On Crashing the Barrier of Meaning in Artificial Intelligence

Melanie Mitchell

■ *In 1986, the mathematician and philosopher Gian-Carlo Rota wrote, "I wonder whether or when artificial intelligence will ever crash the barrier of meaning" (Rota 1986). Here, the phrase "barrier of meaning" refers to a belief about humans versus machines: Humans are able to actually understand the situations they encounter, whereas even the most advanced of today's artificial intelligence systems do not yet have a humanlike understanding of the concepts that we are trying to teach them. This lack of understanding may underlie current limitations on the generality and reliability of modern artificial intelligence systems. In October 2018, the Santa Fe Institute held a three-day workshop, organized by Barbara Grosz, Dawn Song, and myself, called Artificial Intelligence and the Barrier of Meaning. Thirty participants from a diverse set of disciplines — artificial intelligence, robotics, cognitive and developmental psychology, animal behavior, information theory, and philosophy, among others — met to discuss questions related to the notion of understanding in living systems and the prospect for such understanding in machines. In the hope that the results of the workshop will be useful to the broader community, this article summarizes the main themes of discussion and highlights some of the ideas developed at the workshop.*

In 1986, the mathematician and philosopher Gian-Carlo Rota wrote, "I wonder whether or when artificial intelligence will ever crash the barrier of meaning" (Rota 1986). Here, the phrase "barrier of meaning" refers to a belief about humans versus machines: Humans are able to actually understand the situations they encounter, whereas even the most advanced of today's artificial intelligence (AI) systems do not yet have a humanlike understanding of the concepts that we are trying to teach them. That is, the internal representations learned by (or programmed into) AI systems do not capture the rich meanings that humans bring to bear in perception, language, and reasoning.

This lack of understanding may underlie current limitations on the generality and reliability of modern AI systems. While deep neural networks, trained via supervised or reinforcement learning, perform remarkably well on many problems in computer vision, natural language processing, and other domains central to AI, these systems remain brittle compared with human intelligence. Even the most successful deep networks can fail in unexpected ways when faced with inputs that differ, even in small degrees, from their training regime. Moreover, such networks struggle in making conceptual abstractions, and are vulnerable to adversarial attacks that do not affect humans. Researchers are still debating whether such limitations can be overcome with more data, or with additional network layers, or whether something more fundamental is missing.

For decades, the AI community has scrutinized questions concerning machine understanding, exploring related ideas such as strong versus weak AI, symbol grounding, and the general area of common-sense knowledge. Questions about the definition — and necessity — of humanlike understanding in machines have become ever more centrally important as a result of the recent successes and broad real-world deployment of deep learning systems.

In October 2018, the Santa Fe Institute held a three-day workshop, organized by myself, Barbara Grosz, and Dawn Song, called Artificial Intelligence and the Barrier of Meaning. To spur discussion, the following questions were given to the participants ahead of the workshop: By what mechanisms do humans and other natural information-driven systems extract meaning from data or experience? Can insights from such systems be used to improve AI? To what extent do current-day AI systems need to understand the situations they deal with to perform reliably, particularly in situations outside their training regimes? To what extent do systems need to understand in order to be able to explain their decisions and predictions? Does a lack of understanding make data-driven AI systems (for example, deep networks) susceptible to adversarial examples? Is there a way to defend against such attacks without imbuing such systems with humanlike understanding? How do we determine if a system is actually understanding?

Thirty participants from a diverse set of disciplines — AI, robotics, cognitive and developmental psychology, animal behavior, information theory, and philosophy, among others — met to discuss these and related questions (a list of the participants is given in figure 1). The workshop combined short talks with extensive small and large group discussions. In the hope that the results of the workshop will be useful to the broader community, this article summarizes the main themes of discussion and highlights some of the ideas developed at the workshop.

Correlates of Understanding

Is it true that existing AI systems lack humanlike understanding and face a barrier of meaning that limits their generality and robustness? While all of the workshop participants agreed with the intuitions behind such claims, the terms *understanding* and *meaning* are ill-defined. Marvin Minsky called such mental terms “suitcase words,” ones that are packed to the breaking point with different meanings (Minsky 2006).

Rather than proposing specific definitions, the workshop participants collectively listed a set of correlates of understanding in humans and other living systems, and discussed how these correlates contrast with today’s predominant AI systems. The following list of such correlates (and other relevant points of discussion) attempts to capture the flavor of the brainstorming discussion at the meeting.

Core Knowledge

Understanding is built on a foundation of innate core knowledge. Unlike most current AI systems, humans and other animals seem to come into the world with (or develop very early on) a healthy dose of intuitive physics: how objects behave individually, how they interact with other objects, and the possible effects of such interactions. Moreover, humans and animals seem to possess an innate metaphysics: They are born with (or develop early on) the very notions of discrete objects, relationships, events, and indeed causality itself. Humans, being thoroughly social organisms, also have an innate or early-developed intuitive psychology, one that includes concepts of what communication is for, and a basic theory of mind for other humans. The extent to which nonhuman animals have something like a theory of mind is unclear, but communication is also either innate or developed early in many animals. In humans, and to at least some extent in other animals, such core knowledge forms the foundation for future understanding, inference, and common sense. What form such inductive biases take in various living systems, how such biases are represented in the brain and body, and how they emerge and develop in early life are fundamental and largely open questions in neuroscience and cognitive science.

Abstraction and Generativity

Supervised machine learning (ML) typically focuses on training and test data coming from the same distribution. This requires function-fitting interpolation. In contrast, humans and most other animals are able to extrapolate — that is, to adapt what they have learned to diverse situations. This is accomplished via the abilities to build abstract representations, and to make analogies mapping these representations to new situations. Abstract representations and analogy, combined with core knowledge, allow organisms to learn concepts from a small number of examples, to imitate and generate behavior at a conceptual

Gary Bengier, *Bengier Foundation*
Joshua Bongard, *University of Vermont*
Rodney Brooks, *Robust.AI*
Jessica Flack, *Santa Fe Institute*
Mirta Galesic, *Santa Fe Institute*
Dileep George, *Vicarious*
Alison Gopnik, *University of California, Berkeley*
Barbara Grosz, *Harvard University and Santa Fe Institute*
Julia Hirschberg, *Columbia University*
Jürgen Jost, *Max Planck Institute for Mathematics in the Sciences and Santa Fe Institute*
Yarden Katz, *Harvard University*
Garrett Kenyon, *Los Alamos National Laboratory*
Douwe Kiela, *Facebook AI Research*
David Krakauer, *Santa Fe Institute*
Brenden Lake, *New York University*
Percy Liang, *Stanford University*
Alan Mackworth, *University of British Columbia*
Melanie Mitchell, *Portland State University and Santa Fe Institute*
Tom Mitchell, *Carnegie Mellon University*
Christopher Mole, *University of British Columbia*
Cris Moore, *Santa Fe Institute*
Melanie Moses, *University of New Mexico*
Bruno Olshausen, *University of California, Berkeley*
Irene Pepperberg, *Harvard University*
Fernando Pereira, *Google*
Bart Selman, *Cornell University*
Cosma Shalizi, *Carnegie Mellon University and Santa Fe Institute*
Michael Strevens, *New York University*
David Wolpert, *Santa Fe Institute*
Chris Wood, *Santa Fe Institute*

Figure 1. Workshop Participants.

level, to transfer knowledge between modalities, to perform flexible planning, and to generate possible futures and counterfactuals, among other abilities central to our notion of understanding. One workshop participant hypothesized as follows:

> How could we tell if a machine could understand? If the machine can perform an action when asked (for example, jump when told "jump!"); recognize the concept enacted by others (point to who or what is jumping); enact the command with an alternate motor system (for example, use two fingers to demonstrate jumping); and conform to the concept's tacit physical context (for example, jump without hitting the ceiling) as well as its social context (for example, jump gently near humans).

Active Perception, Learning, and Inference

Several workshop participants contrasted the passive, feedforward, and supervised nature of current learning and inference in neural networks with the importance of active mental processes in natural intelligent systems. Here, *active* means that the system itself dynamically seeks out information and continually uses the information it finds — along with prior knowledge or biases — to help direct further information-seeking. Perception, learning, and inference are active processes that unfold dynamically over time, involve continual feedback from context and prior knowledge, and are largely unsupervised.

Object-Based, Causal Models

In contrast with models that solely perform classification or action selection, understanding involves building causal models of objects, relationships, actions, and entire situations, and flexibly using these models to predict and act in the world. Here, the term *object* refers to any discrete conceptual entity, and causal implies that a model captures spatio-temporal relationships of causality among parts of a situation. Such models are built on top of the core knowledge described above.

Metacognition

Understanding seems to require not only causal models of the world, but causal models of our own thinking. The ability to model one's own thinking processes is known as *metacognition*, and allows us to explain and predict our own thought processes and decisions, and map them onto the thought processes of others. Moreover, metacognition is what makes active perception, learning, and inference possible in that it allows an organism to know that it needs additional information to solve a problem; metacognition guides the type of information that is sought and where (or from whom) it must be obtained.

Embodiment

The embodied-cognition hypothesis states that understanding in living systems arises not from an isolated brain but rather from the inseparable combination of brain and body interacting in the world. Supporters of this hypothesis argue that a disembodied brain (analogous to most of today's AI systems) cannot achieve humanlike (or animal-like) understanding. This hypothesis has long been debated in many fields, but over the last decades evidence has emerged from neuroscience, psychology, and linguistics that supports the essential role of the body in virtually all aspects of thinking. In neuroscience, such evidence includes the discovery of mirror neurons as well as the surprisingly extensive connections between motor and cognitive areas in the brain in both humans and nonhumans. In developmental psychology it appears that cognitive development is often triggered by a child's developing motor skills (for example, a child develops understanding about other people's goals only when the child herself is able to start reaching for objects; Robson and Kuhlmeier 2016). Cognitive psychology has provided evidence that even abstract concepts are understood via mental simulations of physical actions (Barsalou 1999). In linguistics, the analysis of metaphors has indicated that abstract concepts are often understood via physical metaphors (for example, social interaction is grounded in the perception of temperature: "she greeted me warmly" or "he gave me the cold shoulder"; Lakoff and Johnson 1980; Williams and Bargh 2008). These are just a few examples of interdisciplinary evidence for embodied cognition. However, these ideas remain controversial and the notion of embodiment needs further clarification and refinement.

Evolutionary Considerations

In our workshop discussions, we focused as well on the evolution of neural structures and perceptual capabilities that make embodied understanding possible. Embodiment itself is not sufficient for understanding: while Autonomous robots routinely perceive features of their environment and integrate this information in neural networks to achieve a specified goal, autonomous cars and vacuum cleaners have not yet achieved humanlike understanding. A shared brain morphology and organization gives humans, and to some extent other animals, a common structure to translate signals perceived about the external environment into an internal representation that appears essential to understanding. As one example, there is evidence that an evolved set of neural circuits underlie human and animal intuitive understanding of numbers. The way the brain encodes numbers may explain why the number line is such an easily grasped metaphor (Dehaene 2011).

If a common internal organizational structure for information representation and processing is at the core of understanding, can AI surmount the barrier of meaning without sharing this underlying neural architecture and its evolutionary history? Perhaps the particular underlying structure of the brain is not as central to understanding as the evolutionary process itself. In evolutionary robotics, both the structure of the internal representation and the embodied agent's responses to the environment emerge from repeated interactions in that environment (Hecker and Moses 2015; Nolfi et al. 2016). Such an evolutionary approach may be a path toward understanding even if that understanding is encoded in structures very different than the human brain.

Sufficiency of the Information Processing Metaphor

In our workshop, discussion of the embodiment hypothesis led to a related discussion about the sufficiency of the information processing metaphor. In AI, and in much of cognitive science, intelligence is typically framed as information processing. The idea is that cognition is a form of computation — a set of operations on inputs (for example, perceptions) that gives rise to outputs (for example, motor activities). At our meeting, the discussion around embodiment gave rise to questions about whether this framing of intelligence is sufficient to explain and capture the notion of *understanding*. Meeting participants gave examples of *Caenorhabditis elegans*, jumping spiders, and other simple creatures for which extensive data exists about the brain, but we still don't understand how brain processes give rise to behavior. It may be that the pure information processing metaphor is not sufficient, and that other frameworks might give the insights that we need. Some proposals for such frameworks include free energy principles, control theory, dynamical systems theory, and ideas from biological development. It may be that an entirely new framework is needed for a full account of cognition.

Are We Misframing the Learning Problem?

Beyond the correlates of understanding, a major discussion topic of the workshop was whether, if the goal is to create understanding in machines, the current framing of how learning takes place in AI is misguided. Supervised machine learning (ML) is often framed in terms of distributions over the training and test data. In fact, the theoretical basis for much of ML requires that training and test examples are independently and identically distributed. In contrast, human learning — and teaching — is active, sensitive to context, driven by top-down expectations, and transferable among highly diverse tasks, whose instances may be far from independently and identically distributed. Moreover, some workshop participants argued that human learning focuses intentionally and preferentially on non-independently and identically distributed samples. A developmental psychologist at our meeting gave the example of *Motherese*: the language samples that mothers (unconsciously) target to their babies. Studies have shown that Motherese does not consist of independently distributed samples of phonemes, but rather pushes extremes — phonemes that are close to the phonetic boundary. Babies seem to selectively pay more attention to these edge cases than to normal language, and to readily generalize from them.

Going further, modern AI systems often focus on the optimization of a cost function. It's unclear what should be optimized to achieve the kinds of correlates of understanding described in the previous section, or even if optimization itself is the right framework to be using.

Another topic of discussion at the workshop was the relatively short and narrow life experienced by ML systems. ML systems are typically trained on narrow problems, using highly restricted datasets that are not necessarily ecologically relevant for developing understanding. Moreover, most research ML systems are short-lived; they are created for a particular set of experiments (or a paper) and then disappear. Unlike living intelligent systems, these ML programs do not experience pressure to develop abstract, transferrable representations. Indeed, it may not be possible to develop humanlike abstract representations without the kind of developmental trajectory that human infants experience. Alan Turing himself proposed something similar: "Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's?" (Turing 1950). Indeed, creating a program with the commonsense abilities of an 18-month–old baby is currently the focus of a multiyear Defense Advanced Research Projects Agency effort (Turek 2018).

Benchmark Datasets as Drivers of Research in AI

Modern AI research, particularly in the field of ML, often focuses on benchmark datasets. Examples of widely used benchmarks include ImageNet (Deng et al. 2009) and Microsoft COCO (Lin et al. 2014) for object recognition, the Stanford Question Answering Dataset (Rajpurkar et al. 2016) for question-answering, and the Workshop on Statistical Machine Translation datasets for machine translation (Luong and Manning 2016). Several discussions at our workshop focused on the role of such widely used benchmarks in promoting and testing systems for visual and language understanding.

Several workshop participants argued that, while such benchmark datasets have done a great service in pushing the field forward, there are some downsides to the strong focus on such datasets. Due to the incentives the field puts on successful performance on specific benchmarks, sometimes research becomes too focused on a particular benchmark rather than the more general underlying task. For example, while performance of deep neural networks on various ImageNet tasks approaches human level performance, the more general task of object detection and visual recognition more generally remains far below the level of humans. Many of the articles published using ImageNet focused on incremental improvement on the all-important state of the art rather than giving any insight into what these networks were actually recognizing or how robust they were. Moreover, work that explores novel, interesting ideas but does not meet the state of the art is often hard to publish at top conferences. In short, benchmarks can have the effect of pushing the research community into an *exploit* rather than an *explore* mode of research.

Another example is the widely-used Stanford Question Answering Dataset benchmark (Rajpurkar et al. 2016) for natural-language question-answering. While AI systems quickly achieved superhuman performance on this dataset, the more general task of question-answering remains very challenging for machines. It seems that there are exploitable biases that allow systems to use what one workshop participant called "cheap tricks" to perform well on the Stanford Question Answering Dataset (and related natural language processing datasets) without actually understanding the text in the way human readers do.

Another striking example of the existence of exploitable biases is on the Winograd Schema dataset (Levesque et al. 2011). This task and dataset were explicitly constructed to avoid such biases; in proposing this task, the authors write: "We want multiple-choice questions that people can answer easily. But we also want to avoid as much as possible questions that can be answered using cheap tricks (aka heuristics)." However, it seems that this dataset does contain subtle exploitable biases that allow statistical models to do well without what we generally think of as understanding (Sakaguchi et al. 2019). One workshop participant commented that "in any set challenge there are likely to be cheap tricks that simple algorithms can exploit."

How should we design benchmark datasets that promote deeper machine understanding? Some workshop participants argued that benchmark datasets should be relatively small, and that projects should focus on testing on many independently created datasets rather than a single benchmark. Others argued that it is important that any given benchmark doesn't stay around too long, new benchmarks should be created constantly, and previously published systems should be evaluated on these new benchmarks rather than just put to rest. Others argued that AI research should not focus on benchmarks at all, and that the community should be challenged to think about how research might be carried out without competitions on benchmarks.

Conclusions

The discussions at this workshop were an attempt to make sense of understanding in both living systems and in machines. Understanding is an ill-defined quality that seems to be a fundamental part of the robust, general intelligence we see in humans and other thinking systems. Our limited conception of what understanding actually involves makes it hard to answer basic questions: How do we know if a system is actually understanding? What metrics can we use? Could machines be said to understand differently from humans? 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? One workshop participant commented: "In my opinion, the obligation is upon those who believe that understanding is some unified, generalized process to show how it is such in human cognitive and neuroscience data."

This echoes the decades-old thoughts of AI pioneer Marvin Minsky:

> Though prescientific idea germs like "believe," "know," and "mean" are useful in daily life, they seem technically too coarse to support powerful theories. . . . Real as "self" or "understand" may seem to us today . . . they are only first steps towards better concepts.

Minsky went on, pointing out that our confusions about these notions

> . . . stem from a burden of traditional ideas inadequate to this tremendously difficult enterprise. . . . [T]his is still a formative period for our ideas about mind. (Minsky 1980)

Of course it is not unusual for ill-defined concepts to be central in science: Think, for example, of *gene* in biology or *force* in physics, both of which are still in the process of being fully understood. New ideas at the boundaries of neuroscience, cognitive science, and AI may allow us to make further scientific progress on the inadequate ideas Minsky describes. While our workshop discussions reminded us of how far there is to go to understand understanding, it also made clear the importance of interdisciplinary collaboration to overcome the barrier of meaning.

Acknowledgments

Many thanks to all the workshop participants for their contributions to the ideas and discussion described in this article. Additional thanks to the following people for their comments and contributions to this document: Barbara Grosz, Percy Liang, Tom Mitchell, Christopher Mole, Melanie Moses, Irene Pepperberg, Michael Strevens, and Chris Wood. The workshop described here was made possible by generous funding from the Santa Fe Institute, the *Journal of Artificial Intelligence*, and the National Science Foundation (Division of Information and Intelligent Systems Award 1832717). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Barsalou, L. W. 1999. Perceptual Symbol Systems. *Behavioral and Brain Sciences* 22(4): 577–660.

Dehaene, S. 2011. *The Number Sense: How the Mind Creates Mathematics*. Oxford, UK: Oxford University Press.

Deng, J.; Dong, W.; Socher, R.; Li, L. J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the Institute of Electrical and Electronics Engineers (IEEE) Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, 248–55. Piscataway, New Jersey: IEEE. doi.org/10.1109/CVPR.2009.5206848.

Hecker, J. P., and Moses, M. E. 2015. Beyond Pheromones: Evolving Error-Tolerant, Flexible, and Scalable Ant-Inspired Robot Swarms. *Swarm Intelligence* 9(1): 43–70.

Lakoff, G., and Johnson, M. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.

Levesque, H. J.; Davis, E.; and Morgenstern, L. 2011. The Winograd Schema Challenge. In *Logical Formalizations of Commonsense Reasoning: Papers from the 2011 Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposium*. Edited by E. Davis; P. Doherty; and E. Erdem. 47. Palo Alto, CA: AAAI Press.

Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C. L.; and Dollár, P. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision (ECCV 2014)*, 740–55. Berlin: Springer.

Luong, M.-T., and Manning, C. D. 2016. Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Vol. 1: Long*

Papers, 1054–63. Stroudsburg, PA: Association for Computational Linguistics.

Minsky, M. L. 1980. Decentralized Minds. *Behavioral and Brain Sciences* 3(3): 439–40.

Minsky, M. L. 2006. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*, 95. New York, NY: Simon and Schuster.

Nolfi, S.; Bongard, J.; Husbands, P.; and Floreano, D. 2016. Evolutionary Robotics. In *Springer Handbook of Robotics*. Edited by B. Siciliano, and O. Khatib. 2035–2068. Berlin: Springer.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. *SQuAD: 100,000+ Questions for Machine Comprehension of Text*. arXiv preprint arXiv:1606.05250. Ithaca, NY: Cornell University Library.

Robson, S. J., and Kuhlmeier, V. A. 2016. Infants' Understanding of Object-Directed Action: An Interdisciplinary Synthesis. *Frontiers in Psychology* 7: 111.

Rota, G.-C. 1986. In Memoriam of Stan Ulam: The Barrier of Meaning. *Physica D. Nonlinear Phenomena* 22(1–3): 1–3.

Sakaguchi, K.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2019. *WinoGrande: An Adversarial Winograd Schema Challenge at Scale*. arXiv preprint arXiv:1907.10641. Ithaca, NY: Cornell University Library.

Turek, M. 2018. *Machine Common Sense*. Arlington, VA: Defense Advanced Research Projects Agency. www.darpa.mil/program/machine-common-sense.

Turing, A. M. 1950. Computing Machinery and Intelligence. *Mind* 59(236): 433–60. doi.org/10.1093/mind/LIX.236.433.

Williams, L. E., and Bargh, J. A. 2008. Experiencing Physical Warmth Promotes Interpersonal Warmth. *Science* 322(5901): 606–7. doi.org/10.1126/science.1162548

Melanie Mitchell (mm@pdx.edu) is professor of computer science at Portland State University and the Davis Professor at the Santa Fe Institute.