James Brown
Carr Essay Submission

Learning in a Community of Black Boxes

In one of the most famous essays about computing and machine intelligence, Alan Turing made it clear that he didn't always know what his machines would produce: "Machines take me by surprise with great frequency." Turing responded to Ada Lovelace's argument that computers can only ever do exactly as they are instructed by insisting on a computational machine's capacity to surprise. The unpredictability of computing is something that often gets lost in our contemporary conversations, conversations that are focused on how humans can use computers to solve the world's problems. Turing's observation reminds us to embrace a more uncertain approach to computing by granting that we are not in control. Humans do not merely use machines. They are also used by them. As any programmer will tell you, the constraints of hardware and software shape how one addresses problems and solutions, and this isn't even unique to computational machines. Upon using a typewriter, Nietzsche suggested that "our writing tools are also working on our thoughts." He saw the typewriter actively shaping his philosophical writings. When we program computers, we collaborate much more than we control.

The idea that we collaborate with and relate to computers is also in keeping with another key idea in Turing's essay: that machine intelligence will be the result of *educating* machines. Turing argued that any attempt to imitate an "adult human mind" should recognize how that mind is shaped. For him, the components of that process included its initial state, its education, and its various other life experiences. Turing suggested that our best approach would be to simulate a child's mind and then to educate it: "Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain." Turing argued that a child brain is like a blank notebook with "little mechanism" and thus could be "easily programmed."

We might take issue with any number of assumptions made by Turing about the human mind, especially the idea that education without life experience would be sufficient to simulate an adult mind. But I am most interested in linking his insistence that machines can take us by surprise directly to his description of the education of a computational machine. Anyone who has programmed a computer is familiar with Turing's surprises, and so is anyone who teaches for a living. Students in my classes take me by surprise with great frequency. I don't mean for this to be a boast about the brilliance of my students (they often *are* brilliant) but rather in the sense that, like computers, they process information, lessons, and ideas in ways that I never would have expected. And they do so in a way that often seems quite black boxed. Like any learning system (machine or human), much of their process is opaque to me. I can ask them why they wrote something in a particular way or how they developed an idea, but that won't necessarily shed direct light on their inner learning processes. This is a problem shared across the humanities and sciences—we don't really know what's going on inside the human mind. Neuroscience,

cognitive science, psychology, and even my own field of rhetoric join numerous other fields in speculating about this curious black box. And yet we still teach, and people still learn. Uncertainty is forced upon us. Teaching and learning are partial, cloudy processes.

What if we treated our computers in a similar way? What if we simultaneously granted both the idea that machine learning processes are largely opaque *and* the idea that this does not prevent us from speculating about how they are making decisions? This would mean embracing a certain lack of control when it comes to interacting with computation, and this ethic may be exactly what we need as we try to build an educational and civic response to our contemporary challenges. When it comes to machine learning, we should let go of the dream of perfect control and knowledge, embracing Turing's surprises. However, this is not the approach we read in the headlines of *TechCrunch*, which recently exclaimed that "machine learning can fix Twitter, Facebook, and maybe even America" (November 26, 2016). And it is also not what we find in the pages of books like Pedro Domingos' *The Master Algorithm*, which aims to build a single "universal learner" that will solve a range of problems, from curing cancer to predicting stock market crashes.

Luckily, a more uncertain approach to machine learning is both available and viable. In "How the machine 'thinks': Understanding opacity in machine learning algorithms," Jenna Burrell provides a more circumspect take on machine learning that begins from the assumption that such systems are opaque. The opacity that is of most interest to Burrell is not rooted in one's lack of programming knowledge or in our inability to access proprietary systems. Instead, her focus is on how machine learning systems operate at scales foreign to the human mind. For instance, Burrell describes a machine learning system trained to recognize a particular kind of spam—the "Nigerian 419 scam." Based on training data, this system learned that the five words most associated with a spam message were "our," "click," "remov," "guarante," and "visit." Some of these make sense to the human mind, since we know that spam messages often have typos or encourage us to click on links. But what of the first word on that list? Few of us would expect "our" to be so highly associated with a spam message. For Burrell, this is just one example of how the processes of machine learning systems are fundamentally opaque to humans: "When a computer learns and consequently builds its own representation of a classification decision, it does so without regard for human comprehension." Machine learning systems don't always clearly reveal their models to us.

So, how do we know much about any learning system, human or computational? Machine learning algorithms raise this question in a particularly stark way. Even those who train machine learning systems don't often fully understand why they make the decisions they do. Again, this is perhaps what humans and computers have in common. We think of humans as being completely different from computational machines, and we are often caught up in the game of drawing lines between the two. This is understandable, as automation continues to reshape the workforce in drastic ways, but what do we gain if think of computers differently, as being alongside us (rather than at our beck and call) as we tackle difficult problems?

One of those difficult problems and one that has been the focus of much current discussion regarding machine learning is "fake news." As large tech companies try to contend with a flood of misinformation, many are suggesting that machine learning algorithms will play a role in flagging fake content and helping people parse information. Many of those same discussions have also insisted that humans will need to play role in this process as well. This is no doubt true. However, the fake news problem will likely not be solved by merely using machine learning algorithms or humans to point out what's "fake" and what's "real." To approach the problem this way is to forget the opacity of both human and machine learning processes. We may have no clear sense why a machine learning system has flagged content as fake, and we also may have no good answer for why content that is quite easily debunked continues to spread.

In the case of both algorithms and people, learning systems are complex and opaque, and when it comes to humans it is an oversimplification to think that pointing out that content is "fake" will deter the spread of lies. Political ideologies, not to mention a range of other cultural factors, shape what we believe and what we don't. How such political ideologies are constructed and maintained is another opaque problem, and that should not prevent us from trying to understand it. However, it might help us if we begin from the assumption that all learning follows confusing and often contradictory pathways. So, while using machine learning to help identify credible content should certainly be part of our efforts, it is not sufficient. We will also need to face up to the idea that we can't solve the problem of opaque learning processes by only working on better algorithms or more data, and we certainly can't perfectly control or program the processes of an SVM or a Jill Stein voter.

This might all sound completely bleak, but any single one of our attempts to teach and learn is built on this uncertain foundation. We should embrace it and not ignore it. In some sense, all we can do is allow the black boxes to be opaque, feed them data, observe how they respond, and then start the process all over again. "Fake news" is a pedagogical problem, and it is just as much about how we teach people as it is about how we teach machines. But I would argue that we should not be trying to get to the truth of our various black boxes (human and machine). In fact, assuming that this is possible actually stops us from doing the difficult work of grappling with how learning actually happens.

* * *

The question that prompted this essay asked about what it will mean to be human in the age of machine learning and what that new notion of humanity will mean for creativity, identity, love, communication, and community. Each of these terms opens up into massive questions that obviously cannot be solved with a single essay. My goal is more modest. I want to encourage us to rethink our traditional understanding of these terms. Creativity, identity, love, communication, and community. We tend to think of creativity and identity as having some clear origin. We consider love to be an expression of feelings that begin within us. We assume that community and communication are based on clearly agreed upon meanings.

But if we're really honest with ourselves, we can see that each of these terms is more about relations than internal states, more about speculation than certainty. Each of these concepts emerges at the edges of our relations to one another.

Like creativity, identity, love, communication, and community, learning sits at the threshold, offering us only momentary glimpses at how it operates. This is the case for both machine and human learning, and it means that we should love and care for our world and our tools in ways similar to how we care for other people. The opacity of learning processes is a foundational problem of both human and machine learning, but it should not and does not prevent us from attempting to solve deep and complex problems. Our impulse to break open black boxes in an attempt to understand how they work or to treat learning as an ultimately achievable engineering problem is driven by a fiction of control, by the idea that if we can just get to the bottom of things, we can gain pure insights into what motivates people and machines. But how we understand ourselves and others is always a result of opacity and speculation, and this is where we might begin our conversations about how machines and humans learn together.