Edge

To arrive at the edge of the world's knowledge, seek out the most complex and sophisticated minds, put them in a room together, and have them ask each other the questions they are asking themselves.

https://www.edge.org/conversation/tom_griffiths-aerodynamics-for-cognition
Printed On Tue August 21st 2018

CONVERSATION : MIND

Aerodynamics For Cognition

A CONVERSATION WITH Tom Griffiths [8.21.17]

Includes *Edge* Video and Audio

It's very clear that in order to make progress in understanding some of the most challenging and important things about intelligence, studying the best example we have of an intelligent system is a way to do that. Often, people who argue against that make the analogy that if we were trying to understand how to build jet airplanes, then starting with birds is not necessarily a good way to do that.

That analogy is pretty telling. The thing that's critical to both making jet airplanes work and making birds fly is the structure of the underlying problem that they're solving. That problem is keeping an object airborne, and the structure of that problem is constrained by aerodynamics. By studying how birds fly and the structure of their wings, you can learn *something important about aerodynamics. And what you learn about aerodynamics is equally relevant to then being able to make jet engines.*

The kind of work that I do is focused on trying to identify the equivalent of aerodynamics for cognition. What are the real abstract mathematical principles that constrain intelligence? What can we learn about those principles by studying human beings?

TOM GRIFFITHS is a professor of psychology and cognitive science and director of the Computational Cognitive Science Lab and the Institute of Cognitive and Brain Sciences at the University of California, Berkeley. He is co-author (with Brian Christian) of *Algorithms to Live By*. **Tom Griffiths's *Edge* Bio page**

AERODYNAMICS FOR COGNITION

I work on computational models of cognition, which means that I'm interested in understanding how people do the amazing things that we do, like learning from small amounts of data, figuring out causal relationships, identifying languages—things that computers have traditionally found hard to do. The way that I think about motivating that kind of research is in terms of making computers better at solving those kinds of problems.

Recently, I've also been thinking about a different way in which that's a relevant enterprise. With all of the successes of AI over the last few years, we've got good models of things like images and text, but what we're missing are good models of people. If we look at the kinds of AI systems that are being built and the kinds of data that people want to understand, often those data have to do with human behavior. We're trying to understand why people do what they do and what the cognitive processes are that underlie the data we find in the world that are a consequence of human behavior.

This enterprise is important for a couple of reasons. It gives us the tools to make sense of these data that are becoming an increasingly important part of our lives. Also, having good models of how people think and behave is relevant to helping AI systems better understand what people want.

My approach is to try and understand the computational structure of the problems that people have to solve. If we're trying to understand how people, say, learn a new causal relationship, how do we formalize that? How do we turn that into a math problem? That's the kind of thing we can imagine getting a computer to solve.

Once we figure out the structure of those problems, we can figure out a good way of solving them. And there we draw on tools that come from AI, statistics, and machine learning as the basis for coming up with hypotheses about how human cognition might work. Using those insights, we can run experiments that test predictions that come out of those models, and then we can use that as a tool for digging deeper into how human cognition works and how people solve those kinds of problems.

There are two ways in which we're drawing on these computational tools to make sense of human cognition. One way is to characterize what we call inductive biases, which are the things other than the data that lead people to reach good conclusions about the processes that might have generated those data.

If we try to formulate a problem like, say, learning a language, the way we do this is by getting some data—hearing what people are saying around us—and then trying to make sense of those data by entertaining different hypotheses about the structure of the language and the processes that might have produced those data. We can formulate that as a kind of statistical problem, where you take the data and try to evaluate which of these hypotheses are right.

What's amazing about human learning is that people are able to solve that problem remarkably well. We're able to learn languages, infer causal relationships, learn new words, learn new categories from small numbers of examples where there's not enough statistical information to allow you to have any kind of certainty. The way that we explain that kind of intelligence is in terms of having something that allows us to narrow down the space of possibilities, something that allows us to make good guesses and to come up with good answers even though we don't have all the information that we need. Machine-learning researchers call those things that influence our conclusions inductive biases.

One enterprise that we engage in is trying to understand the inductive biases that inform human cognition. How is it that people are able to make these inferences? What are the expectations that we have about how the world works—about the structure of languages, about what words might mean, about how physical objects interact that allow us to infer causal relationships? How is it that those things guide the inferences that we make?

One of the things that we try to do in our research is identify what those human inductive biases are like. We've identified a set of experimental methods that we use for solving that problem, using ideas that come from Bayesian statistics as a tool for characterizing what those human inductive biases are like.

For example, one of the things that makes people good at inferring causal relationships is that we have strong expectations about how causality works. If you take a statistics class and you learn about how you're supposed to detect a relationship, normally, the methods that you're using don't make a lot of assumptions about the nature of the causal relationship. All you're looking for is some kind of pattern of dependency between two variables. But if you tell a person to figure out if A causes B to happen, then people have a strong expectation about what that means. They think that if A causes B, what that means is A occurring increases the probability that B occurs by a lot. So, if A causes B, and if A happens, then it's really quite likely that B will happen.

Those two constraints—(1) the assumption that causes are generative, that they produce their effects and increase their probability, and (2) that causes are near deterministic, that if causes occur, they produce their effects with very high probability—really simplify the problem of trying to figure out whether causal relationships are present. You don't need as much data to figure out whether a relationship like that exists. You can just see a few examples and that's enough to establish for you that, in fact, there is an underlying causal relationship.

The other aspect of the work that we've been doing takes a step away from that abstract framework of trying to understand how people reason by thinking about the raw structure of the computational problems involved. More recently, we've been focusing on another aspect of human cognition that is a crucial part of human intelligence, which is our ability to program ourselves effectively.

One of the mysteries of human intelligence is that we're able to do so much with so little. We're able to act in ways that are so intelligent despite the fact that we have limited computational resources—basically just the stuff that we can carry around inside our heads.

But we're good at coming up with strategies for solving problems that make the best use of those limited computational resources. You can formulate that as another kind of computational problem in itself.

If you have certain computational resources and certain costs for using them, can you come up with the best algorithm for solving a problem, using those computational resources, trading off the errors you might make and solving the problem with the cost of using the resources you have or the limitations that are imposed upon those resources? That approach gives us a different way of thinking about what constitutes rational behavior.

The classic standard of rational behavior, which is used in economics and which motivated a lot of the human decision-making literature, focused on the idea of rationality in terms of finding the right answer without any thought as to the computational costs that might be involved.

This gives us a more nuanced and more realistic notion of rationality, a notion that is relevant to any organism or machine that faces physical constraints on the resources that are available to it. It says that you are being rational when you're using the best algorithm to solve the problem, taking into account both your computational limitations and the kinds of errors that you might end up making.

This approach, which my colleague Stuart Russell calls "bounded optimality," gives us a new way of understanding human cognition. We take examples of things that have been held up as evidence of irrationality, examples of things where people are solving a problem but not doing it in the best way, and we can try and make sense of those. More importantly, it sets up a way of asking questions about how people get to be so smart. How is it that we find those effective strategies? That's a problem that we call "rational metareasoning." How should a rational agent who has limitations on their computational resources find the best strategies for using those resources?

We're used to making decisions. The reason why this is an instance of metareasoning is that now we're making decisions about how we're going to make decisions. My graduate student Falk Lieder and I have been exploring how we can understand human strategy choice and the ways in which people end up making decisions from this perspective of meta-level rationality.

One of the consequences of thinking about this is that we gained new insight into some of those classic human irrationalities, the kinds of things that were explored by Kahneman and Tversky in the heuristics and biases literature. We can say that some of the things we do seem like pretty good strategies for solving problems.

One classic example of this is what's called the availability heuristic, which is making an estimate of the probability of something based on the examples that you can recall from memory. That can result in biases because things like plane crashes, terrorist attacks, shark attacks—the things that are very salient to us and stick out in our memory—people overestimate their probability as a consequence. We can show mathematically that following a strategy like that is a good way of making use of limited computational resources.

If you're trying to evaluate the expected utility of an action and you're only going to be able to consider a few different possible outcomes, then a good way to minimize the variance in your estimate—trying to get a less noisy estimate of that expected utility—is to sample events based not just on their probability but also on their utility. Something that is very bad is something you should over-represent when you're trying to evaluate making a decision because that's exactly the kind of thing that is going to have a big impact on your assessment of the relevant utility.

Taking this perspective gives us a few different practical insights. One insight is about how you can go about being more rational yourself. How do we do a better job of solving the problems that come up in the context of our own lives? One angle on that is maybe relax a little, don't feel so bad about how well you're doing at solving those problems. A lot of the strategies that we end up using represent a good point on that tradeoff between effectively using the resources that we have and making errors.

The other thing that comes out of this is that we can understand why we're making those errors, and understand that as the consequence of those limited computational resources. That suggests that if you want to change the way people behave, then trying to teach them the exact right way of solving the problem, if it's too computationally costly, isn't going to be very effective.

Another strategy might be to say that the reason we're having the problems we're having is because the algorithms that we're using are biased in the particular environment we're in, or because we're not able to devote the computational resources to allow us to plan more effectively. Maybe the place to intervene is on providing essentially those resources.

If you've got a human being and a computer, the computer has the computational resources, but the human being is the one who is going to make the decision. How do you combine those things together? One way might be to make a computer chip that you stick inside your head that augments the onboard computational resources that you have. That's something which is a little further in the future than we might like, so we've been focusing on a different way of doing this. Again, this is with my student, Falk Lieder. We've been looking at how you can use the computer to change the environment that the human being is in so that that human being ends up making better decisions.

We already do this to some extent. If you've ever used the strategy of gamification, where you're using an app or something that gives you points for completing tasks, or if you make a to-do list and you get satisfaction from checking things off, what you're doing is essentially using this external device as a mechanism for changing the environment that you're in.

We can go further than that. If you've got a computer that has information about the structure of the problem that you're solving and can communicate that information back to you through a mechanism like gamification by giving you rewards, then we can build a system that will help guide people to making more effective decisions by modifying their local reward function.

If you've got a decision problem and you're able to use a computer to solve it, then we have worked out the optimal gamification scheme, the right way of transferring information the computer has about what the best actions are into information that you can provide to people in the form of points or modifying the rewards that they get for doing something. So people end up solving the problem as well as we might hope.

An example of this is if you have a task that you're trying to complete. By getting all the way through the task, you get some reward at the end, but each of the steps in that task is something painful or frustrating for you, something like writing a book or another long-term project with these local costs. One way that you can make it easier to get to that point is by taking that reward you get at the end and spreading it back through time so that you're getting some incremental reward for completing each of those sub-tasks. That's a simple example because it's relatively straightforward to figure out how you should be rewarding yourself along the way for achieving smaller goals. But we've shown that the strategy is something that can be used for relatively complex sequential decision-making problems.

Another example is if people are only able to plan a few steps into the future. You have a computer that's solving a problem in a way that allows it to plan arbitrarily firing into the computer. We have a scheme for taking those solutions that the computer gets and putting them back into the problem as rewards along the way that allow it to correct for the fact that, even if the human being is only planning a few steps into the future, that human being wants to end up achieving the ideal outcome.

The most challenging problems that human beings solve, both from a human perspective and from a computer perspective, are the problems that involve other people. These are things like trying to figure out what another person's behavior means. Are they acting in a way because they like you or they don't like you? Are they doing that because they don't remember you? Are they making a decision because they prefer one thing over another? Trying to reason, from the actions people take to the mental states that they have, and trying to work out what the consequences of those mental states are is something that can be taxing for humans, but it's something we do automatically when we're interacting with another person. A lot of what you're doing is trying to reason about the mental states that they have. It's also something that is key to being able to operate in a society.

As you walk around and interact with other human beings, you are making inferences about the preferences of those human beings and, normally, doing your best to accommodate those preferences. The preferences can be as small as someone wanting to go in a particular direction, so you're just making sure that you're not blocking their way. They can also be as important as they prefer to stay alive, so you're doing your best not to interfere with that aspiration. Understanding how people make those kinds of inferences is something that is important for getting insight into potentially how to help people navigate some of those things in their own lives; it's also critical to being able to make computers that interact with humans in ways that are beneficial for both.

One of the interesting things about that problem is that when you formulate it as a statistical problem, it's a problem where you're getting data. You're seeing the way in which the person is behaving and you're forming hypotheses about what they want, what they prefer, and why they're doing the things that they're doing. In order to make that mapping from hypotheses to data and the reverse inference from data back to hypotheses, you need to have what a statistician would call a forward model or a generative model. That model tells you that if somebody believes this, then they will act in this way. If they want this, then they will act in this way. If this is their goal, then they will act in this way. We need to be able to make predictions about the data that we get based on the hypotheses that we're entertaining to reason backwards from the data back to the hypotheses.

That sets up this problem of trying to understand human behavior as a rational process or a boundedly optimal process of people trying to achieve their goals through their actions as a key ingredient of being able to work backwards and figure out what people are desiring or trying to do.

There are two interesting perspectives that this provides when we think about both cognitive development in children and social cultural development, where societies hopefully get better at solving different kinds of problems. When we start to think about what this implies about children, the first principle is that we should try to interpret children's behavior as the consequence of some kind of rational process. This may seem like a big ask, insofar as children are notoriously irrational in the sense of being highly variable—running around not doing things you want them to do, and so on. Some of that is because they have a different perspective on the world; they're operating from different information that leads them to act in different ways. But that kind of variability is exactly what we might expect out of an organism that is designed to solve problems rationally over the course of its entire lifespan.

One of the ideas that shows up in machine-learning research is the idea of the explore-exploit tradeoff, where you are trying to solve a problem in which you're going to have the same set of options repeatedly. For example, you've got a set of places where you could go for dinner, say, if you're living in a city. You're going to have the same set of options tomorrow. The explore-exploit tradeoff comes up because, when you're deciding to go out to dinner, you could either go to a new restaurant or you could go to a restaurant that you already know is good.

When you're trying to make a decision in that situation, you have to tradeoff these two things: Do I gather more information about the world, which might be useful for me when making a decision about restaurants in the future, and might ultimately maximize my utility in terms of dining at these restaurants? Or do I exploit the knowledge that I have in order to have the best dinner that I already know I could have tonight?

What machine-learning algorithms do when they're solving this problem is recognize that the thing you should be doing is exploring more when you first arrive in the city and exploiting more the longer you are in the city. The value of that new information decreases over time. You're less likely to find a place that is better than the places you've seen so far, and the number of opportunities that you're going to have to exploit that knowledge is decreasing.

That tradeoff also appears in human decision making over our lifespan. If we're going to face similar kinds of decisions—we have the same kinds of objects in our environment, which is something that is going to be relatively constant throughout our life—then we want to weight our exploring to the first part of our lifespan, and weight our exploiting to the second part of it.

My colleague Alison Gopnik, who has been pursuing this, has a hypothesis about cognitive development. When we look at children, that variability and randomness that we see is exactly a rational response to the structure of the problems they're trying to solve. If they're trying to figure out what are the things in their environment that they will most enjoy, then putting everything in their mouth is a pretty good strategy in terms of maximizing their exploration.

Working with Alison and our students, we've done a few studies that have looked at how this picture of human learning as solving a statistical problem changes as we look at individuals that are at different points in that developmental process. When you think about solving inductive problems and characterizing inductive biases, thinking about them in statistical terms says that there's a principle of conservation of learning, that you can only be good at learning certain kinds of things.

If learning well is a matter of having biases that point you towards particular solutions, then being pointed towards one solution is going to point you away from another one. If as adults we're converging on a more tightly wound model of how the world works, that's the thing that gives us the maximum inductive leverage in the world that we're operating in, then what we expect to see is that children will be much more flexible as learners. Children are not going to be as strongly committed to the kinds of hypotheses that we are. That's basically what we see across a few different kinds of tasks.

One example in the context of causal learning is that adults have an expectation that if you've got two things that could potentially cause something to happen, normally, those two causes operate independently. So, if I put some things on my blicket machine, and these things make the machine light up and play music, the assumption that adults will make by default is that each of those things have the capacity to make the machine light up and play music; those things were acting independently to produce that effect. That's a good assumption in the world that we live in.

If you flip a light switch, those switches are things that directly affect the light. There's a relative amount of independence of causes in the world we live in. It turns out that if you have a causal system that doesn't work like that, one where you take two objects and put them on the machine in order for it to light up and play music, then kids can figure that out quicker than adults do. That's something which violates the inductive biases of adults, but kids haven't acquired those same kinds of biases. As a consequence, they're faster learners.

That's consistent as well with this explore-exploit framing. As they're starting out, kids have a much more diffuse expectation about how the world works, and that gives them more flexibility to discover different kinds of relationships that could exist.

~ ~ ~ ~

At Berkeley, I'm affiliated with psychology, cognitive science, neuroscience, and computer science in one way or another. Those are all audiences that this work connects to. Most of what we do is write scientific papers that introduce those ideas to those audiences. We grapple with these deep questions about how human cognition works, how we understand the things people are doing, and how we can make people better at solving those problems.

We've recently started to reach out to a broader audience. This is a consequence of the fact that we're at a moment where there's a unique opportunity for psychology and cognitive science to have a broader impact. In the technology industries right now, there's a lot of data on human behavior. When you go to a website, often the company that has put out that website is collecting information about what you look at and what you click on. They're trying to figure out information about you that they can use to show you the right ads and make recommendations of the right products. They use information about your behavior to make inferences about your preferences and desires, and then figure out how they can best satisfy those (and take some of your dollars in the process).

How do I make recommendations to somebody? How do I identify people who other people will want to be friends with? How do I figure out, based on their actions, what people are interested in? How do I figure out what kinds of things they will apply a tag to, what kinds of images they'll label in a particular way? These are all problems that are fundamentally psychological problems. But the way that they're being tackled is largely as computer-science problems.

There's an opportunity that goes in both directions, in the direction from academia to industry and from industry to academia. The reason I say that is because most of these kinds of data are being used in a relatively superficial way. To give an analogy: The current state of data science is in the state that psychology was in during the first half of the 20th century.

In the first half of the 20th century, it was disreputable to try to study how the mind works because minds were things that you never saw or touched or intervened on. What you could see was behavior and the environment that induces that behavior, so the behaviorist psychologists said, "Let's get rid of the mind. Let's just focus on these mappings from environment to behavior." That's where a lot of behavioral data science is. If I show you this, then you click on this. If you've seen these webpages, then you're likely to go to this webpage. It's a very behaviorist conception of what underlies the way that people are acting.

In the 1950s, a new way of thinking about psychology and cognitive science was introduced, which was to talk about how minds work. The thing that made that possible was mathematics, having good formal mathematical theories that could be used to describe how you could have an intervening variable between the environment and behavior. What cognitive scientists and psychologists are experts at is figuring out the structure of those intervening variables, putting the right things between environments and behavior. I see an opportunity there for making data science richer, and to engage more with the models of cognition that would hopefully result in more effective predictive models as well.

On the other side, going from industry to academia, a lot of the data that's being collected by these companies is being kept as proprietary data; it's not necessarily used in ways that will give us the kinds of scientific insights that it might support.

For example, I have two daughters, and when my first daughter was born, my wife and I started using an app to keep track of her sleep. After doing this for a while, I realized that the company that had that app had more data on infant sleep than every study that had ever been run by psychologists. There's a huge opportunity there to understand things about development, but more broadly how people learn and think by using these sources of data.

Psychologists don't normally think about using data like this. The way that a psychologist answers a question is by running an experiment, maybe with some undergraduates, maybe online, and using the results of that experiment to tease out a particular hypothesis. The analogy I make is that this is much more like astronomy. You don't get to intervene; you only get to observe. The observations you get are very large scale and noisy. But that doesn't mean there's no scientific value in those observations. It sets up a new set of challenges for how we pursue psychology in the 21st century, which are about how we make the most of these rich but complicated datasets that characterize the nature of human behavior.

We started a few enterprises that try to engage with that. My postdoc Alex Paxton and I have been working on a website called dataonthemind.org, which collects a huge amount of behavioral datasets that have been publicly released. Those publicly released datasets have been tagged with different aspects of cognition. If you're a psychologist and you want to understand how attention works, you can go to the website, click on "attention," and it shows you a whole list of datasets that we think tell you something about human attention. Then you have to figure out how to use this to answer research questions.

That bridges what we call the "imagination gap," which is the gap between wanting to solve a problem and being able to imagine how to use these different sources of data to solve that problem. The second gap we call the "knowledge gap," which explores how to get the skills to be able to do that. We also have been putting together video tutorials about how to work with large datasets.

The third gap is what we call the "culture gap," which is just helping people to recognize that this is a good way of doing psychological research, and on the other side, helping people in industry to recognize that there's value in working with academic psychologists and cognitive scientists to try to solve these kinds of problems. That's where we are now, starting to reach out to companies and say, "The kinds of data that you have would be scientifically useful, and we also think that the kinds of science that we do can be useful from a business perspective."

Rather than trying to do this in a way where we're focused on one particular company, the advantage of being in academia is that we can work with many different companies. We don't have to commit to having one kind of data. Part of my motivation in doing this is that I have also worked in machine-learning research. Machine learning has gone through this rapid transition over the last decade. Maybe ten years ago, most people who were doing machine-learning research were in academia, with some presence in industry. Over the last ten years, machine learning has become more and more important for companies. As a consequence, there's been this big shift of machine-learning researchers from academia into industry. We're now at the point where there are certain kinds of problems where, if you want to work on them, you pretty much need to be in industry to do so because that's where the datasets and the computational resources are.

Ten years into the future, when I'm thinking about what would be the next thing that might be like that, I think of the social sciences, including psychology and cognitive science. At the moment, most research being done in those disciplines is being done in academia, but the kinds of data that companies have are becoming increasingly important to being able to answer certain kinds of questions.

For me, there's a goal of trying to preempt getting into a situation where the only way to answer those questions is by making a commitment to work at a particular company, by trying to establish some norms about how those kinds of data are shared, used, and made available to academic researchers. Dataonthemind.org is a mechanism that hopefully will be a way of doing that.

~ ~ ~ ~

I grew up in Australia. I was born in London, and my parents moved to Australia when I was eight years old. I did my undergraduate degree at the University of Western Australia, in Perth, which has a reputation as being the most isolated capital city on earth. It's a long way from anything else, but it's also a great place to grow up. In Australia, in the last year of high school, you have to make a decision about what you want to study at university. It was 1994, I was sixteen years old, and I had no idea what I wanted to do. I knew that I liked math, but I certainly didn't want to make a commitment to doing that for the rest of my life. I said, "Okay, I'll study the things that we don't know anything about—philosophy, psychology, anthropology." That was what I went to university to do.

A couple of years into that degree, I was reading a philosophy book by Paul Churchland called Matter and Consciousness. Right at the back of that book, there's this chapter on neural network models, and I was amazed. It was like, okay, this is fantastic: You can use mathematics to describe things like how brains and minds work. I decided right there that that's what I wanted to do. I spent the summer reading all sorts of books about neural networks and mathematical models of cognition.

On the first day of the semester, I cornered the guy who I'd identified at the university as working on that topic convinced him to let me into his lab. While I was working there, I got the chance to get involved in research and studying these kinds of things. I knew that there was a lot that I wanted to learn about computer science, and statistics, and these other disciplines.

Then when I applied to graduate school, I went to Stanford University, where I worked with Josh Tenenbaum. In the process of doing my PhD, I had the chance to do a Master's degree in statistics, which was statistics, computer science, and machine learning. That gave me the tools to be able to do the kind of research that I do today. I worked with Josh at Stanford and MIT. Then I went to Brown University, where I started to teach, and I met some of the colleagues who continue to be good friends and collaborators today.

~ ~ ~ ~

The real important ideas here are that, first of all, we can learn things that are relevant to making computers better at solving problems and smarter by studying human cognition. That's a view that has oscillated in the AI community in terms of how much people believe that or not. In the early days of AI, it was closely tied to cognitive science. In the 1950s, the very first AI paper was also the first computational models of cognition paper. Allen Newell and Herb Simon did this work on the Logic Theorist, which was a theorem-proving system, but it was inspired by how humans solve that kind of problem. Those two disciplines were tied together from the start.

It's very clear that in order to make progress in understanding some of the most challenging and important things about intelligence, studying the best example we have of an intelligent system is a way to do that. Often, people who argue against that make the analogy that if we were trying to understand how to build jet airplanes, then starting with birds is not necessarily a good way to do that.

That analogy is pretty telling. The thing that's critical to both making jet airplanes work and making birds fly is the structure of the underlying problem that they're solving. That problem is keeping an object airborne, and the structure of that problem is constrained by aerodynamics. By studying how birds fly and the structure of their wings, you can learn something important about aerodynamics. And what you learn about aerodynamics is equally relevant to then being able to make jet engines.

The kind of work that I do is focused on trying to identify the equivalent of aerodynamics for cognition. What are the real abstract mathematical principles that constrain intelligence? What can we learn about those principles by studying human beings?

Over the last few years, there have been significant advances in AI, in particular, in solving certain kinds of problems. There are problems that involve doing things with images, text, and problems that involve learning to play games or other kinds of reinforcement-learning problems, where you have an agent who's just getting a reward for pursuing different strategies, which also translates to things like robotics. In each of those domains, there have been huge advances as a consequence of using neural network models that are very large, that are trained on very large amounts of data that take advantage of large amounts of computation.

Where I'd say the challenge lies is in seeing to what extent that same kind of modeling perspective can help us solve problems that require reasoning, not about images, text, or reward, but about things like human behavior. One of the ways in which human beings still outperform computers is in being able to solve problems of reasoning about why you did the thing you did, what you're going to do next, what the underlying reasons were behind things that you did.

For those kinds of problems, it seems like the symbolic nature of being able to think about the thoughts that another person is having seems like an intrinsic aspect of it. It also requires having a way of reasoning about this link between goals and behavior, and that's something which is traditionally filled by a model of rational action, that you wanted to do a particular thing and that's the reason why you did it. That's a rational consequence of the desires that you had leading you to act in a particular way.

Having an understanding of what the nature and limits are of human rationality is critical to being able to make computers reason about those kinds of things. And making computers reason about those kinds of things is critical to having computers interact with humans in ways that are mutually beneficial, that are engaging with some of these concerns people have about things like AI safety.

It's more important than ever to understand what makes people behave in the ways that they do and to be able to describe that in mathematical terms because it gives us the tools for building that bridge between humans and machines.

There's a new set of challenges raised by machines becoming more intelligent. Some of those challenges—you could think about these as psychological challenges—involve how to interact with those intelligent machines in a way that's effective, and understand how it changes the way we conceive of ourselves. Those are things that I've been thinking about increasingly. They're things that I'm not particularly worried about. We as human beings are used to being surrounded by intelligent systems whose thoughts are opaque to us. It's just that normally those intelligent systems are human beings.

One of the things we need to be able to do is establish enough context and comprehensibility in the ways that machines act. Human beings are able to use the resources that we're used to using for reasoning about other people as mechanisms for reasoning about the actions that those intelligent systems are going to take. One of the challenges there is that if it's possible to make machines that are more intelligent than people, you start to run into these issues of it being hard for us to be able to reason about the motives that underlie their behavior.

You can already see this in restricted domains; for example, in the AlphaGo system that DeepMind had for playing Go, which is something where the moves that it makes are things that can seem relatively opaque to a human being because those moves are motivated by, many steps into the future, resulting in a slightly increased probability of winning the game. That's a form of motivation that outstrips the cognitive capacities that we have.

This is a moment where we are starting to recognize that we're going to need to interact with systems that, at least in restricted domains, are going to be smarter than us. Thinking about how to design those interfaces between humans and machines in ways that make it possible for us to interact with those systems in a way that allows us to function effectively is an important research challenge, and a significant social challenge.

What’s Related

People

Tom Griffiths

Professor, Psychology and Cognitive Science, University of...

Mentioned

Daniel Kahneman

Recipient, Nobel Prize in Economics, 2002; Eugene Higgins...

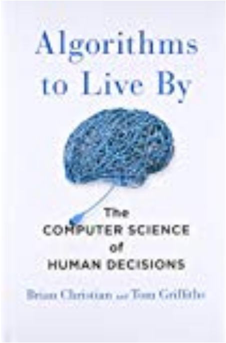
Alison Gopnik

Psychologist, UC, Berkeley; Author, The Gardener and the...

Beyond Edge

Tom Griffiths' Website

Books



Algorithms to Live By: The Computer Science...

By **Tom Griffiths** Hardcover

[2016]

Tags

AI

bounded optimality

cognition

machine learning

psychology

reasoning

John Brockman, Editor and Publisher
Russell Weinberger, Associate Publisher
Nina Stegeman, Associate Editor

Contact Info:editor@edge.org

Edge.org is a nonprofit private operating foundation under Section 501(c)(3) of the Internal Revenue Code.