OPTIC

Susanna Schellenberg, a dialog between philosophy, neuroscience and AI

Susanna Schellenberg, Professor of Philosophy and Cognitive Science at the Rutgers University, analyzes how the perception of the world around us is based on discriminating abilities in the neurosciences. Hence the need to develop a philosophy of perception. Interview at the frontier of science and phenomenology.

What's at the core of perception, and how, according to that core, do we get knowledge from our environment?

I argue that perception is at its core a matter of employing discriminatory capacities, that is, capacities to discriminate objects, properties, and events in our environment. We have perceptual capacities to discriminate red patches from blue patches, high notes from low notes, loud noises from quiet noises, the smell of cheese from the smell of coffee and millions more. Perceiving a red coffee cup on a wooden desk is a matter of discriminating the red color of the coffee cup from the brown color of the desk, the shape of the cup from the surround shapes, and many more such capacities. In employing such capacities, we see the coffee cup.

In neuroscience, it is standard to analyze perception as fundamentally a matter of discriminating. But in philosophy, this key aspect of perception has barely been given attention. I am trying to change that.

As I argue, seeing a red cup is a matter of employing discriminatory capacities, for example, the capacity to discriminate red from other colors and the capacity to discriminate cup shapes from other shapes.

Now experience can lead us astray: we can suffer hallucinations and illusions, and our perception is riddled with biases of various kinds. So, given how unreliable perception is, it's an important question as to how perception could give us knowledge. On the other hand, it's clear that if anything can give us knowledge then it's perception.

Perceptual capacities function to discriminate and single out objects, properties, and events of a specific kind. The perceptual capacity to discriminate and single out red fulfills its function if it is employed while in fact discriminating and singling out something red in the environment.

So, if I see a red cup by employing my perceptual capacity to discriminate red, then the function of that capacity is fulfilled. In such a case, I gain knowledge of the cup. And moreover, my perceptual state gives me justification for believing that there is a red cup in front of me.

What about cases of hallucination and illusion? From the perspective of the experiencing subject, they can be indistinguishable from cases of perception. Take a case in which someone hallucinates a red cup. So it seems to her that there is a red cup in front of her, despite the fact that there is no red cup there. To make the case extreme, let's assume that for this subject, the hallucination is indistinguishable from a perception. I argue that in such a case, she employs the very same perceptual capacities that she would be employing, were she perceiving. But since there is no red cup there, she fails to discriminate a red cup. So, the capacities she employs do not fulfill their function. Because of this failure, the hallucination cannot provide her with knowledge of her surroundings. And she would not even know that she doesn't know.

In cases of perception, hallucination, and illusion, the same mechanism is activated in our mind: employing discriminatory capacities. This explains why from our perspective these experiences can seem exactly alike. However, there are differences. In perception, the capacities employed fulfill their function: we discriminate what it seems to us is present. In hallucination and illusion, they don't: we fail to discriminate what it seems to us is present. So perceptual capacities are at the root of a unified theory of perception.

In that sense, can we also consider artificial intelligences combined with sensors as perceptual systems?
I think AI systems combined with sensors are a kind of perceptual systems. Like our perceptual systems, they take up information from their surround via discriminatory mechanisms. While there are significant differences in the physical implementations between AI systems and human beings, the underlying mechanism of discrimination is the same.

We all agree that a human with an auditory implant is still a human. Indeed, with regard to its mechanism of discrimination, auditory implants are similar to our auditory system.

If this is right, then at least when it comes to perception, we are not that special. This shouldn't bother us too much. After all, science tells us we are not different in kind to other animals when it comes to perception. Perception is a lowly mechanism that we share with other animals and, as I argue, also AI systems. There are many aspects in which we are categorically different from AI systems, for example with regard to our capacity for creativity and our ability to experience emotions—our own and those of others.

Now, to push the differences and similarities between AI systems and humans, let's imagine a human with ever more implants. Is there a point when she is no longer human and more of a robot? Let's consider an extreme case in which every single cell of the body has been replaced with an implant. We could even assume that she is now a physical duplicate of a robot created in a computer lab. These are deep problems. But my point here is simply that when it comes to perception, we are not that different form other animals and AI systems.

Could an artificial consciousness be possible?

If you take my view of perceptual consciousness as constated by employing perceptual capacities and, from the previous question, that AI systems with sensors can be though of as perceptual systems, then yes, AI could be perceptually conscious.

I've always been a bit surprised by the intense interest in consciousness. I am more interested in how our mind does the amazing things it does, than whether or not it is conscious while at it. Our perceptual systems process an astonishing amount of information in a fraction of a second. Approximately 50% of the human brain is devoted to visual processing. That allocation of our resources speaks to the complexity of the task at hand. A miniscule amount of the information processed bubbles up to the conscious level. A lot of the information is used to guide our action and is available for our cognitive system without ever bubbling up to the conscious level.

One thing that is certain: we are far away from developing conscious AI systems. There is recent evidence that machine learning has stalled and despite enormous efforts we are struggling developing AI systems that have the linguistic capacities of pre-schoolers. We should not worry so much about the singularity[1] or AI becoming conscious or killer robots. What we should be worried about are biased algorithms. They are here now. They have enormous implications in our lives.

Speaking about biased algorithms, with ever better sensors and computing algorithms, how is it that there are still biases?

First, it's important to note that all complex recognitional systems are riddled with biases. Both the human mind and AI operate on massive quantities of data, and whenever you have a mismatch between the quantity of inputs and the processing power available, you must simplify, and consequently some information is lost. This process generates biases.

It is important to recognize that some biases are unproblematic. They make these systems more efficient. For instance, the human perceptual system has a bias that light comes from above, and that moving objects are solid: we thus sometimes duck when a moving shadow coincides with a gust of wind.

[1] Wikipedia definition: "The technological singularity—also, simply, the singularity—is a hypothetical point in time at which technological growth becomes uncontrollable and irreversible, resulting in unforeseeable changes to human civilization."

However, some biases are deeply problematic and can be greatly harmful to marginalized and disenfranchised groups. Since all complex recognitional systems have biases and at least some of these biases are unproblematic, one big challenge is to differentiate problematic biases to the unproblematic ones. One possible approach is to look at the outcome. If the bias is harmful to a group of people, then the algorithm has to be fixed to eliminate the bias. If it is not harmful, then it can be left.

This is the current approach in dealing with algorithmic bias. While it has its place, I believe this approach is too little, too late. I am currently investigating ways of eliminating biases at an earlier stage, that is, ways in which algorithms can be developed so that the harmful biases don’t crop up in the first place.

A different point is that, contrary to what is typically assumed, most algorithmic biases are not “top-down biases”. They don’t stem from the programmers’ beliefs and background views and how those beliefs and background views affect the choices she made when designing the algorithm. To be sure, such top-down biases exist and they are a big problem. However, a similarly big problem are “bottom-up biases” not only in AI but also in the human mind. These kinds of biases derive from the incoming data, its processing at the lowest level, the patterns that the system detects in the data, and the classifications and correlations it makes.[2]

How can the philosophy of perception and so-called “hard sciences” foster a constructive dialog?

Philosophy has always been intertwined with other sciences.
Historically, there are three big questions that have motivated research on perception. First, how does perception justify beliefs and yield knowledge of our environment? This question has almost exclusively been addressed in philosophy. Second, how does perception bring about conscious mental states? This question has also primarily been addressed in philosophy. Indeed, except for a few exceptions, neuroscientists state we’re nowhere close to understanding what it means to be conscious in terms of brain mechanisms. Third, how does a perceptual system accomplish the feat of converting varying informational input into mental representations of invariant features in our environment? The latter question has mostly been addressed in neuroscience, cognitive psychology and psychophysics.

A core assumption in my research is that the answers to these three questions are not independent of each other and that to make progress in understanding the nature of perception, we need to study it in a more integrated manner.

[2] More about Susanna Schellenberg’s analysis of algorithm biases https://www.newstatesman.com/science-tech/2020/04/how-biased-algorithms-perpetuate-inequality

Hence the title of my last book “The Unity of Perception” in which I develop a unified set of tools to address all the three questions in a single theory that is conceptually disciplined and empirically constrained.
In the other direction, neuroscience is a very new field that can benefit what philosophers are good at, that is: analyzing new concepts neuroscience generates and articulating them.

Will this be the subject of your coming book?
Sure! I am working to lay out all the different ways in which AI systems and the human perceptual and cognitive system can be biased. Here are a few key distinctions:
One way for a system to be biased is if the incoming data is biased. A second way is due to how the features of the incoming data are linked and classified in the processing stage. A third is in how the output is interpreted. Another key distinction is the one I mentioned earlier between top-down biases and bottom-up biases. Bottom up biases are underexplored and need to be better understood. One further key distinction is between training-sample bias and feature-linking bias.

Training-sample bias is due to biases in the data on which an algorithm is trained. An example is Google’s speech recognition software. Initially it worked much better on male voices than on female voices. It turned out that reason for this was that it was trained predominantly on male voices. So unsurprisingly it worked much better in the range of a typical male voice. This kind of bias is easily fixed. All it takes is to expose the algorithm to lots of female voices. This kind of bias is also easily avoided. All it takes is to choose unbiased training samples. But of course, there will always be hard choices when choosing training samples. Google speech recognition works terribly for Scottish accents. Given how few English speakers have Scottish accents and given that there are many different kinds of Scottish accents, hard choices need to be made as to how much effort should be put into making speech recognition systems work for Scottish speakers.

In whatever way this problem is solved, the negative fallouts are obviously not on the same level as the fallouts form biases in algorithms used in criminal sentencing, parole, job applications, loan applications, health care, and ad generating systems. The biases in these applications are primarily feature-linking biases. Features get linked and that create biases. Such feature-linking biases are much harder to fix.

Here is an example of a feature-linking bias: If one googles an African American sounding name one is more likely to get an ad for a criminal background check than if one googles a name typically given to European decent babies. The reason for this is that the AdSense, the Google ad algorithm, detected a pattern that people were more likely to do a criminal background check after having googled a name if that name was one typical in the African American community. It then generated ads accordingly. This is a highly damaging bias, since it doesn't just perpetuate existing biases in our society, it amplifies them.

Given algorithmic biases, we have reason to be skeptical that computer generated decisions are more objective than human decisions! I think a lot of work needs to be done on dealing with these issues.

Interview by Lauriane Gorce, Scientific Director of the Institut de la technologie pour l'humain – Montréal