Do brain cells need to be connected to have meaning?

December 4 2012, by Lisa Zyga

Roy proposes that the only difference between distributed representation and localist representation brain models is that localist neurons have meaning by themselves, and distributed neurons do not. He argues that experimental evidence supports the view that localist neurons are widespread throughout the brain, in contrast with the connectionist brain model in which a pattern of neuronal activity is needed to represent a concept. Credit: SW Ranson

(Medical Xpress)—The classic theory of the brain is one of connections, in which the brain consists of a network of neurons that interact with each other to allow us to think, see, interpret, and understand the world around us. In this model, called distributed representation, an individual neuron by itself has no inherent meaning, but only contributes to a pattern of neuronal activity that has meaning. For example, a certain

pattern of many neurons fires when you think "dog" and another pattern for "cat."

"The belief in distributed representation theory is that a concept or object is not represented by a single neuron in the brain but by a pattern of activations over a number of neurons," explains Asim Roy, a professor of information systems at Arizona State University, to *Medical Xpress* . "Thus there is no single neuron in the brain representing a cat or a dog. Proponents of this theory claim that a cat or a dog is represented by its microfeatures such as legs, ears, body, tail, and so on. However, they think that neurons have absolutely no meaning on a stand-alone basis. Therefore, they go further and claim that these microfeatures are at the subsymbolic level, which means that meaning arises only when you consider the pattern of activations as a whole. Therefore, there are no neurons representing legs, ears, body, tail, etc. The representation is at a much lower level."

Roy is among a number of scientists working in the fields of neuroscience and artificial intelligence (AI) who suspect that the brain may not be as connected as distributed representation suggests. The basis of their alternative model, called localist representation, is that a single neuron can represent a dog, a cat, or any other object or concept. These neurons can be considered symbols since they have meaning on a stand-alone basis. However, as Roy explains, this doesn't necessarily mean only one neuron represents a dog; such "concept cells" are high-level neurons, which fire in response to the firing of an assortment of low-level neurons that represent the legs, ears, body, tail, etc.

"In localist representation, there could be separate neurons for a dog and a cat, and also neurons for legs, ears, body, tail, etc.," he said. "It's very similar to the model in my paper for word recognition, which is an old model from James McClelland [Chair of the Psychology Department at Stanford University] and [the late pioneering neuroscientist] David

Rumelhart. You have low-level neurons that detect letters of the alphabet and then high-level neurons for individual words. So letter neurons and word neurons, they both exist."

The origins of this dispute between localist and distributed representation goes back to the early '80s, to a dispute between the symbol processing hypothesis of artificial intelligence (AI) and the subsymbolic paradigm of connectionists. In the past 30 years, the debate has only intensified.

Not so different after all?

Staunchly on the side of the symbol model, Roy has published a paper in a recent issue of *Frontiers in Cognitive Science* in which he makes two main claims that he thinks will ramp up support for localist representation. First, he proposes that distributed representation and localist representation models are essentially the same, with just one small but important difference: localist neurons have meaning by themselves, and distributed neurons do not. Traditionally, the two models have been thought to have inherent structural differences. Roy's second claim is that localist representation and its symbolic, meaningful neurons are widespread throughout the brain. Up to now, even the strongest proponents of localist representation considered that the brain may use symbolic neurons only in some areas at certain levels of processing.

In regards to his first point, he explains that several misconceptions of the two models have led scientists to assume that they differ more than they actually do.

"The first misconception is that the property where 'each concept is represented by many units, and each unit represents many different concepts' is exclusive to distributed representation," he said. "I show that

that property is actually a property of the model that one builds, not of the units. A second misconception, which is partly related to the first, is that a localist unit should respond to one and only one concept. I show that that is not true either, that localist units can indeed respond to many different higher-level concepts. All these false notions haunt localist representation, and the first thing I did was show that they are false notions. And you can show them to be false only if you stick to the basic property of localist units, that they have 'meaning and interpretation on a stand-alone basis.'"

If Roy is correct, it would mean that many of the arguments used against localist representation – in particular, against University of Bristol Psychology Professor Jeff Bowers' "grandmother cell theory" – are invalid. (Put simply, grandmother cells are high-level concept neurons.) But perhaps more importantly, Roy's interpretation also means that any model built with distributed neurons can be built with localist neurons, since there is no structural difference. In other words, a model in which a neuron responds to multiple concepts can be either distributed or localist.

A neuron for everyone and everything

This interpretation clears the path to Roy's second claim, that the brain processes information using symbols, not subsymbolic connections. He explains that experimental support for symbol-based localist representation is robust, with some of the earliest evidence coming from studies of the visual system.

"There's more than four decades of research on receptive fields in the primary visual cortex and even in retinal ganglion cells that shows that the functionality of the cells in those regions can be interpreted," Roy said. "Researchers have found cells that detect orientation, edges, color, motion, and so on. David H. Hubel and Torsten Wiesel won the Nobel

Prize in physiology and medicine in 1981 for breaking this 'secret code' of the brain."

The discovery of these vision cells is just one piece of neurophysiological evidence suggesting that individual neuron cells have meaning and interpretation. Roy also cites several recent studies that have identified individual neurons in the hippocampus and the medial temporal lobe that represent specific objects or concepts and do not depend on the activity of other neurons. For example, in 2005, neuroscientists discovered that an epilepsy patient had one neuron cell that fired whenever a photo of Jennifer Aniston was presented. Various photos showing the blonde actress in different poses and from different angles all elicited a response from the same concept cell, a neuron in the hippocampus.

"Concept cells were also found in different regions of the medial temporal lobe," Roy said. "For example, a 'James Brolin cell' was found in the right hippocampus, a 'Venus Williams cell' was in the left hippocampus, a 'Marilyn Monroe cell' was in the left parahippocampal cortex and a 'Michael Jackson cell' was in the right amygdala."

Roy thinks that one of most supportive studies of his argument is the Cerf experiment from 2010. In this experiment, Moran Cerf, a neuroscientist at New York University and UCLA, asked epilepsy patients to look at several different images on a screen while the researchers attempted to identify one neuron in the medial temporal lobe that independently fired for each of the different images. One of the images was then randomly selected to become the target image, and patients were shown the target image at 50% visibility and a distractor image at 50% visibility and asked to focus their thoughts on the target image. The visibility of the target image increased when the firing rate of the previously identified target neuron increased compared to the firing rate of the distractor neuron. By focusing on the target images, the

patients could increase the target neuron's firing rate, with 69% of the patients succeeding in making the target image 100% visible.

In Roy's perspective, these results suggest that the neuron the researchers originally identified as the representative neuron for the target image was indeed a localist neuron. In other words, when that neuron fired, it had one specific meaning: the patient was thinking of the target image.

Roy emphasized that he did not look exclusively for studies to support his claim and ignore studies that contradicted it; he says he found no evidence that might contradict his claims.

"Although I have not exhaustively searched this literature, from what I looked at, there was not much to 'pick and choose' from," he said. "In the paper, I have cited some recent studies. And although I have not covered the universe of single cell studies on insects, animals, and humans, the ones I have looked at don't contradict my broad claim.

"There are some studies that show that a population of neurons has meaning," he acknowledged. "But that doesn't contradict my theory. For example, one can read the outputs of cells representing legs, ears, body, tail, and so on, and say that represents a cat. However, that doesn't contradict the claim that all of these cells have meaning and interpretation on a stand-alone basis, even though only when their outputs are combined can you say that it's a cat."

Future developments

All this evidence further solidifies Roy's impression that the brain is a system of symbols rather than a network of connections. If he's correct, then it would have implications for our understanding of the brain and future AI developments.

"The brain would need fewer connections with localist representation than with distributed representation," he said. "There is efficiency and filtering associated with localist representation. We can quickly filter out aspects of a scene without further processing. And that saves computations and energy consumed. Our brains would be exhausted if they didn't filter out irrelevant things quickly."

Applying the brain's symbolic representation to create AI systems may sound more straightforward than attempting to build AI systems using a subsymbolic mode, but it's far from simple.

"Localist representation may sound simplistic, but we are still struggling with the mathematics to replicate those functionalities, even for the visual system," Roy said. "So maybe it's not that simple."

Commentary on Roy's paper by David Plaut

David Plaut, Psychology Professor at Carnegie Mellon University, carries out research using the connectionist framework for computational modeling of brain functions. He has found issues with a few ideas in Roy's paper, starting with the fact that Roy frames the argument on neural representation differently than how it's usually framed.

"Asim's main argument is that what makes a neural representation localist is that the activation of a single neuron has meaning and interpretation on a stand-alone basis," Plaut said. "This claim is about how scientists interpret neural activity. It differs from the standard argument on neural representation, which is about how the system actually works, not whether we as scientists can make sense of a single neuron. These are two separate questions."

Plaut also thinks that Roy needs to clearly define what he means when he

says that a neuron has "meaning and interpretation."

"My problem is that his claim is a bit vacuous because he's never very clear about what a coherent 'meaning and interpretation' has to be like," he said. "He brings up some examples that he claims are supportive of neurons having meaning and interpretation, such as in the medial temporal lobe and hippocampal regions, but never lays out what would count as evidence against his claim. On his view, if we can't yet characterize the function of a neuron, it just means we haven't figured it out yet. There's no way to prove him wrong."

In fact, Plaut thinks that much of the experimental evidence that Roy cites as support for his view may not be as supportive as Roy claims.

"If you look at what he says 'meaning and interpretation' is supposed to be coding for, if you look into the examples he gives, they're not actually quite like that," Plaut said. "If you look at the hippocampal cells (the Jennifer Aniston neuron), the problem is that it's been demonstrated that the very same cell can respond to something else that's pretty different. For example, the same Jennifer Aniston cell responds to Lisa Kudrow, another actress on the TV show *Friends* with Aniston. Are we to believe that Lisa Kudrow and Jennifer Aniston are the same concept? Is this neuron a *Friends* TV show cell?"

He notes that there are other examples; for instance, there is one neuron that fires for both spiders and snakes, and another neuron that fires for both the Eiffel Tower and the Leaning Tower of Piza – somewhat related concepts, perhaps, but still with quite distinct meanings.

"Only a few experiments show the degree of selectivity and interpretability that he's talking about," Plaut said. "For example, Young and Yamane published a study in 1992 in which, out of 850 neurons, they found only one that had this high level of selectivity, while the other

cells had varying degrees of responses. If we ignore what the vast majority of what neurons are doing, it's selection bias. In some regions of the medial temporal lobe and hippocampus, there seem to be fairly highly selective responses, but the notion that most cells respond to one concept that is interpretable isn't supported by the data."

Commentary on Roy's paper by James McClelland

As mentioned above, one of the papers that Roy cites is coauthored by James McClelland, a psychology professor at Stanford University whose work has played a pivotal role in developing the connectionist framework. In response to Roy's paper, McClelland explained why he still favors the distributed representation model:

"Roy's paper lays out his claim that the brain uses localist representation – the view that individual neurons in the brain have 'meaning and interpretation' on a stand-alone basis – and contrasts this with the distributed representation view – the view that each neuron participates in many representations, and that it is therefore not possible to determine what concept is being represented by looking at the activity of a single neuron. Although my collaboration with David Rumelhart exploring neural networks began with the exploration of localist models (McClelland & Rumelhart, 1981), we soon became convinced that the localist view is unlikely to be correct (McClelland & Rumelhart, 1985). Here I briefly explain why I still hold the distributed representation view.

"One problem with localist representation is the question, when to start and when to stop using a localist representation. Suppose I encounter a new kind of bread – one baked in thin sheets with sesame and cardamom seeds. In order to understand that this new kind of bread might smell or taste like, I would likely rely on representations of other kinds of bread and of sesame and cardamom seeds, and also on my knowledge of other kinds of foods in thin sheets that I may know about. I already have a

great deal of knowledge about this thin bread, having never encountered it before. Did I already have a localist representation for it, or did I compose my understanding of it out of knowledge I had previously acquired for other things? If the latter, what basis do I have for thinking that the representation I have for any concept – even a very familiar one – as associated with a single neuron, or even a set of neurons dedicated only to that concept?

"A further problem arises when we note that I may have useful knowledge of many different instances of every concept I know – for example, the particular type of chicken I purchased yesterday evening at the supermarket, and the particular type of avocados I found to put in my salad. Each of these is a class of objects, a class for which we may need a representation if we were to encounter a member of the class again. Is each such class represented by a localist representation in the brain? The same problem arises with specific individuals, since we know each individual in many different roles and phases. Do I have a localist representation for each phase of every individual that I know? Given these questions, my work since the 1985 paper has focused on understanding how the brain may use what it has learned about many different and partially related experiences, without relying exclusively on localist representations. On this view, the knowledge arising from an experience is the set of adjustments made to connection weights among participating neurons – neurons that participate in representing many different things.

"Roy lays out several lines of argument in support of his point of view. Perhaps the central argument is that recordings from neurons show that the neurons in some parts of the brain have what some might consider to be surprisingly specific responses. Let us discuss one such neuron – the neuron that fires substantially more when an individual sees either the Eiffel Tower or the Leaning Tower of Pisa than when he sees other objects. Does this neuron 'have meaning and interpretation independent

of other neurons'? It can have meaning for an external observer, who knows the results of the experiment – but exactly what meaning should we say it has? An even harder question is, what meaning does the neuron have for the individual in whose brain it has been found? Let's take the simpler question first.

"First, for the external observer: it should be apparent that the full range of test stimuli used affects what meaning we assign to such a neuron. The Japanese neuroscientist Keiji Tanaka found neurons in monkeys' brains that others had called 'monkey paw detectors' and others they might have called 'cheshire cat detectors,' but he then constructed many special test stimuli to use in testing each neuron. He found that the neurons generally responded even better to schematic stimuli that were not recognizably paws or cats but had features in common with them. Such neurons surely participate in representing cats or paws but may also participate in representing other objects with similar shape features. Critically, however, the response of the neuron is difficult to pin down in simple verbal terms and neighboring neurons have similar responses that shade continuously from one combination of features to another. Is the same true of the Eiffel Tower/Leaning Tower of Pisa neuron? In the context of these observations, the Cerf experiment considered by Roy may not be as impressive. A neuron can respond to one of four different things without really having a meaning and interpretation equivalent to any one of these items.

"Second, to the individual in whose brain the neuron has been found: Roy's analysis ignores the question of how a neuron assigned to represent a concept is then used by the observer to mediate use of the observer's knowledge of the concept. This is the issue my colleagues and I have sought to explore with explicit models that rely on distributed representations over populations of simulated neuron-like processing units. While we sometimes (Kumeran & McClelland, 2012, as in McClelland & Rumelhart, 1981) use localist units in our simulation

models, it is not the neurons, but their interconnections with other neurons, that gives them meaning and interpretation. The sight of a picture of Saddam Hussein brings to mind heinous crimes against the citizens of Iraq and Kuwait, not because a particular neuron is activated but because it (and many other neurons) participates in activating other neurons that are involved in the representation of other heinous crimes and/or in verbal expressions and imagined scenes involving such crimes. And it participates in activating these other neurons because of its connections to these neurons. Again we come back to the patterns of interconnections as the seat of knowledge, the basis on which one or more neurons in the brain can have meaning and interpretation.

"In our work we have proposed that different parts of the brain rely on representations that differ in their relative specificity (McClelland et al, 1995; Goddard & McClelland, 1996). The Medial Temporal Lobes are thought to represent items, locations, events, and situations in terms of sparse patterns of activation, but even here each neuron is thought of as participating in many representations. Even here, the principles of distributed representation apply: the same place cell can represent very different places in different environments, for example, and two place cells that represent overlapping places in one environment can represent completely non-overlapping places in other environments. Other parts of the neocortex of the brain are thought to rely on denser distributed representations, where a somewhat larger overall fraction of the neurons are activated by a particular item, location, etc. There is a lot more to understand about these representations. Studies involving very small numbers of neurons may be misleading in this regard. Progress will depend on recording from large numbers of neurons, so that we can more readily visualize the activity across the entire population."

Roy has responded to Plaut's and McClelland's comments here.

More information: Roy, A. "A theory of the brain: localist

representation is used widely in the brain." *Frontiers in Cognitive Science*.

Kumaran, D. & McClelland, J. L. (2012). "Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system." *Psychological Review*, 119, 573-616. DOI: 10.1037/a0028681

Rogers, T. T. & McClelland, J. L. (2004). Semantic Cognition: A Parallel Distributed Processing Approach. Cambridge, MA: MIT Press

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). "Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory." *Psychological Review*, 102, 419-457

McClelland, J. L. & Rumelhart, D. E. (1985). "Distributed memory and the representation of general and specific information." *Journal of Experimental Psychology: General*, 114, 159-197

McClelland, J. L. & Rumelhart, D. E. (1981). "An interactive activation model of context effects in letter perception: Part 1. An account of Basic Findings." *Psychological Review*, 88, 375-407