Artificial Intelligence and the "Barrier" of Meaning Workshop

Questions from Participants

Alan Mackworth: Humans use language, diagrams and sketches to communicate meaning. How should agents represent, learn and use the knowledge required to create and interpret diagrams? I have proposed using constraint satisfaction ["On Reading Sketch Maps", 1977] and FOL reduced to model-finding in propositional logic ["A Logical Framework for Depiction and Image Interpretation", with Ray Reiter, 1990]. These approaches are partially adequate for addressing the issues of knowledge representation and use but do not address the learning issue. On the other hand, contemporary approaches to machine learning do not seem useful either. Could agents learn the knowledge required to create and interpret meaningful diagrams?

Alison Gopnik: We know that young children are the best learners in the world, and in particular infer abstract concepts and structured meanings from concrete, probabilistic and apparently chaotic data patterns. Moreover, they modify and revise those concepts based on new evidence in radical ways. How is that possible? How do children go from the particular sensory data at their eyes and ears to the abstract causal structures involved in intuitive theories? How can they revise those concepts in ways that generate new concepts? How do they use them as basis for language?

Barbara Grosz: What level of recognition of intentions — of people and other agents with whom they communicate or work — do AI systems need to derive meaning and to act appropriately in the situations in which they find themselves?

Brenden Lake: Why are people smarter than machines?

Bruno Olshausen: What does it mean to 'perceive'? There are by now numerous physiological, psychological, and modeling studies of perceptual processes. Typically these studies frame perception as an input-output processes - i.e., a neuron computes some function of the input and passes that on to another stage of processing, or an observer makes a yes/no or same/different judgement about a stimulus. But perception is something more profound than this. Observe for example the image below - at first glance, without knowing what it is, one sees only a collection of amorphous black and white splotches with no clear meaning. But with a subtle hint (look for a cow) the pieces begin to fall into place and form a coherent whole, with immense amounts of detail about the 3D shape and texture of the object, as well as the photographic process. This is perception. It is a rich and entirely subjective experience, far beyond what can be communicated verbally through the linguistic bottleneck. This makes it difficult to approach scientifically, or to capture computationally as is currently being attempted with multilayer neural networks. The current models essentially "dumb down" perception in order to make it a tractable problem, but end up throwing out the baby with the bath water in the process. I would like to discuss what would constitute a good model of perception - i.e., what are the functional and phenomenological requirements of a theory of perception? Examining this question in a more forthright manner is important for making progress both scientifically and technologically.

Christopher Mole: A question that interests me is whether we should distinguish between 'crashing the barrier of meaning' (in Rota's phrase), and surmounting that barrier cleanly. We know that Artificially Intelligent systems can now accomplish tasks in which the criteria for success are meaning-involving (tasks such as picture captioning, and translation between natural languages). Perhaps these accomplishments show that the barrier of meaning has been crashed, but — since the methods by which these accomplishments have been made involve the brute force of data — these do seem to be cases of crashing that barrier, rather than surmounting it. There is a widely-held suspicion that these accomplishments make no contribution to our understanding of how intelligent agents are able to respond to meanings, as such. My question is whether such suspicions concerning the explanatory import of data-driven methods for dealing with meaning are merely the result of a prejudice, or whether they can be put on a rationally respectable footing.

Chris Wood: Claude Shannon's landmark 1948 paper, "A Mathematical Theory of Communication", took the radical step of defining information in a manner that completely eliminated meaning from the definition. Focused primarily on the engineering problem of designing effective communication devices, Shannon defined information in terms of the signal actually sent relative to the ensemble of possible signals that could have been sent, completely without regard to the meaning of the signal or its alternatives. This separation of information from meaning, of signals from semantics, has allowed "Shannon Information" to become the dominant mathematical and scientific characterization of information, extensively applied not only to engineering problems, but to a wide variety of phenomena in the physical, biological, and social sciences.

My question is "can we adopt an analogous strategy for meaning to that Shannon employed so successfully for information?". The key elements of Shannon's strategy were to define "information" in a manner that was: (a) independent of the nature of the signals involved (plaintext or code written on a piece of paper, telephone or telegraph signals, or today's SSL encrypted internet traffic, et al.); (b) independent of any semantic meaning that may be associated with those signals; and (c) independent of any assumptions about the nature of the sender and receiver of the signals apart from their ability to send and receive them as required by the theory.

So, analogously, can we strip "meaning" down to its essence and define it in such a way that is: (a) independent of the specific system and signals involved; (b) independent of any linguistic content of the signals in question; (c) that depends only (or at least primarily) upon the logical relationships among inputs, system, and outputs; and (d) with all that paring away is still more than "just" information?

David Krakauer: In 1953 in his Philosophical Investigations Wittgenstein penned one of his more pessimistic, anthropocentric aphorisms, "If a lion could talk, we could not understand him" . One assumes that his "meaning" (always enigmatic) would be no less true if he had chosen a flea or a platypus. What Wittgenstein meant was presumably what Searle meant by the Chinese Room Argument. We might call Wittgenstein's phylogenetic prequel, the Vienna Zoo Argument. Independently, as is so often the case in the pursuit of meaning, Jorge Luis Borges had invented the Buenos Aires Zoo Argument in his poem, The Other Tiger: "It strikes me now as evening fills my soul/That the tiger addressed in my poem/Is a shadowy beast, a tiger of symbols". Somewhat earlier, Da Vinci writing in his encrypted journals made the uncomfortable observation that his "Mentula" had "a mind of his own". Something we now think about in terms of the neural firewall separating the voluntary and involuntary nervous system. Let's call this the Vitruvian Man Argument.

From phylogenetic to ontogenetic rule systems —across species, within species, and within individuals — there are barriers to accessing meaning. These are all instances of "Type systems" related to type theory, which restrict operations to given variable types in order to minimize errors. That is, there are distinct rule systems that are not fully inter-translatable operating in adaptive systems that restrict access across all domains that engage in information processing. Hence Chinese rooms, Viennese Zoos, and autonomic nervous systems are necessary features of intelligent systems. These barriers are more like Dykes and Levees and less like international borders and walls — they "Screen off" computations in order to maximize the benefits of modularity.

David Wolpert: I suggest that the whole issue of "assigning meaning" to the decisions of ML algorithms is in large part a category error, without any more substance than the problem of "assigning meaning" to the actions of a jet engine, or of any other technology that we the users of the technology may not understand.

More precisely, I suggest that the primary reason we currently care about this issue at all is the unease that the members of the public have with new technologies based on ML. However, Joe and Jill Q. Public are always uneasy with any new technology to which they are being asked to delegate some control over their physical well-being. That was just as true with the introduction of the steam-driven train and propeller-driven airplane, as it might be with things like autonomous vehicles, ML-driven medical diagnoses, or other ML-based technologies.

The uneasiness with those earlier technologies was due in part to not being able to "assign meaning to the actions of those devices", i.e., due to lack of understanding of those devices. And just as Joe and Jill quickly got over whatever uneasiness they may have had about those earlier new technologies, so Joe and Jill would (I predict) quickly get over their uneasiness with new technologies rooted in machine learning.

The only exceptions that I could imagine, and that would be quite rare if they occur at all, would be regulators who might demand that machine learning experts themselves can explain the
decisions being made by machine learning algorithms in scenarios that affect human lives. (An analogy with old technologies is the FAA demanding that aerospace engineers can do V&V on their designs.)
The public will not care about understanding these technologies any more than they currently care about understanding the technologies in a jet engine.

In short, machine learning very quickly will be seen as just a tool in an engineer's work box, without even any need for that engineer who is using this tool to understand precisely how it works (never mind the public understanding how it works).

Dawn Song: How do we frame the question of learning to incorporate more naturally the notion of generalization?

Dileep George: How are concepts represented, acquired and inferred? How can we build concrete computational models for ideas like perceptual symbol systems, image schemas and mental imagery that have been discussed extensively in cognitive science? Is there a way to combine these with ideas from probabilistic language of thought to learn grounded concepts? How can we bring grounded concepts in contact with language? In what way can we measure and share progress on these questions?

Douwe Kiela: Can artificial agents ever hope to learn to understand meaning passively, or is language learning inherently active?

Fernando Pereira: Meaning comes from active experience of the real world, not from translation from a (natural) language to a (formal) one.

Garrett Kenyon: Is predicting the future a sufficient criteria for acquiring an understanding of the world?

Irene Pepperberg: Nonhumans clearly can engage in abstract thinking and symbolic representation; sometimes their levels of understanding and meaning are easily transferred across domains, but at other times can be constrained by the context of their learning. We know some techniques that work to expand and improve nonhumans' abilities, but what additional techniques might exist and could such findings be used to inform aspects of AI?

Jessica Flack: My view is that understanding in science comes through triangulation using multiple representations of the world---typically, the data, mathematics, and natural language. In principle this triangulation can occur within a single mind or collectively but in practice the collective aspect seems to be central. Does this characterization miss any essential features?

Josh Bongard: Can a current-day deep learner (or a future one with arbitrarily more computational power and data resources to draw on) "understand" external phenomena; can those phenomena "mean" anything to the learner? If not, what other branches of AI and cognitive science investigation --- embodiment, predictive coding, free energy principles, integrated information theory, theory of mind -- may endow future machines with subjectivity, and thus understanding and meaning, and if so, how?

Julia Hirschberg: How can we understand cultural bias in AI? Does knowing why an algorithm is biased fix the problem?

Melanie Mitchell: I have two questions: (1) Human understanding in vision, speech recognition, language, and so on seems to require vast amounts of "commonsense" knowledge, as well as the abilities to form new concepts via abstraction and analogy. Can AI systems be as reliable and robust to error and adversarial attack as humans without such knowledge and abilities? (2) Is "superintelligent" AI possible, even in principle? Or will the "limitations" of human thinking (e.g., speed of reading and listening, slowness at arithmetic, cognitive biases, lack of "rationality", need for sleep) be inevitable emergent properties of any human-level intelligence?

Melanie Moses: Does meaning exist only in human interpretation of language and symbols? In animal communication, particularly social animals, is there meaning in communication or only probabilistic responses to cues? How much mental processing is necessary to associate meaning with cues and symbols?

Michael Strevens: What is the difference between merely representing some aspect of the world, as a thermostat represents temperature, and truly understanding what it is that you are representing?

Mirta Galesic: I am interested in how human groups establish and share meaning of historical and current political events, and their own reactions to these events. Is accurate understanding of events and of the underlying complexities necessary for good group and individual performance? Can groups function well without understanding why they react to events in a particular way? Are there any implications of the level of understanding achieved in human social systems for the level of understanding that we should expect from AI systems?

Percy Liang: In the last decade, we've seen AI systems obtaining human parity across a wide array of benchmarks (e.g., image classification on ImageNet, question answering on SQuAD). But it's clear that these systems don't truly "understand" anything or generalize to genuinely novel situations with the nimbleness of humans. Then the question is how we should evaluate AI systems to test true understanding?

Rod Brooks: This is meta, about us humans as researchers. If we take Lakoff & Johnson seriously, then much of our thinking is based on using metaphors of place, state and change of state at place, movement through time from place to place, and, overall, a countable world. That mode of thinking fails us in understanding quantum mechanics. It, however, is certainly the basis of metaphor that Turing used for computation. Are we blinded by such metaphors in our research definitions of meaning and understanding?

Tom Mitchell: How is it possible that an artificial neural network, which is essentially propositional, or the human brain which is also built of neurons, can represent relational and first order knowledge? Interestingly, we now have computer programs that successfully predict fMRI neural representations of individual word meanings from their corpus statistics, but we still do not understand how the human brain represents the difference between "Mary hit Sue." and "Sue hit Mary."

Yarden Katz: How can we develop a notion of scientific understanding that incorporates the scientist? In other words, can we sketch an account of understanding that respects the fact that scientists are historically situated and embodied, and that this is inextricably linked to scientific theorizing (and consequently, to what might be considered "understanding")?