EPIC2020
Ethnographic Praxis in Industry Conference

**CATALYST**

# Scale and the Gaze of a Machine

RICHARD BECKWITH, *Intel Labs*
JOHN SHERRY, *Intel Labs*

*Scale suffuses the work we do and, recently, has us considering an aspect of scale best suited to those with ethnographic training. We've been asked to help with scaling up one of the latest blockbusters in high tech – deep learning. Advances in deep learning have enabled technology to be programmed to not only see who we are by using facial ID systems and hear what we say by using natural language systems; machines are now even programmed to recognize what we do with vision-based activity recognition. However, machines often define the objects of their gaze at the wrong scale. Rather than "look for" people or objects, with deep learning, machines typically look for patterns at the smallest scale possible. In multiple projects, we've found that insights from anthropology are needed to inform both the scale and uses of these systems.*

*Keywords: Deep Learning, Human Scale, Ethnographic Insights*

## PEOPLE THINK AT A HUMAN SCALE

When we talk about "human scale", we refer to the sizes of objects and spans of time that people tend to think about. We humans don't *have* to think on the human scale. We *can* think on the scale of the universe or the atom. However, thinking at the human scale is natural; it is what allows us to collaborate; it allows us to see the reasons in another's acts; it supports our sociality. Although we can argue with an imposition of "rationality" on broad of swaths human thought (e.g., Malinowski 1922/1984), we also must admit that it is typically rather easy to attribute a rationale to what a person has done. We naturally "see" what other people are doing; machines do not.

Why don't machines just see like humans? Humans program the machines after all. The reason is that machines would need to be programed to see at a human scale and, at this point in time, that hasn't been the case. It's quite hard and there are alternatives. Machines have been programmed to a surprising level of accuracy, to be sure, but that's not enough. You can be accurate and yet not correct. The human ability to see what others are doing – this "vision" – is not the same as being able to describe the outward behavior that people have engaged in. The social sciences became convinced of that disconnect following the fall of behaviorism. Now, the social sciences rarely provide an "objective" description of the "behaviors" of others, rather, we offer what might be called a "preferred description." (Searle, 1983). Someone might describe another's behavior as alternating movement of the legs across a floor, but this would likely not match how the person would describe it themselves. An observer might say that a subject has walked to the north, which may be true, but the walker may not even have known the direction. It's more likely that the person being observed had thought that they were walking to the exit. "Walking to the exit", then, is the preferred description and these descriptions are easy for humans to generate about each other. It seems fairly obvious that a person watching that walker would say the same thing, and perhaps this is what Malinowski may have had in mind – that he could look at

Trobriand Islanders and their culture and imagine why they would travel great distances to bring some long-held possession to be held by another. That attribution is thinking at a human scale (e.g., Dennett on the "intentional stance", 1978). It's so much easier to collaborate with, to trust, another whom you can at least convince yourself that you can understand. So, it can be a real problem when "thinking machines" don't think like us.

### Machines Don't Have to Think in Human Ways

One of the reasons that our technology company hires social science types is to help to design technologies such that they are better partners. It used to be we were asked to help make purely responsive computers that would fit with people's lives. Now, the computer can take initiative (Console, et al. 2013) and fitting in is so much more significant. New technologies promise to be more connected to their environment and better able to understand and interact with people in more natural ways. That promise is where the problems start. It's frequently the case that "high technology" is designed in a decidedly non-human way and we're here to tell the choir that these machines can be harder to collaborate with and harder to trust than people. In many ways, what we are trying to do in our work is to help to create technology that can truly participate at the human scale or to point out when machines are incapable of working with that way people.

We'll detail some examples from the technology literature and briefly describe some cases that we're working on, but before that, we'll lay out a technology domain to which we will restrict our focus, one that is not only salient these days but which also highlights the value of the social sciences for technology development, namely artificial neural networks or, more simply, "neural nets".

## NEURAL NETS

Neural nets are the "iron horse" of the 21st century. OK, maybe "neural net" is just a similarly inapt metaphor. Iron horses weren't remotely horses and neural nets aren't remotely brain-like. Despite not being horses, railroads have been remarkably useful as a means of transport. They deliver goods, simplify travel, and can be quite reliable. Neural nets can be remarkably useful, too. As many people know, neural nets are terrific at finding pictures of cats (Le, et al., 2012). Moreover, neural nets are driving significant innovation in the computing industry. They have enabled improved multimodal sense-making and understanding (Owens and Efros, 2018), automated speech recognition (Chan et al., 2016), and natural language processing (Vaswani et al., 2017) and, then, there's that near magic we see with computer vision; and, it goes well beyond cats (Krizhevsky et al., 2012).

### AlphaGo

For famous example, AlphaGo, which debuted in 2016, was built on a neural net that was programmed to play the game Go (Silver et al., 2016). Go is a two-player game where players capture space by putting colored playing pieces on a game board. At its debut, AlphaGo beat a world champion Go player in four out of five games. This was a surprise to nearly everyone, including the AI community, because Go is considered much harder than

chess for a computer and computer scientists had worked for years on computer chess before being able to beat a human champion.

While there are lots of different aspects to the game and the program, we want to focus on just one aspect here – the Go board, its moves, and how AlphaGo sees them. First, let's consider how humans see Go. The Go board is a grid of lines that form 19 squares across, 19 squares deep, and has 361 intersections. (361=19x19) These intersections are where a player puts the playing pieces – "stones" – which are black for one player and white for another. Players take turns in placing one of their stones on an empty intersection. The goal of the game is for a player to build continuous walls of their stones around sections of the board such that their walls enclose more space than their opponent's. When a player puts down a stone, it is to either build a wall of their own pieces or block their opponent from building a larger enclosure. Any surrounded stones of an opponent are taken as prisoners. Each completed game takes about 250 turns. It will be relevant in the paragraph after the next to have noted here that, for humans, reading the current paragraph once or twice would allow a person unfamiliar with Go to not only play the game but also create a functional board with playing pieces.

Humans see a Go board as a 19x19 grid on which walls are built with stones. That's not how AlphaGo sees the game. AlphaGo sees the Go board as one long vector with a separate element for each of the 361 intersections. The training data for AlphaGo consist of game length sets of these vectors with each consecutive vector in the set representing each subsequent move in a game. As AlphaGo sees it, each move in the game is represented by a new vector that is different from the previous vector by one element (i.e., the new stone) or more if an opponent has been surrounded and their pieces taken as "prisoners". The bottom line is that people playing Go see the building of walls around sections of the playing surface; AlphaGo sees patterns in a series of vectors.

Before playing a game against a person, AlphaGo will look at, literally, millions of games to see what patterns emerge in the vectors. Once AlphaGo has seen millions of games that were played, it can figure out how to win. More specifically, AlphaGo can figure out which next step (i.e., which change in one element) is most likely to lead to a win and with each step chooses the move it believes will get it closer to a win. In order to learn to play at the level it played, AlphaGo needed to see millions of games that had been played. Interestingly, in order to play at all, AlphaGo would likely have needed to have access to nearly as many completed games. This requirement of seeing millions of games, it must be noted, is simply not true of humans who can learn the game quickly (as noted in a previous paragraph) and people are unlikely to ever encounter a million games in their lifetime let alone by the time they've played their first opponent.

## Feature Engineering

To be perfectly honest, almost none of that is central to the argument we want to make. What we care about most is that the two-dimensional 19x19 grid on the board on which a person sees walls, AlphaGo sees as a simple line with pieces of data about the state of each cell (black, white, or empty) which forms patterns with the state of the board in nearby lines. That AlphaGo sees the state of the board as linear is quite significant since a line can have no walls. AlphaGo simply finds patterns in the sequence of changes between the lines within a game.

One can imagine that engineers didn't have to spend much time figuring out that a vectorized representation of a two-dimensional board was going to be good enough. They still had a single variable for each intersection and only three different states of those 361 intersections. Noticing patterns across elements isn't likely to be outside the ken of an artificial neural network and, frankly, there isn't much else going on in the training data that the machine would need to notice or would be distracted by. The system only needs to know possible next steps and the likelihood that a change in arrangement on the board will lead to a winner. So, the feature engineering for AlphaGo would have been fairly simple. Nevertheless, feature engineering is an important part of any neural network or machine vision system and is nearly always much more complex than what we've seen with Go.

In fact, deciding which features to include in training a neural net can be quite difficult especially in areas like vision or language which so often seem magical. Because of this difficulty, engineers have discovered ways to allow a program to find its own features. This is called "automatic feature engineering". Despite the fact that automatic feature engineering has some fairly significant issues, in many ways, it is the magic of vision and language neural networks and underlies the ability to find so many cats. Yet, it can lead to a particularly pernicious type of problem – inferences based on spurious correlations.

Spurious correlation errors are one of the more significant side effects of automatic feature engineering. Obviously, spurious correlations are not just a problem for deep learning. People fall prey to spurious correlations, too. Consider for example, the recent conspiracy theory holding that 5G radio towers cause Covid-19. The best evidence that proponents have for this theory are geographic heatmaps showing that, in February and March of 2020, Covid hotspots and the then-current 5G deployments lined up quite well. The correlation between maps looked compelling, and without a more sensible explanation, 5G could seem like a reasonable-enough theory. The reason for maps lining up, according to experts, was that Covid was hitting urban areas hard and urban areas are also where 5G rolled out first. The correlations between Covid and 5G were spurious. What is important to note here is that we can see the sense of people's mistaken explanations – "the maps lined up so well"; there is a transparency to the error.

Often, transparency of errors isn't the case with deep learning. In fact, sometimes the errors generated with deep learning seem inexplicable. Research on attacks against deep learning systems can demonstrate how opaque the reasons for an error can be. For example, researchers have created patterned eye-glass frames that will fool a state-of-the-art facial recognition system created with automatic feature engineering (Sharif, et al. 2016). This system was trained to recognize different celebrities. The automaticity in the facial recognition system had the system look for pixel-level differences between a number of photos that were labeled with different celebrity names. As with the Go board, the system looked at each picture as a long vector. That is, photos were seen as a long line of pixels. In these digitized photos, the pixels are row after row of dots, each of which is one color, not unlike the Go board with its 19 rows of 19 columns and three states per element. Photos are just more complex than a Go board: more rows, more columns, and more states per element. Instead of Go's three states, the colors of a photo can include 100s of options or more. So, an image is, like the Go board, seen as a vector, but a much longer vector with much more varied contents.

The complexity of digitized images means that there is a greater chance of spurious correlations. The photos of celebrities offered spurious correlations aplenty. The

researchers in this study found that they could design a set of colorful eyeglass frames, each of which appeared to have a random design, but the design would match a pixel pattern associated with a particular celebrity. The researchers discovered patterns that would fool the vision system into believing that one person was another. For example, despite the fact that the system was excellent at recognizing photos of Reese Witherspoon, a picture of her was mistaken for Brad Pitt when she was pictured wearing the Brad Pitt glasses (or other celebrities when other glasses were used). [We suppose we should mention that to most people, these two celebrities don't look much alike.] Any person wearing the Brad Pitt glasses would look like Brad Pitt as far as the system was concerned. Brad Pitt was identified by the pattern of pixels in the eyeglass frames (there were certainly other "random" patterns of pixels that happened to be associated with Brad Pitt but those on the glasses were sufficient for identifying him.) Despite being state-of-the-art, the facial recognition system fell for a spurious correlation. However, unlike the similarity of the maps of 5G and Covid, the correlation that the system found between name and pixel pattern was not something that a person could ever see. The patterned glasses don't even remotely look like Brad Pitt or any of his features. The errors would make more sense if the researchers had deployed prosthetic chiseled chins to make someone look like Brad Pitt. People simply don't hypothesize identity of others based on random patterns in pixels.

## PROBLEMS FOR ETHNOGRAPHERS

So, now we've covered neural nets and feature engineering and the problem with spurious correlations and can now turn to projects we've worked on to highlight some of the issues that ethnographers are best able to deal with.

### Communication versus "Natural Language" Networks

One of the projects that we are now working on is a system that will use deep learning to translate between American Sign Language (ASL) and English. The idea is to find patterns in videos of people signing and relate those patterns to simultaneous English translations. The videos we are using sometimes have ASL translated to English and, other times, English translated to ASL. In all cases, these videos include ASL and English that are intended to express the same content. The goal is to have an "end-to-end" system that learns from videos of signing and an associated translated text of the spoken language used as a "label" for the signed content. Tens of thousands of these labeled videos are required for the system to begin to learn to translate.

Given that the system's input streams include raw video, it will not be surprising to hear that the system will be looking at the video as a sequence of vectorized images with the top left corner of the video being the first element in the vector and the bottom right pixel being the last. The features that the system will discover are like those of the celebrity ID system – in that they are patterns of pixels associated with some label.

A knowledgeable signer of ASL would look at the video and see a sequence of meaningful tokens (i.e., morphemes) composed of a set of language specific building blocks (i.e., phonemes) but the deep learning system has automatic feature engineering and is learning without seeing or knowing anything about phonemes or morphemes and, further, is not being programmed to acquire them. By focusing on pixels and patterns of pixels, the

system is far simpler to program. By focusing on these language independent features (i.e., pixels), problems with spurious correlations are rife. Systems in the future will be able to learn morphemes and phonemes first and acquire the language with that "knowledge". This is the only way to avoid the problem of "spuriosity". But this is only the beginning; the problem with pixels goes further than that.

Ethnographers trained in microanalysis can say more about what a knowledgeable user of ASL sees or how a fluent signer would construct and understand meanings. What microanalytic techniques brought to the study of communication was to show where relevant data had been ignored in trying to assign meaning: The weight of conversation is not carried only by syntactically words; there are non-linguistic gestures, postures, and eye gaze (Birdwhistell 1970, Kendon 1967, Schegloff 1998). There's intonation, pitch excursion, and volume. Conversation even moves forward with what is not said (Watzlawick, 1967). These all fly under the banner of "microanalysis". What microanalysis brought to the more strictly behavioral concerns of the time was a research program that asked what needed to be considered in the way that people construct and understand meaning when they communicate. This methodology is associated with anthropology as much as communication theory; both areas study meaning and the technologies and techniques with which meaning is shared. There can be no question that a fluid and facile interpreter will need to consider these cues nor that a system meant to interpret must also consider them. A job for the ethnographer working with deep learning is discovering both the right level of analysis and an ontology that makes sense…and then advocating for them.

## The Interpretive Stance and Machine Vision Networks

Part of the magic of these deep learning systems is not only that they can work at all but also how well they work once they do (remember all those cat photos). Part of the problem, is that when they make an error, it will not be an error that a person is likely to be able to understand. That is, it won't fail in a human way and a person working with it is unlikely to be able to determine what data it considered and how it was analyzed while making an inference. When the system offers a solution, a user may find it difficult to know that it has failed. Simply put, when the errors are not on a human scale, it is difficult for a person to be able to correct it, to work with it.

## How Do You Work with Failure?

Arguably, effective translation is crucial, and errors could be life threatening. However, it is also the case that, in an operational system, an ethnographer will have insured that conversational methods of correction would be in place. Perhaps an example where the system performs as an autonomous tool would help to highlight the potential risk of our misunderstanding how a machine sees. Here's another example from the tech literature.

The boffins have taken deep learning's most common machine vision training set (i.e., ImageNet (Deng, et al., 2009)), played with something quite like the Brad Pitt eyeglasses noted above, and come up with something diabolical (Athalye, et al., 2018). While it doesn't include celebrity photos, ImageNet is a database of one million images of many different classes of objects. This database is used by many deep learning practitioners to build systems that identify new images of the object types included in ImageNet (like turtles and

rifles). In this case, the boffins trained up a network so that it had world-class performance in identifying the object categories.

One of the object categories that is relevant for this story is that of "rifle". Rifle plays the role of Brad Pitt here. What's interesting is that these researchers used the seemingly random pattern of dots/pixels associated with "rifle" and did something akin to what the other researchers did with the Brad Pitt pixels. Instead of eyeglasses, they manipulated a view of a 3D toy turtle with this random dot pattern. Then, they rotated the turtle and placed the dots such that from every angle, the toy turtle looked like a rifle to the network. To the human viewer, the coloring wound up looking a bit like turtle camouflage. So, instead of a person wearing colored frames on a pair of glasses and then looking like Brad Pitt, a toy turtle was misidentified as a rifle. One can imagine negative consequences that could follow from having a child bring such toy into a protected area…very negative consequences and the reason for the error would not be at all obvious to those protecting that area. Because of the way pixel-based systems work, one would hope that a security detail would never rely on one. (Of course, police departments do use deep learning-based machine vision already (Harris, 2019).) Clearly, there's more for the ethnographer to do.

## Collaborating with a Deep Learning System

People expect that others, whether a person or a system, will see things as they do. We can learn a lot about how we see things by thinking about how we live and work with others. It's really important that we communicate and agree on what things are. If we see something, we expect that an intelligent other will see the same thing and call it by the same name. If someone calls something by a name we know, we expect that thing to be what *we* would call that name.

Communicating people don't necessarily agree on everything but, at least where collaboration is concerned, we usually mean the same thing with words. Formally speaking, ontologies do not have to be identical, merely sufficiently overlapping and with a method for finding and resolving difference, if need be (Ludwig, 2016). Some remaining differences are fine as long as we understand what they are.

For example, the Kahluli of Papua New Guinea consider the male and female birds of paradise to be different species and this is entirely consistent with their knowledge that the two come together for breeding (Feld, 1982). This isn't likely to be of much consequence when dealing with a Kahluli person. If you want to see a male bird of paradise, you simply ask to see one. It doesn't matter that it isn't considered the same type of bird as the female. In fact, this is kind of what much ethnography has always been about: How should we understand others outside our group? Classic Ethnography is rife with examples of ontologies that aren't shared. Ethnographers can work with that and explain how to understand each other.

In this way, ethnography, like anthropology more generally, assumes a rejection of radical incommensurability. This rejection of incommensurability means that the concepts deployed by one entity (individual or collective) can be understood by another. When someone discusses their family, say, they may have in mind a different set of people from what another might assume but, if the two share sufficient beliefs, they can discuss the boundaries of the concept of "family". Foucault would have called this an episteme (1971) and Kuhn, a paradigm (1962). What's important is that our categories are fluid and we can

work within and, to a great degree, between them. Ethnography assumes a level of commensurability sufficient that someone could explain another in terms that are understood.

## Practically Incommensurable and Practically Inscrutable

Unlike the subjects of ethnographic work, systems created using deep learning are practically incommensurable because they are practically inscrutable. That is, in practice, such systems work with very different concepts from the people who work with them and it will take a lot of work to get to a point where differences can be discovered and resolved.

*Incommensurability*

If you imagine that a system is observing as you would and "describing" those observations in terms that you would use, a deep learning system could easily be seen as an unreliable observer; however, nothing is further from the truth. The system is quite a reliable observer; under the same conditions, it will come to the same conclusions. It is the expectations of the naïve user that is a problem because the borders of the system's concepts are considerably different from our own (consider Brad Pitt or the rifle). Still, how can you rely on someone who tells you things that you know are simply wrong? How can you work with someone you don't understand and with whom you cannot negotiate a shared meaning? It is only by coming to understand the other's constraints. An example from our work might help.

We work with another project that uses machine vision. This one watches factory workers. The goal of this system is to improve safety while, at the same time, facilitating training, automating record keeping, and increasing efficiency. The video cameras constantly observe and record. The way this system works is that it has been trained to recognize the steps in procedures undertaken by skilled technicians on the factory floor. These technicians are taught a particular plan, composed of a set of steps, done in a particular order. The system learns to recognize them. If you think this sounds like it could lead to Taylorism run amok, you won't be the first. Watching people and watching how they are doing what they are doing is important for safety and practical training but could also be seen as something that would provide management with an unwelcome gaze over the worker.

When we keep in mind the difference between a plan and a situated action (Suchman, 1987), we know that as good as a plan may be, a person may need to veer from that plan to account for local conditions. So, when the local situation requires it, an intelligent being will find a way to reach the appropriate end state despite having to change some part of a plan. This is not a situation that a typical deep learning-based system can account for. One thing that a machine vision cannot do is to recognize something new. It does not recognize novel actions for what they are, it simply recognizes that they are not the expected step.

Because of this, one of our roles here was to explain to management why they shouldn't always have access to what the machine "sees". An example came up in our work. During an observation, we saw an expert "going through the steps" when someone came up to them with a problem. This was standard protocol where someone with a problem should come to someone more senior for assistance. This new problem was solved and the expert returned to his task. This diversion would, of course, have caused the lengthening of the

time of that interrupted step, not to mention the overall process. Some members of management wanted to know what was happening every time the system didn't see what was expected but this is the sort of naïve error that would cause disruption in the work being done.

*Practically inscrutable*

Developers often say that one simply can't understand how a deep learning system works. It is difficult, to be sure, but the workings of the system could be understood. Jose Hanson (Hanson and Burr, 1990) argued years ago that because neural nets are implemented on state machines, we know that they *can* be understood: one state leads to the next by virtue of an explicit command and there is a set of input data; each can be clearly seen. It just takes a lot of time to analyze, a whole lot of time. It took Google weeks to figure out how AlphaGo came up with one of its moves and explain why it was able to beat the world champion Go player using that move. The important point, though, is that they *could* explain it. It *was* possible. It was just ridiculously hard. A non-expert could not be expected to interrogate a system in any kind of reasonable time. Experts can't even do this quickly. So, how do we interact with a machine?

With inexplicable ontologies derived from patterns in pixels, understanding is surrendered to a mostly "well-performing system" built in a way to ease machine processing.

## Human Scale: Description and Explication

Another way of looking at the previous examples is as a "failure of description". The system in the factory setting had an incomplete description of the technician's job. Going and helping another technician is actually a prescribed part of the job; it's merely infrequent. But as far as the system was concerned, prolonged absence from the process it knows is a problem like any other. So, a problem is seen where none actually exists because the system hadn't been trained to recognize this option (or myriad others). All of the possible actions that might be correctly undertaken by a technician are not possible to train the system to recognize because there are countless correct things to do and ways to do them. Instead, what the system can do is to learn a limited set of actions that could be undertaken and watch to see when they are done correctly. There are many valuable services that such a program can provide but 24/7 understanding of everything it sees is not one of those.

We also see this failure of description in the case of ASL. The system had not been trained to recognize the data relative to the categories of human perception, the types of data that make human language possible even when we're not explicitly aware of them. That is, phonemes and morphemes are important ways that humans see language even when we're not aware of them. And we noted that there are other types of data that the system won't see. Research in ethnomethodology, conversation analysis, and embodied interaction have demonstrated that we are signaling each other in many ways, often unconsciously, but those signals are nonetheless important for the interpretation of meaning. This may include such factors as subtle body positioning, direction, timing and coordination of gaze, and a host of other signals that happen too quickly or subtly to be easily described, but which nonetheless affect communication. The problem is, except for such micro-analytic work, those signals

are rarely even acknowledged and, to our knowledge, have never been included in a deep learning-based natural language system.

While (at least, heuristics for) each of these communicative categories could be learned by the system, it could only happen by resisting the emerging standard for such deep learning systems. Seeing ASL as a set of vectors of pixels, simply doesn't bode well for bringing this up to a human scale. Pixels are too fine a scale. Humans think of and see things in ways that are difficult to find in sets of pixels.

The challenge that failures of description present for deep learning systems, then, is that these systems will always be hamstrung.

A way out: changing how we design DL systems. Rather than designing a system as though it completely describes a process (e.g., servicing a tool or translating ASL), we should be developing systems that watch for events in the environment and provide further information in ways that recognize a potential insufficiency and are always compatible with the possibility of error. This is how we can provide a reasonable user experience in the face of deep learnings benefits and limitations. Spurious correlations will still happen, or maybe more correctly, "meaningless" event detection will still happen. Consider Geertz's discussion of the meaning of a wink (1973). Sometimes a wink will just be dust in someone's eye.

The implication here is that our DL systems must be bounded and targeted at the kinds of recognition tasks that make their level of activity more commensurate with human understanding and assessment. That means any individual DL system will perform a task that, if it produces a result that is not meaningful or useful to the human, the human user doesn't require hours of analysis to figure out what happened, but rather can disposition the result quickly, and in a way that the system can learn from.

## SUMMARY

The intent of this paper was to argue that one of the most significant recent directions in technology – deep learning – has flaws that are best addressed by those trained in ethnographic methods. Who better than ethnographers to advance the cause of human scale?

A generation (or two!) ago, ethnographers were brought into technology development in order to help people make products that fit people so that businesses could "scale up" their offerings and make them relevant for the whole world. However, once they were inside the corporation, so many more problems were revealed to be within the ethnographer's domain.

Atomic units are used to simplify programming. Pixels are used for images and spectrographic-style frequency analyses for speech sounds. It does simplify programming, too. It's just that it is the wrong level of abstraction for dealing with people.

The work we presented here was to say, in part, how we might create machine learning that works well but, beyond that, it's also about developing AI systems that can be more easily understood by people. Much of today's deep learning consists of the type of system that Latour could point to as being particularly rife with blackboxing (1999); because it is practically impossible to know how they work. Successful scaling-up of the technology should not mean that no one will have access to the methods behind the madness.

By getting the scale right for human understanding, we can hope to have more control over the gaze of the machine. This may slow down both processing and even system creation; it could even mean that a given system would not be as broadly applicable. But it

would be a better system, working at a more human scale, and would enable more fundamental interaction with the system itself.

**Richard Beckwith** is a Research Psychologist at Intel Corporation's Intel Labs. He is a psychologist who studies the impact that emerging technologies have on those upon whom they emerge and helps to ensure that technology designs can support people in the way that they should.

**John Sherry** is the director of the User Experience Innovation Lab in Intel Labs. This organization focuses on the human dimension of machine learning technologies from diverse perspectives, to better imagine and prototype new technological possibilities, and anticipate the alignments necessary for those to become reality.

## REFERENCES CITED

Athalye, Anish, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. 2018. "Synthesizing Robust Adversarial Examples". Accessed [24 Aug 2020]. https://arxiv.org/pdf/1707.07397.pdf.

Birdwhistell, Raymond. 1970. *Kinesics and Context: Essays on Body Motion Communication*. Philadelphia: University of Pennsylvania Press.

Chan, William, Nadeep Jaitly, Quoc Le and Oriol Vinyals. 2016. "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, 2016, pp. 4960-4964, doi: 10.1109/ICASSP.2016.7472621.

Console, Luca, Fabrizio Antonelli, Giulia Biamino, Francesca Carmagnola, Federica Cena, Elisa Chiabrando, Vincenzo Cuciti, Matteo Demichelis, Franco Fassio, Fabrizio Franceschi, Roberto Furnari, Cristina Gena, Marina Geymonat, Piercarlo Grimaldi, Pierluige Grillo, Silvia Likavec, Ilaria Lombardi, Dario Mana, Alessandro Marcengo, Michele Mioli, Mario Mirabelli, Monica Perrero, Claudia Picardi, Federica Protti, Amon Rapp, Rossana Simeoni, Daniele Theseider Dupré, Ilaria Torre, Andrea Toso, Fabio Torta, and Fabiana Vernero. 2013. Interacting with social networks of intelligent things and people in the world of gastronomy. ACM Trans. Interact. Intell. Syst. 3, 1, Article 4 (April 2013), 38 pages. DOI:https://doi.org/10.1145/2448116.2448120

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Fei-Fei Li. 2009. ImageNet: A Large-Scale Hierarchical Image Database. IEEE Computer Vision and Pattern Recognition, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

Dennett, Daniel. 1978. *Brainstorms*. Montgomery, VT: Bradford Books.

Feld, Steven. 1982. *Sound and Sentiment: Birds, Weeping, Poetics, and Song in Kaluli Expression*. Philadelphia: University of Pennsylvania Press.

Foucault, Michel. 1971. *The Order of Things: An Archaeology of the Human Sciences*. New York: Pantheon Books.

Geertz, Clifford. 1973. "Thick Description: Toward an Interpretive Theory of Culture." In *The Interpretation of Cultures*, 3–30. New York: Basic Books

Handle, Richard. 2009. "The Uses of Incommensurability in Anthropology." *New Literary History*, Vol. 40, No. 3, pp. 627-647.

Hanson, Stephen José & David Burr. 1990. "What Connectionist Models Learn: Learning and Representation in Connectionist Networks." *Behavioral and Brain Sciences*, 13(3), 471-489. doi:10.1017/S0140525X00079760

Harris, Mary. 2019. "Amazon Encourages Police to Use Untested Facial Recognition Technology." 24 May, 2019. Slate website. Accessed [18 August, 2020] https://slate.com/news-and-politics/2019/05/facial-recognition-police-officers-hillsboro-oregon-amazon.html

Kendon, Adam. 1967. "Some functions of gaze-direction in social interaction." Acta Psychologia 26: 22-63.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. "Imagenet Classification with Deep Convolutional Neural Networks." Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, Pages 1097–1105.

Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. University of Chicago Press.

Latour, Bruno 1999. *Pandora's Hope: Essays on the Reality of Science Studies*. Cambridge, Massachusetts*:* Harvard University Press.

Le, Quoc V., Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeff Dean, and Andrew Y. Ng. 2012. "Building high-level features using large scale unsupervised learning." In Proceedings of the 29th International Conference on International Conference on Machine Learning. Omnipress, Madison, WI, USA, 507–514.

Ludwig, David. 2016. "Overlapping Ontologies and Indigenous Knowledge: From Integration to Ontological Self-Determination." Studies in History and Philosophy of Science 59: 36-45.

Mahmood, Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. 2016. "Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition." In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications.* 1528–1540. DOI:https://doi.org/10.1145/2976749.2978392

Malinowski, Bronislaw. 1984. *Argonauts of the Western Pacific: An Account of Native Enterprise and Adventure in the Archipelagoes of Melanesian New Guinea.* Prospect Heights, Ill.: Waveland Press.

Owens, Andrew and Alexei A Efros. 2018. "Audio-visual scene analysis with self-supervised multisensory features." Proceedings of the European Conference on Computer Vision. 639-658.

Schegloff, Emanuel. 1998. "Body torque." Social Research 65: 535-596.

Searle, John. 1983. *Intentionality: An Essay in the Philosophy of Mind.* Cambridge: Cambridge University Press.

Silver, David, Aja Huang, Christopher J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* 529, 484–489. https://doi.org/10.1038/nature16961

Suchman, Lucy. A. 1999. *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge: Cambridge University Press.

van der Wal, Reine. C., Robbie M. Sutton, Jens Lange, and João Braga, J. 2018. "Suspicious binds: Conspiracy thinking and tenuous perceptions of causal connections between co-occurring and spuriously correlated events." *European Journal of Social Psychology*, 48(7), 970–989. https://doi.org/10.1002/ejsp.2507

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is All You Need." In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc. 6000–6010.

Watzlawick, Paul, Janet Beavin Bavelas & Don D. Jackson. 1967. *Pragmatics of Human Communication: A Study of Interactional Patterns, Pathologies, and Paradoxes*. New York, NY: Norton.