more time. So this is some evidence that these two types of cues are doing different things in your head Any questions about that?

And both of these types of things happen all the time in everyday life. So there's very frequently transient things happen, attention is drawn to some location-- because that's how we talk about it. Other times there'll be something that's more symbolic that indicates to you-- so, for instance-- I don't know-- you see a turn signal. It indicates the thing's going to turn to the left, and so you're paying attention over there. Yeah.

AUDIENCE: Is there are always these binary difference between the exogenous and the other one? Or could you imagine an arrow that's pointing upward, but the visual stimulus is immediately there as well? So it's orienting you. is there some of, I guess, gray area?

JOSH MCDERMOTT: Yeah. I mean, I think that often in a lot of real-world scenarios, you're getting combinations of these two things, where there'll be some transient, but you also just-- because of your understanding of the situation, you know that something's going to happen in a particular area. I don't know. Let's say you're watching baseball. And so the pitcher is throwing the ball. So you know that it's going to be caught in some region around the strike zone.

But then the catcher puts up the mitt, and you hear the sound of the ball slapping in. And then you get this more exogenous cue for where it's happening. So that's like a combination of things right. So I think that kind of stuff happens a lot. Yeah. I don't think these are necessarily mutually exclusive.

So that's a very common distinction in the world of attention, exogenous and endogenous attention. So these kinds of cuing studies, they gave rise to this idea that you could think of attention as a spotlight. And so that was very persistent for several decades.

And then people got very interested in whether or not you could actually attend to multiple things at the same time. And so the paradigm that is shown here became wildly popular in the world of visual attention. This is what's often called multiple object tracking.

And so what happens in this paradigm is there's a bunch of objects that are on a screen. And you are cued to attend to a subset of them. So in this case, the red ones are the ones that you're supposed to attend to. So you get this initial cue. Then the cue disappears.

And at this point, all of the objects are basically identical. So the only thing that distinguishes them is the fact that you know that you're supposed to be paying attention to a subset of them. Then they start moving around. So they move around for a while.

And then at the end, one of them is going to turn red again. And you have to say, is the one that turned red, is that part of the initial set of red balls, or is it not? So it's a measure of the extent to which you can keep track of all these different objects. Let's see if you can do it.

So you need to fixate that little square at the center. The flashing ones are the ones you got to pay attention to. And now they're going to move around.

So did you feel like you could pay attention to a bunch of them? Yeah. You also probably noticed that gets pretty hard when two things kind of overlap. So one of the things that you're trying to track is-- crosses paths with one of the other ones. That gets kind of hard. So there's all kinds of interesting stuff that happens here.

But people have some ability to do this. This paradigm was introduced by Zenon Pylyshyn, who is a famous cognitive psychologist at Rutgers. This is a graph that shows the proportion of errors that are made as a function of the number of targets that people are tracking.

And you can see that people get worse as the number of targets increases. But the point is that the proportion of errors is still relatively low, even out to five things. So people seem to have the ability to track multiple things at the same time.

And so this paradigm kind of became a little cottage industry. And there were many, many, many experiments measuring how people could do this under different conditions and things like that. So the general conclusions from this are that people can attend to multiple locations at once.

So attention is not just a spotlight, maybe there are multiple spotlights. It's still the case, though, that in a lot of cases, the spotlight metaphor works pretty well. So there's often kind of one main thing that you're attending to. Any questions about multiple object tracking? Yeah.

AUDIENCE: Is it learnable. Like can you train to get better at it?

JOSH MCDERMOTT: So can you train to get better? So-- probably. I'm pretty sure that there's a study that I know about, which I hope is not apocryphal, but I believe these experiments were ran on-- I think it was like the Canadian Olympic basketball team.

So basketball is a great example where you maybe have to do something kind of like this, because there's all these different players, and you need to keep track of who's where and stuff like that. And so I believe that there was a finding that the capacity for multiple object tracking was better on people who were really good at basketball. So I think there's been some studies like that showing that some people are better than others.

I don't know how trainable that is. It could also just be that the people who are good at multiple object tracking end up being the ones that are good at basketball or something. So yeah.

The other thing that is related to this is there's, again, another small cottage industry of studies on people who play lots of video games. And in general, the finding is that people who play lots of video games seem to do better on a lot of visual attention tasks. I don't know if they were ever tested on this one in particular, but some of the other effects that we will see later in the lecture show benefits from video game playing.

So lots of attention research has been conducted in vision, but it's also-- it's really important for hearing as well. You remember back when we were talking about audition, we talked a lot about the cocktail party problem, where there's one person you're trying to understand, so you have to pay attention to their voice, omit these distractors. And so one situation where attention may be important is where there are concurrent sources that are similar, and you have to keep track of one that you're trying to understand.

And so this is a graph that is showing you the trajectories that are taken by two voices that are-- two people talking at the same time through a space of features that we think are important for voices. So F0 is the fundamental frequency. Everybody remembers the fundamental frequency, right? It's thought to be like the correlate of pitch. It's the rate of repetition.

And then the other thing-- another thing that's really important about voices are the formants. So the first two formants define vowels, the first order, so F1 and F2. And so when you talk, you're constantly modulating your fundamental frequency and your formants. And so you can think of a voice as kind of moving around in that three-dimensional feature space.

And so these are example trajectories. The yellow line is for one voice, and the blue line is for another voice. And so the point is just that they're all kind of tangled up. So this is like two-- I think there are two female speakers that are talking at the same time.

And so naively, you might think that if you want to follow what one person is saying, you might need to track that voice as it moves through the feature space. And so this is an example task that was devised to look at this. I'm going to play you an example of the stimulus, but these are synthetic voices that are continuously voiced vowels.

So they move around in this space. And there's going to be two of them. And this is a task that was devised to measure whether or not you can track one of the two voices. And so the way that this works is you initially get a cue, and that tells you which of the two voices you're supposed to listen to.

And that cue is the very initial part of the target voice, so the green one in this case. So then you get a mixture of two voices. And so you can see this is like two harmonic things on top of each other. It's just spectrograms, obviously.

And then after the mixture, you get what's called a probe. And so the probe is the end portion of one of the two voices. So half the time, it will be the end portion from the green voice, the target voice. Half the time it will be the end portion from the other voice that you're supposed to ignore. And so you just have to say, yes, the probe is from the cued voice, or, no, it isn't.

And the idea is that because these things kind of are circling around each other in the same part of the space, in order to perform this task, you have to be able to track this thing over time to essentially be able to connect the cued portion of the voice to the end portion of the voice. So you can try it for yourself. You're initially going to hear the cue, and then you're supposed to track the cued voice.

And then listen to how it ends. And then you get a probe, and you have to say whether the probe was the one at the end or not-- was the end of the target voice, rather.

[AUDIO PLAYBACK]

[OVERLAPPING VOCALIZATIONS]

[SINGLE VOCALIZATION]

[END PLAYBACK]

JOSH MCDERMOTT: Yes or no?

AUDIENCE: No.

JOSH MCDERMOTT: Yeah? OK, good. Some of you're not sure. Here's another one.

[AUDIO PLAYBACK]

[OVERLAPPING VOCALIZATIONS]

[SINGLE VOCALIZATION]

[END PLAYBACK]

JOSH MCDERMOTT: Yeah. So in both of those cases, it was. So, interestingly, it turns out that this task is much easier for musicians. And I think many of you who were smiling are probably musically trained.

But people can do this. So one question is like, well, it seems like in order to do this task, you would have to track this thing over time with your attention. And one way to potentially measure this, and a conventional way to measure whether you're attending to something, is to see whether there is some other kind of attentional benefit to attending to that thing. So in other words, are you better at noticing stuff about that thing?

And so a way that this was looked at in this particular paradigm was by adding a little bit of vibrato to one of the voices. So the vibrato could either appear on the cued voice or on the uncued voice. So you have to perform this tracking task, but then in addition, you have to say, yes, there was vibrato or no, there was not. And so half the time, there would be vibrato on one of the two voices, and half the time there wouldn't be.

And so what this graph is showing-- so this is how good people are at detecting vibrato. So this is d prime-- remember, sensitivity, right? And when the vibrato appears on the cued voice, performance is better than when it appears on the uncued voice.

But the other thing that's kind of interesting is that if you take the people who are performing this task and you split them up into the ones who are good at the tracking task, and not so good-- so that's the good streamers and the poor streamers. The good streamers show a pretty big advantage for detecting the vibrato in the cued voice compared to the uncued voice, whereas the poor streamers don't.

So that really indicates that your ability to track this thing over time does really depend on being able to selectively attend just to that thing. And that makes you more sensitive to its features compared to the other stuff. And this is just showing that advantage for detecting the vibrato in the cued voice is present throughout the whole time the stimulus is on. So another kind of paradigm for studying this kind of stuff, in this case in hearing.

So in vision, you often are attending to particular spatial locations, but not always. You can also attend to features. So what I want you to do here is attend to the blue elements. And when you attend to the blue elements, you become very aware that there's this diagonal organization to the blue elements.

Now I want you to attend to the red elements. And you become aware that there's a circular organization. Now I want you to attend to horizontal. And you become aware that there's another diagonal organization in the opposite place.

So attention is pretty flexible. You can attend to locations, also to certain classes of features. And one paradigm for studying this type of attention and its interaction with location is visual search. This is another kind of classic paradigm

So you're all going to do these tasks. In this particular task, you have to say whether there is a blue dot.

AUDIENCE: Yes.

JOSH MCDERMOTT: OK. A little slow, but you got it right So the result of doing experiments like this, and this is kind of obvious when you look at this display. This is super easy. As soon as it pops up, you see there's a blue dot.

So for some kinds of targets and scenes, visual search is always fast. And we say that in these situations, that the target kind of pops out of the display. Remember back when we were talking about grouping, we also talked about this phenomenon of pop-out, where you have one element that kind of differs from all the rest on some simple dimension, and it pops out. It's really obvious.

So the way that you objectively measure this is by measuring how long it takes you to detect the target as a function of the number of items that are in the display, which is often called the set size. So this is how you analyze visual search experiments, reaction time versus set size. And in a situation like this, where pop-out occurs, the graph is flat.

So the standard explanation of this is that for some kinds of properties, a unique value in the image will draw attention. So it serves as an exogenous cue. So the fact that that one thing was blue and everything else was yellow immediately draws your attention to that location.

So other kinds of things that have the same effect are a single large item among small ones, or a single curved item among straight ones. There's lots of things like that. And the diagnostic property is that the search slope of that results graph that you make of reaction time versus set size, the search slope is zero. So that's what corresponds to pop-out. But not everything is-- not everything pops out.

So is there a blue vertical line? That was slow. So that's much harder than saying whether there was a blue thing amongst yellow So this is what is called a conjunction search. So conjunction searches are often pretty slow.

So conjunction searches, that just means that the target is not defined by a unique value of one feature, it's defined by a conjunction of features. So you can see that there's blue stuff and vertical stuff, and the target is just the only one that is blue and vertical. So in some kinds of situations like this, search is slow. And the diagnostic results graph that you would get in this situation-- again, it's reaction time versus set size-- is that the reaction time increases with the number of elements in the display with-- that is the set size

So it's almost as though you have to look around at all of the elements until you find the one that contains this combination of properties. And so the more things there are, the more things you have to look at. And so the reaction time kind of gets higher

So here are some more examples. So in these feature search examples, the three on the left, again, it's kind of immediate. In these different conjunction search examples where you have to find the red vertical, each time it takes you takes you a minute to see it. These spatial configuration searches, again, it's a sort of conjunction, a certain combination of features. Finding the T takes a little while.

So again, the results of these experiments would be expressed as reaction time versus set size graphs. In the cases where there's what's called a feature search, you get pop-out and the slope is zero. In these other cases, the slope is positive.

So there's two lines in these graphs, because in order to actually do this experiment-- so normally, the way the experiment is done is you say whether the target is there or not. So sometimes the target will be present, and sometimes it will be absent. And so there are two results graphs here. One for the trials where the target is present, and one for where the target is absent.

And you will notice that the slopes are twice as big when the target is absent as when the target is present. Does anyone want to posit an explanation for why that would be the case?

AUDIENCE: To double-check.

JOSH MCDERMOTT: Say it again.

AUDIENCE: Double-check.

JOSH MCDERMOTT: Double-check. What do you mean by that?

AUDIENCE: You have to double-check that it's not there.

JOSH MCDERMOTT: All right. What's double-check mean, though?

AUDIENCE: Look once, look twice.

JOSH MCDERMOTT: What do you think?

AUDIENCE: Well, you can short circuit and finish searching early once you find it. You only have to look through half the image, potentially.

JOSH MCDERMOTT: On average.

AUDIENCE: Yeah.

JOSH MCDERMOTT: Yeah. That's right. So under a model where you're kind of looking around at every element in the image, then if the target is there, on average, you'll find it after looking through half of the things. Whereas if it's not there, to be sure, you got to look at everything. That's probably what you meant by double-check.

So this is the phenomenon, that in some cases, the reaction time doesn't scale with the number of items, and in other cases, it does. So the standard explanation for these slow searches is that you have to combine properties in order to detect the target, and the combination process is not automatic. So it's commonly proposed that you need attention in order to do that combination. So the spotlight of attention is often proposed to weld these different elementary features together.

And this makes search a serial process because the spotlight of attention under this explanation can only be in one place at a time, so you have to move that around. And every time it's on something, you get the features combined. And so if you look at what these slopes are like, you can infer that the spotlight travels at about 50 milliseconds per item.

Here's another example. Find the red verticals. This one is really annoying. There's a couple of them there. Yeah. Yeah. It's a tough conjunction search.

So this, on the one hand-- it's a very interesting phenomenon. So we're finding that there's this structure in the visual system, and that's kind of interesting. But it's important also because it has a lot of real-world relevance.

So oftentimes the things that you're looking for in the real world-- and you do visual search all the time. Where are my keys? Like where did I leave my coffee cup? Where's my kid's left shoe? It's like visual search is a big part of life.

And oftentimes what you're looking for is defined by conjunctions. So if I ask you to find the faucet here, raise your hand when you see the faucet. Yeah. See, it takes a little while. So visual search, it's an important thing, and a classic paradigm.

So visual search became very popular around 1980, and the popularity in part was due to some very influential papers by Anne Treisman, who was a very important cognitive psychologist who worked at Berkeley, and then Princeton. She died a few years ago. And the general proposal was that there are these maps of different features in the brain, these elementary features, orientation, color, size, stuff like that-- spatial frequency-- that we're kind of loosely associated with the kinds of things that we think of as being extracted by early vision.

So you have these different maps, and attention kind of serves to bind different features together. So without attention, all you have are these separate maps of features, and that makes it difficult to tell whether you have a conjunction of features. So the proposal was that somehow-- again, these were not very mechanistic explanations. They're very abstract. But somehow or another, attention kind of magically enables you to combine these properties. And the conjunction search and feature search data were consistent with that.

So leaving aside the question of how this would actually work, this raised this question of what will happen in conditions where attention is not available to glue stuff together. And so this was studied with experiments like this. And it led to this phenomenon called illusory conjunctions.

So you're looking at displays like this. They're flashed up very quickly. Your task here is to report the digits. So you're looking at the middle, but you got to report, in this case, 3 and 5. So that's your main responsibility in the context of this task, is get the digits right.

So the idea is that causes people to focus their attention kind out in the periphery. But you're also supposed to report the colors and letters. And so you make sure-- you set this up in such a way that people have had a cup of coffee, they're motivated. And you can confirm that they're doing what they're supposed to do by verifying that they're good at reporting the digits.

And if you do this, then you find that the subjects will make errors at reporting the letters and colors. But the errors are not random. Most of the errors, or more of the errors, are what are called conjunction errors, where you get the wrong combination of the color and the letter.

So people generally-- they're not going to-- they don't report a color that's not there. They tend to not report a letter that's not there. But they'll get the letters and color combinations incorrect. So red letter T and a blue X. So this phenomenon of illusory conjunctions was discovered by Anne Treisman, and was taken as support for feature integration theory. Any questions about illusory conjunctions?

So this was very, very influential in the 1980s. Nowadays, I would say the star has fallen. I mean, most people don't really believe very strongly in feature integration theory. One reason for this is that over the years, through studying lots of different types of visual search tasks, lots of complexities emerged that seemed challenging to account for.

So one such complexity is that you can get pop-out for things that probably aren't represented in maps. So 3D shape is a good example. So if you look at this, the one on top, it immediately pops out to you that there's one that's kind of different from all the rest.

And the thing at the bottom is an alteration of that display that kind of kills the three-dimensional interpretation. And there, it's much harder to actually detect the target. So there's lots of examples of pop-out occurring for these kind of complicated, slightly high-level things, and that seemed inconsistent with it.

But it remains the fact-- so it remains the case that-- I mean, illusory conjunctions are a phenomenon that demands an explanation. And I think, in general, it remains the case that attention seems to change the way that collections of what we might consider to be features are represented. And I would say we still don't really have great quantitative or rigorous explanations of why that is, but it's a real phenomenon.

I'm going to end there. When we resume on Tuesday, we will finish up talking about attention.