Thinking from the inside or the outside?

Matthew D. Lieberman

Will machines someday be able to think? And if so, should we worry about Schwarzenegger-looking machines with designs on eliminating humans from the planet because their superior decision-making would make this an obvious plan of action? As much as I love science fiction, I can't say I'm too worried about the coming robot apocalypse. I have occasionally spent time worrying about what it means to say that a machine can think. I would either say that we've been building thinking machines for centuries or I would argue that it is a dubious proposition unlikely to ever come true. What it really comes down to is whether we define thinking from a 3rd person perspective or a 1st person perspective. Is thinking something we can identify as occurring in systems like people or machines but not in ham sandwiches from the outside based on their behavior or is thinking the kind of thing that we know about from the inside because we know what thinking feels like.

The standard definition of thinking implies that it occurs if informational inputs are processed, transformed, or integrated into some type of useful output. Solving math equations is one of the simplest straightforward kinds of thinking. If you see 3 of something and then you see 4 more of that something and then you conclude there are 7 of those things overall then you have done a little bit of mathematical thinking. In addition to you being able to do that, so could Pascal's first motorized calculator in 1642. Those calculators needed the input of a human to get the '3' and the '4' but then could do the integration of those two numbers to yield '7'. Today, we could cut out the middle man by building a computer that has visual sensors and object recognition software that could easily detect the 3 things and the 4 things and then complete the addition on its own.

Is this a thinking machine? If so, then you would probably have to admit that most of your internal organs are also thinking. Your kidneys, spleen, and intestines all take inputs that could be redescribed as information and then transform these inputs into outputs. Even you brain as seen from a 3rd person perspective doesn't deal with information, strictly speaking. It's currency is electrical and chemical transmissions that neuroscientists work very hard to redescribe in terms of their informational value. If pattern *X* of electrical and chemical activity occurs as a distributed pattern in the brain when we think of '3', is that pattern the same as '3' in any intrinsic sense? It is just a convenient equivalence that we scientists use. Electrical impulses in the brain are no more intrinsically "information" or "thinking" than what goes on in our kidneys, calculators, or any of the countless other physical systems that convert inputs to outputs. We can call this thinking if we like, but if so, it is 3rd person thinking – thinking that can be identified from the outside and it is far more common than we would like to admit. Certainly the character of human or computer information transformation may be more sophisticated than other natural occurring forms of thinking, but I'm not convinced from a 3rd person perspective that they are qualitatively different.

So do humans think only in the most trivial sense? From a 3rd person perspective, I would say yes. From a 1st person perspective the story has a different punchline. Around the same time that Pascal was creating the first manmade thinking machines, Descartes wrote those famous words *cogito ergo sum* ('I think, therefore I am'), which, by the way, were cribbed from St. Augustine's writings from a thousand years earlier. For many reasons, I don't believe Descartes had it quite right but with a slight modification, we can make his philosophical bumper sticker into something both true and relevant to this debate about thinking machines.

While "I think, therefore I am" might have a touch too much bravado, "I think, therefore there is thinking" is entirely defensible. When I add "3 + 4", I might just have a conscious experience of doing so and the way I characterize this conscious experience is as a moment of thinking which is distinct from my experience of being lost in a movie or being overcome by emotion. I have certain experiences that feel like thinking and they tend to occur when I am presented with a math problem or a logic puzzle or a choice of whether to take the one marshmallow or try to wait it out for two.

This feeling of thinking might seem inconsequential, adding nothing to the computational aspects of thinking themselves – the neural firing that underpins the transforming of inputs to outputs. But consider this: countless different things in the physical world look like they are transforming inputs that could be described as information into outputs that could also be described as information. To our knowledge, humans and only humans seem to have an experience of doing so. This is 1st person thinking and it's critical that we not confuse it with 3rd person thinking.

Why does 1st thinking matter? First off, it is intrinsic. There is no way to redescribed the ongoing experience of thought as something other than thought. But whether we describe kidneys, calculators, or electrical activity in the brain observed from a 3rd person perspective as thought is arbitrary – we can do it, but we could also choose not to. The only reason we think our brain is doing a special kind of thinking is because it seems to be linked to our 1st person kind of thinking as well. But 3rd person thinking is not intrinsic - 1st person thinking is.

Second and more practically, our experience of our thinking shapes what kinds of thinking we will do next. Did if feel effortful, boring, rewarding, or inspiring to think those last thoughts? That will determine whether and how often we engage in thinking of a certain kind. I'm not suggesting that our 1st person experiences do not also have neural correlates. But no scientist or philosopher can tell you why those neural processes behaving the way they do must necessarily give rise to those experiences or any experience at all. It is one of the 3 great mysteries of the universe (that stuff exists, that life exists; that experience exists).

Will we increasingly be able to create machines that can produce input-output patterns that replicate human input-output patterns? Unquestionably. Will we be able to create machines that go beyond this and produce incredibly useful algorithms and data transformations that humans could carry out on our own and will help improve the quality of human life? We already are and will do more of this each year. Will we be able to create machines that can do 1st person thinking – that can experience their own thoughts as they have them? I don't know, but I'm not terribly confident that we will. While solving this problem would be perhaps the most magnificent achievement of mankind, we must start by recognizing that it is a problem at all. I would love to see 1st person thinking machines, but until we begin to figure out what makes us 1st person thinking machines, everything else is just a glorified calculator.