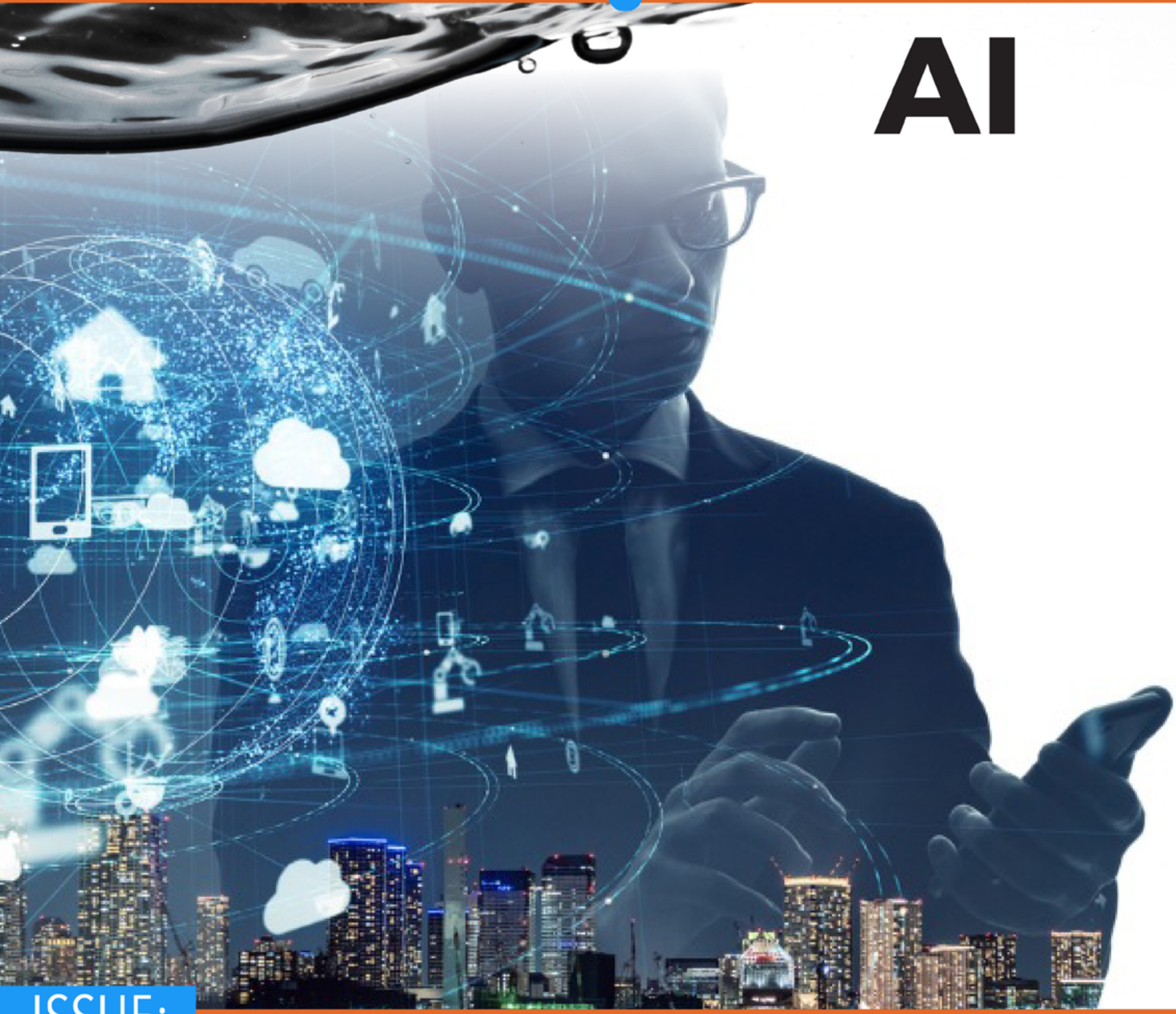
DEEP DIVE INTO:

AI



ISSUE:

1 Explainable Artificial Intelligence

A monthly in-depth report on AI, written by Byron Reese.

SUBSCRIBE NOW

GIGAOM

Issue 1 - Explainable Artificial Intelligence

As artificial intelligence becomes more powerful and is used in more places of greater importance, the question of why an artificial intelligence (AI) makes the recommendation or choice that it does becomes ever more relevant.

The challenges of creating explainable artificial intelligence (XAI) are numerous and potentially insurmountable. Yet, social and legislative burdens are being placed on companies to provide XAI. How will this all unfold? Let's dive in.

What is XAI?

The simplest kinds of AI don't even look like AI. These include such simple devices as a cat food dish that refills automatically or a sprinkler system that comes on when your lawn is dry. In these devices, the logic is quite simple. With regard to the food dish, if the collective weight of the food falls below some threshold, the refilling mechanism is triggered.

But for more complex systems, the very kind we are building today, the logic isn't simple, and in fact might rely on the subtle interplay of thousands of variables. When such systems make decisions that affect people's lives – such as the denial of a home loan – people want a humanly understandable explanation as to exactly why the AI came to the conclusion that it did. These explanations are what we call explainable AI, or XAI. AIs without explainability are referred to as black boxes, a name evoking data going in one side and results coming out the other side, with complete opacity about what is going on in the black box itself.

Explainability isn't a product. It is a feature of a system that uses artificial intelligence. However, it's not a bolt-on feature like running boards on a truck. It is a fundamental design decision that begins with understanding the data that is being used to train AI, then choosing the proper type of decision engine, and finally selecting algorithms that explain decisions after they are made. These three steps are often referred to as data explainability, model explainability, and post hoc explainability. Most of the focus in the AI community has focused on post hoc explainability.

Explainable Artificial Intelligence sounds fairly straight forward until you start asking some basic questions, such as: "What exactly is an explanation?" and "Understandable to which humans?"

How we got here.

The need for XAI was not a top-of-mind issue for most industry experts until relatively recently. For decades, explainability has been a non-issue because the models and techniques we used were simple enough that a conclusion could be understood with a little dedicated inquiry.

In an expert system, for instance, the exact decision tree that the computer went through would be easily available to anyone who wanted to take the time to reproduce what the computer did. Computer systems, until relatively recently, were simply stand-ins for people. We used them not because they came to decisions substantially better than a human would, but simply to do monotonous computation faster and more reliably than a human.

This type of dominant use recently has changed. Machine learning, which focuses on finding patterns in data, has no particular interest in *understanding the data*. This evolution has resulted in the creation of models with complexity vastly beyond human understanding. The situation we find ourselves in is akin to a weatherman being asked to explain why a hurricane took a certain path. To be sure, it is a 'knowable' thing, at least in theory. Well-understood natural forces are the only factors that govern the hurricane's trajectory, but practically speaking, no human is capable of explaining exactly why one particular storm took the path it did, except in the most general of terms.

Machine learning techniques tacitly acknowledge this limitation. Programmers don't generally build systems then try to understand the output; rather they merely test to see if the output is correct. Data sets are often broken up into two halves, a training half and a testing half. Models are developed with the first one, and then that model is run on the second half of the data. If the results from the testing set are consistent with what was observed in the training set, we say that the system works. But never is it a requirement that we understand *why* the model produces accurate results. Thus, decisions are seldom explainable.

Consider this example. You operate a pool cleaning service in Austin, Texas. When you do a search on Google for "pool cleaning Austin," you come up #5, whereas your main competitor comes up #4.

What would happen if you found a Google engineer and put a question to them: "Why do I come up #5 but my competitor is #4?" The engineer would likely shrug and answer, "Who knows? We index fifty billion pages, and you want to know why one is #4 and one is #5? There are thousands of factors that go into the rank, many of which are simply correlations that have been observed. And they change constantly, as do the various weightings of them."

Amir Khosrowshahi

On my podcast **Voices in AI**, Amir Khosrowshahi, a VP at Intel and the CTO of its AI products, summed up the situation we find ourselves in this way:

"There's been a high emphasis on performance of machine learning models, and that's been at the cost of other things, and one of those things is transparency and explainability. I think what's happening now, is that in the process of building machine learning systems, the machine learning researcher has to understand what they're doing, such that they can make better models."

The argument for XAI

We live lives ever more governed by algorithms. The media we consume is suggested to us by algorithms, as are the places we eat. T he products we buy are suggested by algorithms, as are the ads we see. The programs we stream, the movies we go to, and the music we hear, are all heavily driven by algorithms. The temperature our home is kept at may be governed by a smart thermometer. So is the route we take to work, as well as the person we are matched with in dating apps, which resumes we consider for jobs, and a hundred other things in our everyday life.

Some say we shouldn't worry about the machines "taking over" our lives. It is far more likely that control over our lives will be thrust upon them by us, relieving us of the tedium of making untold numbers of decisions every day, often with little more to go on than our gut as our guide. The entertainer Keith Lowell Jensen captured a bit of this when he said of the book *1984*, "What Orwell failed to see was that we'd go out and buy the cameras ourselves and that our biggest fear would be that nobody's watching."

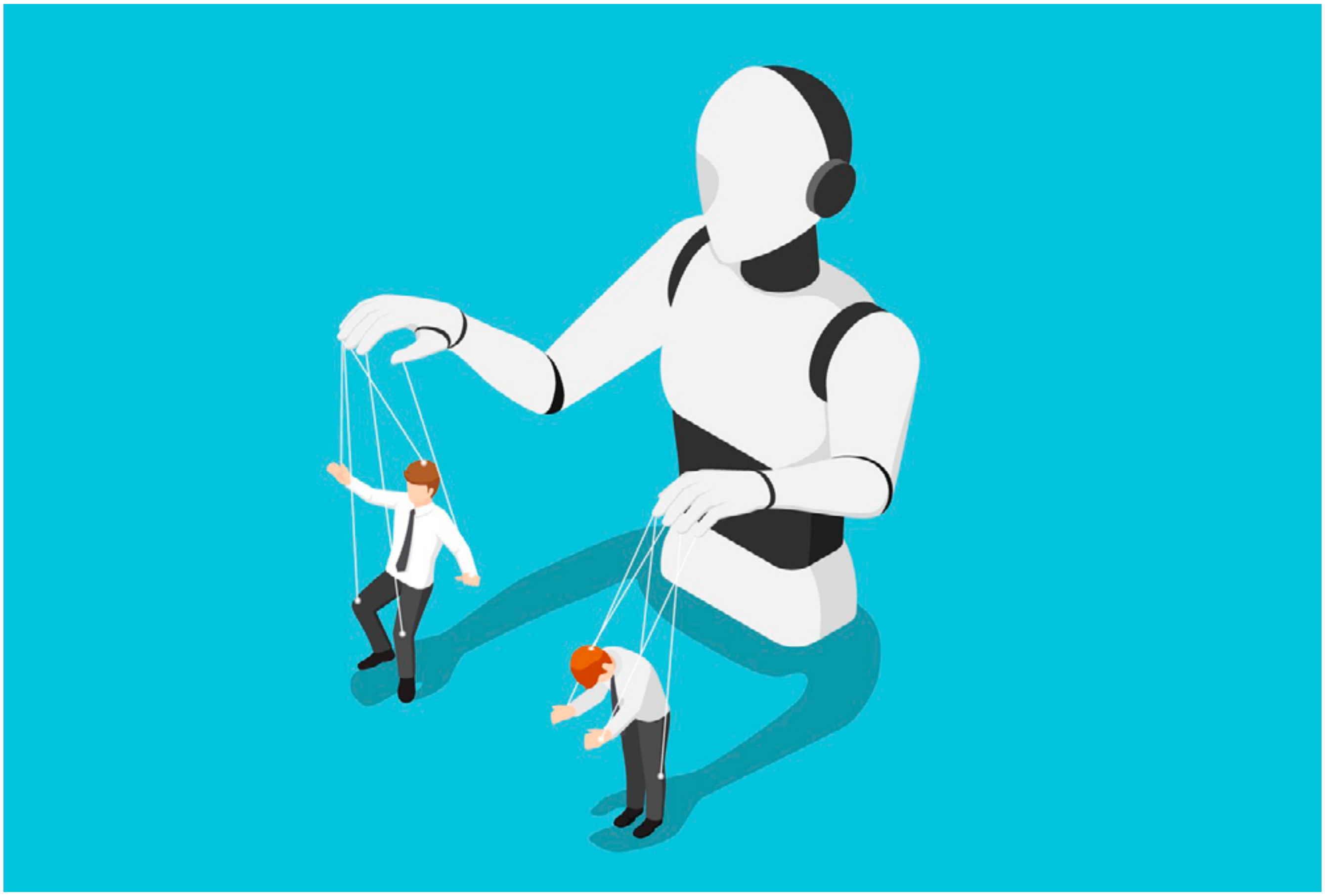
Still, many people are torn between two opposing views about the increased dominance of algorithms over our daily lives. On the one hand, the value of the algorithms, and the potential for them to better our lives, is widely acknowledged. But along with that feeling comes a sense that we have lost something along the way... that some amount of agency over our own lives is gone, and by reason of that, the world has become a less understandable place, where every day more decisions are made *for us* as opposed to *by us*. From this standpoint, the desire to understand how the decisions that govern our lives are made is completely understandable.

But that desire for control is not the only reason people want AI to be explainable. Layered on top of that is a fundamental distrust of the underlying system and those who operate it. There are twin worries that either the algorithms don't work correctly, or that their results are being manipulated by those with ulterior motives.

> ***Is some self-proclaimed know-it-all algorithm giving us flawed advice? Maybe. After all, how would we know? Are we being deliberately manipulated by a puppeteer pulling our strings, tricking us into making choices that are not in our best interest, but in theirs? Perhaps. It is certainly possible.***

It is these concerns that drive the desire for XAI. They are entirely understandable and reasonable, and they are founded on completely legitimate concerns. AIs do have an increasing impact on our lives. They are imperfect and they are sometimes created to further hidden agendas. But explainability has all kinds of significant, and potentially insurmountable, challenges. These challenges are not part of some conspiracy to keep information from people, but reflect a reality that truly explainable AI is at least quite hard, and perhaps even impossible.

Where does our distrust of AI come from? Other technologies have an impact on our lives, but we trust them. Primarily, this lack of trust is a by-product of the newness of the technology. GPS has reached a point where people will often follow it even if they think they know better. But this reliance occurs because GPS has been in widespread use for two decades, and for the most part, our collective experience with it has allowed it to earn a certain amount of trust. If it regularly failed, and sent people to Portland Oregon instead of Portland Maine, or even worse, drove them into a lake, we would be understandably hesitant to blindly follow it.



Unfortunately, AI is new enough that often times it is buggy. In fact, often we are pleasantly surprised when it does work. And its failures are often both embarrassing and widely reported. Consider these recent, high-profile news stories of purported AI failures:

- "Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam" - *The New York Times*, March 19, 2018
- "Chinese businesswoman accused of jaywalking after AI camera spots her face on an advert" - *The Telegraph*, November 25, 2018
- "Passport robot tells Asian man his eyes are closed" - *New York Post*, December 7, 2016
- "Amazon's Alexa started ordering people dollhouses after hearing its name on TV" - *The Verge*, January 2017.
- "Microsoft's racist chatbot returns with drug-smoking Twitter meltdown" - *The Guardian*, March 30, 2016
- "Crime-fighting robot hits, rolls over child at Silicon Valley mall" - *The Los Angeles Times*, July 14, 2016
- "This $150 mask beat Face ID on the iPhone X" - *The Verge*, November 13, 2017
- "IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show" - *Stat News*, July 25, 2018
- "Toddler asks Amazon's Alexa to play song but gets porn instead" - *NY Post*, December 30, 2016

Copyright Warner Bros.

It doesn't help that AI failures are fertile territory for blockbuster movies. A few of the many examples include Bladerunner, 2001, Ex Machina, Metropolis and Will Smith's I, Robot. The situations in these movies are obviously not real data points against AI, but tend to undermine trust in AI due to a human cognitive bias called "reasoning from fictional evidence," whereby these movies make AI seem less reliable.

In addition to distrust of the technology itself, there is often distrust of the motives of the institutions deploying AI. This is also a common theme of movies. In X-Men: Days of Future Past, a powerful corporation, Trask Industries, makes robots called Sentinels, originally created to kill mutants that then came to hunt all of mankind. In addition, there's Cyberdyne Systems which built the Terminators, Weyland-Yutani Corporation that is the evil corporation in the Aliens franchise, and even Stark Industries, the good guys in The Avengers franchise, made an AI named Ultron that decided to destroy humanity after being plugged into the Internet for just a few minutes.

But you don't have to turn to science fiction for stories along these lines. It happens in real life. A few recent headlines will suffice:

- "How Artificial Intelligence is Being Misused to Harm Students" - *Forbes*, July 16, 2018
- "China reportedly using secret AI system to track Muslims" - *New York* Post, April 16, 2019
- "The US Army wants to turn tanks into AI-powered killing machines" - *Quartz*, February 26, 2019
- "The rise of the KILLER ROBOTS: Armed machines 'could guard North Korea border'" - *Express*, August 27, 2017

The difficulties of XAI

So why is XAI difficult? To begin with, not all AI models are hard to explain. Certain models lend themselves to explainability. Others do not. For example, an AI that decides whether you have a cold or the flu might be a simple decision tree, beginning with the question, "Do you have a fever?" After that, it might ask about aches and pains, and a few other variables. If our AIs were this straightforward, explainability wouldn't be an issue. But AI wouldn't be all that powerful either.

However, the recent advances that we have made in AI, the kinds that are making everyone so excited by the technology, don't work this way at all. They rely on taking vast amounts of data and finding patterns in that data. The underlying models are often agnostic to the task being performed. The computer doesn't know if it is learning how to spot cats or cancer. It simply is trying to solve a problem, which is, "Given this set of data and these known outcomes, how could I have best predicted those outcomes from that data?"

There is never a question of *why* that particular data produces that outcome. It is simply a fact that it does. Sometimes a narrative explaining the results can be imposed on data, which we will discuss later, but even in this case, it is not necessarily, nor perhaps even likely, that that narrative is true; that it captures the underlying causality of the real world. It simply provides a story on which to hang the correlations that are discovered by the AI.

In *The Empire Strikes Back*, Yoda famously told Luke that, "There is no 'try'." But with AI, it is not a stretch to say that there is no 'why.' There simply 'is.' This model makes this prediction. Why? Because that's what follows from the data.

To make matters worse, generally speaking, the more accurate an AI is, the less explainable it is. Simple AI, like those that do rudimentary classification or which are based on decision trees, are usually understandable by humans. Generally, linear models and Boolean rule sets are highly explainable. But as you work up in complexity and accuracy, as is the case with graphical models, ensemble methods, and neural nets, explainability becomes far harder.

The argument against XAI

The main argument against XAI runs like this: "If AI is required to be explainable, then we are explicitly limiting the science to merely human-level performance. If a human can understand the model, the human could, in theory, replicate it with pen and paper. This massively shortchanges the technology and shackles an otherwise powerful technology."

The reasoning continues this way:
"This overarching requirement, that any AI must be explainable to be used, would completely stymie the science. All of the things AI can bring us, all of the ways it can improve our lives and even save lives, will be thrown to the wayside all because we have a belief that if we can't understand it then we cannot trust it."

This sentiment was succinctly expressed by Pedro Domingos, AI super guru and author of the book *The Master Algorithm*, when he tweeted in early 2018:

"Starting May 25, the European Union will require algorithms to explain their output, making deep learning illegal."

Groucho Marx famously resigned from the Friars' Club with the statement: "I don't want to belong to any club that would accept me as one of its members." Likewise, many in the AI world wouldn't want to use any AI that humans could understand. What would be the point of that?

In addition, the argument is that since advanced AI is inherently unexplainable, a requirement for an explanation will result in a kind of pseudoscience, that is: things that look like explanations but really aren't. This will give some illusion of explainability but at its core is an assertion that's simply not true. In fact, they do little more than dupe the unsophisticated into thinking they understand the decisions of the AI.

Critics of explainability point out that much of what helps us in our modern world isn't explainable either. We didn't know how aspirin worked when it first came out, nor did we understand how penicillin works, and we still don't know how acetaminophen stops pain today, nor how general anesthetics work. We don't know why placebos work, even when people know they are placebos.

The irony of the situation is that we live in a world where AI might suggest a treatment that we ignore because it isn't explainable, but we might still use a medicine that we don't understand either.

Does AI need to be explainable?

So what is the net of all of this? Does AI need to be explainable?

As host of the VoicesinAI podcast, I asked two giants in the AI field this question, and found their answers to be informative. The first is Manoj Saxena. He is a huge proponent of the need for explainability, and his credentials are beyond impressive. In addition to being a longtime advocate of "Trusted AI," he is the chairman of both CognitiveScale and AI Global, and Manoj was the first GM of IBM Watson.

The following is an excerpt from **Episode 19 of Voices in AI Podcast:**

Manoj Saxena

Manoj Saxena: "I believe 99.9% of the AI companies today that are funded will not make it in the next three years, because they lack some fundamental capability, like explainability. It's one thing to find pictures of cats on the internet using a deep learning network, it's another thing to explain to a chief risk officer why a particular claim was denied, and the patient died, and now they have a hundred-million-dollar lawsuit. The AI has to be responsible, trustworthy, and explainable; able to say why was that decision made at that time."

Byron Reese: "So wouldn't an explainability requirement impede the development of the technology?"

MS: "Or, it can create a new class of leaders who know how to crack that nut. That's the basis on which we have founded CognitiveScale. [Explainability] is one of the six requirements, that we've talked about, in creating enterprise-grade AI. One of the big things–and I learned this while we were doing Watson–was how do you build AI systems you can trust, as a human being? Explainability is one of them. Another one is recommendations with reasons. When your AI gives you an insight, can it also give you evidence to support, 'Why I'm suggesting this as the best course of action for you'? That builds trust in the AI, and that's when the human being can take action. Evidence and explainability are two of those dimensions that are requirements of enterprise-grade AI and for AI to be successful at large."

BR: "Why aren't we further along with XAI?"

MS: "It's like the Web; I keep going back to Internet. We are like where the Internet was in 1997. There were probably, at that time, only a few thousand people who knew how to develop HTML-based applications or web pages. AI today is where the Internet was in 1996 and 1997, where people were building a web page by hand. It's far different from building a web application, which is connecting a series of these web pages, and orchestrating them to a business process to drive an outcome. That's far different from optimizing that process to an industry, and managing it at the requirement of explainability, governance, and scalability. There is a lot of innovation around enterprise AI that is yet to come about, and we have not even scratched the surface yet.

We are in the most dangerous time right now, where the hype about AI has far exceeded the reality of AI. These AIs are extremely unstable systems today. Like I said before, they are not evidence-based; there is no kill-switch in an AI; there is no explainability; there is no performance that you can really figure out."

On the other hand, I put this same question to Intel's Amir Khosrowshahi, mentioned earlier, who replied in Episode 67 of **Voices in AI** Podcast:

"There's a lot of promise in this direction of building tools to understand models and also to build more explainable and transparent models. This is a really difficult challenge, because if you ask yourself if you want an MRI machine and a diagnostic set of software to identify a potential tumor, do you want it to be explainable or do you want it to be high-performance? And I think you would choose the latter. You don't care how it's doing it, you just want it to be really good. So there is a tradeoff there, and we've sacrificed for performance and I think we're going to start catching up on explainability and transparency."

What is an explanation?

As if this situation was not complex enough, there is the additional wrinkle of exactly what an explanation is.

Artificial intelligence is a term without an agreed-upon definition. What constitutes intelligence itself is a hotly debated term, and the nature of AI's "artificialness" is not agreed upon either. It is unclear, for instance, whether AI really is intelligent or whether it can *mimic* intelligence, or even if those two things are the same.

Likewise, the idea of an explanation with regards to AI is equally nebulous.

Consider other kinds of explanations. What if someone were to ask, "Explain to me how a cell phone works." What does that answer look like?

Is it acceptable to say, "Well, it is a device that has been engineered based on principles of physics to interface with a cellular provider to make phone calls and transmit data." Does that tell you anything? If that is unsatisfying, then consider an explanation that goes into the actual physics of how the phone turns your voice into a signal that can be decoded by another cell phone. But to do this, you have to introduce concepts that themselves need to be explained, like radio waves and asynchronous communication and microprocessors. Those in turn require explanation, which can only be had by introducing more concepts that need to be understood.

So what does the statement, "Explain to me how a cell phone works" actually mean?

Let's consider a simpler example: the aforementioned sprinkler system that comes on when your lawn is dry. What would an explanation of that system look like? "The system comes on when your lawn is dry" isn't really an explanation. How dry? What mechanism is it using to measure dryness? Does that system have implicit bias? What part of the yard is it measuring? What about parts of the yard that aren't dry? Do they get watered? If rain is imminent, does it still come on? The question is almost fractal in nature, the closer you look at it, the more questions emerge.

In the end, it is like the saying "Biology tells you you're 70 percent water, chemistry tells you you're 60 percent oxygen, and physics tells you you're 99.99999999 percent empty space." Which of those is an explanation of what you are made of? And are any of them complete?

XAI is actually more complicated than these simple examples. If you are turned down for a loan, then merely an explanation of why the model turned you down isn't actually all that enlightening. The model was, after all, trained on data. And almost certainly a subset of all the data about loan repayment. So how was that data collected? Was the selection of that data itself biased? Was it interpreted correctly? What forms of statistical analysis were used on the data? What models were used to build the AI? How do they work?

Raj Minhas leads the AI research lab at PARC that focuses on people, their behaviors, and interactions with machines. I discussed this question of what an explanation is with him on **Episode 69 of Voices in AI**:

Raj Minhas

"Explainability is sort of like intelligence. It is one of those things which is a simple thing to understand but hard to explain, and so for us we had to sort of be clear about what we mean by explainability, and explainability for us is a way to, not explain how something will generate it, but why.

So you should be able to say "How would I change things for something to be different?" You should be able to talk about, like in the context of drones, for example might be that, the task was for it to fly autonomously and provision a lost hiker, and it didn't do that. It searched and it came back, and you ask that question: "Well why didn't you go this way, why didn't you go that way?" It doesn't need to explain in terms of its parameters and underlying components, but it needs to be able to say in terms of concepts at a higher level and be able to say, 'There was a fire there and because of the smoke, my sensors weren't able to pick up anything, so I didn't know what to do and therefore the decision was to wait out until the smoke subsides,' which is explaining an action in terms of its consequences on what the task was, rather than explaining how the various parameters came together."

On **Episode 43 of Voices in AI**, I chatted about this topic with Markus Noga, who at the time was SVP of Machine Learning at SAP and now is the SVP of the Cloud Platform Business Services at the same company. I found his remarks insightful:

Markus Noga

"I also think that the quest for an explanation is something that is very human. At the core of us is to continue to ask "why" and "how." That is something that is innate to ourselves when we apply for a job with the company, and we get rejected. We want to know why. And when we apply for a mortgage and we can offer a rate that seems high to us and we want to understand why. That's a natural question, it's a human question, and it's an information need that needs to be served if we don't want to end up in a Kafka-esque future where people don't have a say about their destiny. Certainly, that is hugely important on the one hand."

"On the other hand, we also need to be sure that we don't measure ML and AI to a stricter standard than we measure humans today because that could become an inhibitor to innovation. So, if you ask a company why you didn't get accepted for that job offer, they will probably say, 'Dear Sir or Madam, thank you for your letter. Due to the unusually strong field of candidates for this particular posting, we regret to inform you that certain others are stronger, and we wish you all the best for your continued professional future.' This is what almost every rejection letter reads like today. Are we asking the same kind of explainability from an AI system that is delivering a recommendation today that we apply to a system of humans and computers working together to create a letter like that? Or are we holding them to a much, much higher standard? If it is the first thing, absolutely essential. If it's the second thing, we got to watch whether we're throwing out the baby with the bathwater on this one."

What is an explanation?

Do people know why they make the decisions that they do? Probably not. Brain science suggests that the brain makes decisions outside the framework of our conscious mind, then our consciousness rationalizes the decisions and comes up with plausible reasons. In other words, your brain decides for reasons opaque to you, but your conscious mind comes up with a good reason after the fact that you erroneously believe is the "real" reason. It sure doesn't feel like that, but that seems to be going on in our heads.

But putting that aside for a minute, think of the kinds of explanations that we are used to getting from other humans. Are they "explanations" in any meaningful sense of the word?

Question: Why did you fire me?
Explanation: This is a right to work state. Providing details exposes us to litigation.

Question: Why did I get declined for this loan?
Explanation: We did not have confidence that you would repay it.

Question: Why didn't you renew my lease?
Explanation: We're going to do some renovations to the apartment.

Question: Why did you suspend my visa?
Explanation: That's classified.

Question: Why did you prescribe this medication?
Explanation: People like you tend to respond well to it.

Question: Why didn't you accept me to college?
Explanation: There were many qualified candidates.

Question: Why did the officer use deadly force?
Explanation: He thought his life was in danger.

Question: Why did you make that stock trade?
Explanation: I had a hunch.

Question: Why are you sending me to the electric chair?
Explanation: A jury of twelve of your peers says you're guilty.

Question: Why do you rob banks?
Explanation: Because that's where the money is.

Even in the face of life-altering decisions, the explanations we receive from other humans are woefully vague. Sometimes we aren't entitled to an explanation. Sometimes the explanation is a lie. (For example, the landlord may have no intention of renovating the apartment. He just didn't like the tenant.) Sometimes we receive an incomplete explanation. The doctor may have prescribed a medication because others found it useful, but it may also be because the pharmaceutical company offered to take him and his wife to Hawaii. Frequently, people have no idea why they made a decision so they attribute it to a hunch or a feeling.

So the question we have to ask is the one that Markus Noga proposed in the prior section, which is, 'Are we going to hold machines to a higher standard of explanation than we hold people?' This may have unintended consequences.

Max Welling

In **Episode 82 of Voices in AI** I chatted about this with Max Welling, VP Technologies at Qualcomm and Senior Fellow at the Canadian Institute for Advanced Research. His remarks are entirely on point:

> "[W]e often don't even understand why humans make certain decisions. So if you are out to buy a new home, you visit a whole lot of these homes and you look and you feel how that home feels to you and maybe you have a certain list of things you want to check. But mostly you're taking this decision very intuitively, and if you're then asked why did you take this decision, you will come up with some reasons but it's often [true] that they're not the actual reason why you would make the decision. Researchers have compared people who make these decisions intuitively versus trying to approach it logically, and typically you make worse decisions if you really try to sort of make these decisions logically."

Types of explanations

Putting aside the specifics of what is in an explanation, there are a few distinct types of explanations that apply to AI.

The first distinction is between global and local explanations. Global ones are about how the model works. Local ones are about how a single instance of the model came to a conclusion. The difference is: "How does the model tell the difference between a cold and the flu?" versus "Why did it decide Sally has the flu?"

Another distinction is between directly interpretable models and post-hoc ones. Directly interpretable means you can watch the model unfold, or even predict how it will behave, as opposed to post-hoc models in which the output of the model is evaluated after the fact.

Attributes of a good explanation

From a computer science perspective, an explanation describes the exact process a system uses to turn input into output. It literally steps through the code line by line. Nothing less than that is an explanation. But while that sort of explanation is useful to those in computer science, to the rest of the world it isn't very helpful.

One can judge the depth of the explanation using the following criteria:

- **Audience Accessibility.** Can the intended audience understand the explanation? An explanation that would make complete sense to a data scientist might be useless to an end user, or worse, a regulator.

- **Resemblance.** How closely does the explanation match the actual logic of the AI? "Think of the computer as a groundhog that pops his head out of his hole on a certain day. If it's sunny, his shadow scares him and he retreats into his hole. Your grades are like the sun. If they are strong and bright, you go back to school. Otherwise, you're on your own in the world." This may be accessible to the audience, yet say nothing about the neural network that generated the result.

- **Completeness.** Does the explanation provide all of the data that led to the decision or simply enough to satisfy the audience? An employee review AI may recommend firing an employee. Why? Because he was late three times in a month and that is grounds for firing. That justifies the decision but it isn't complete, because the employee had been accused of stealing; customers complained that he was rude and he parked in the manager's space.

- **Accuracy.** Sometimes a surrogate model such as a decision tree is used to explain a complex neural network. If the surrogate produced the same result as the neural network 99 percent of the time, it would be very accurate (and there would be little use in the NN). If it only produced the same result 50 percent of the time, it would be much less accurate.

- **Suitability.** Does the explanation answer the consumer's question? The person who was turned down for a credit card may not care how the AI works. They simply want to know what they have to change to be approved next time.

Places where explanations are most important

The requirement for exact explanation varies widely depending on how the AI is used. Here are some attributes of use cases that demand a higher degree of explanation. If a faulty AI makes a bad restaurant suggestion to you, you may regard this as a tragedy, but certainly not one on par with a bad cancer treatment regimen.

Places where explanations are most essential include:

High stakes decisions

Decisions that have important consequences need more detailed explanations. No one really cares why a streaming service recommends a song. People do care why drones failed in a search and rescue mission. Examples here include law enforcement and national defense.

Customer Demand

Customers are always right. If they won't buy a black-box product, you must build in explainability.

Ultimate Decisions

If the decision is the last step in a process, it is more likely to require explanation than if it is just one in a series of steps. For example, if an AI detects fraud at a bank, an audit team would come in to investigate. Those humans would make the ultimate decision and they could explain it.

Places

Where the AI's decision might be challenged, particularly in court. "What do you mean I didn't get the loan?"

Regulation

If the law says you have to provide an explanation for the AI, you have to provide an explanation for your AI. This puts developers in a difficult position, because whether or not an explanation is sufficient is generally determined by a regulatory board after the application has been deployed. Unfortunately, the level of explanation should be one of the driving factors in the application's design. It is difficult to modify after the fact. Industries where this is largely the case include banking and finance, as well as medicine.

On **Episode 57 of** **Voices in AI** Akshay (Shay) Sabhikhi, the CEO at Cognitive Scale, chatted with me about explainability in regulated industries, which he regards as essential:

Akshay Sabhikhi

"You are in highly regulated industries like healthcare or financial services, and relying on AI to give you insights, whether it is insights going to your financial clients, or your patients or your compliance department. These have to be auditable; they have to be tracked, and you have to be able to have a stream of a connection between why this insight was delivered in the first place … [M]any of the industries we work in regulate even the insights that are being delivered from the AI to the consumer, but bring a human in the loop. That's still important today because you haven't yet trusted the machine."

I pressed him on this, asking if this level of explainability was going to slow down advances in AI:

"You know, not really and I gave you sort of a compliance example about heavily regulated industries. Let's talk about the consumer dimension where you don't have to go as far as to make it explainable. But think about that also for a second. We all know, even as consumers we get these annoying ads that follow us everywhere–if I happened to have shopped for a vacuum cleaner for example, right? And you're annoyed because it's something that's distracting you and you're like, 'man, this stuff is creepy.'+

"Now look, if I was given control even as a consumer, and I'm talking about a pure B2C type of environment, not even a B2B of a way you deal with businesses. Only a consumer-to-consumer or in a use case, you still have to build trust for the consumer. And so we look at explainability, so if I'm putting an insight from someone around, some recommendation for a restaurant or an event or a diet or an activity that they should do, at the end of the day I have so many things being told to me, but if I bring it in my context, and I'll tell you: 'Here's why I think it's really important for you,' I may turn explainability off because I start trusting the system, but initially to get going, what explainability does is build trust and then what that drives is adoption. What adoption drives is now you have more feedback signals and the system gets smarter, right? So I think it's intertwined. In fact we made explainability almost a core component of how we deliver AI and frankly this goes beyond just a business scenario that I mentioned to you.

This particular case for XAI – regulated industries – is unquestionably the biggest driver of explainability in the private sector, given that the medical and financial sectors are among the biggest investors in AI technology.

There are two special cases of explainability that deserve some mention.

The first is national defense. While one might suppose that automated systems that independently make kill decisions would be an area that would necessitate XAI, this probably isn't going to be the case.

If the institutions charged with waging war or protecting a nation from cyber attacks have a choice between an explainable system with some level of performance or a black box with better performance, it is a safe bet they will choose the latter. There are simply some areas where the stakes are so high that any perceived tradeoff of performance for explainability will not be deemed to be wise. Those who manage these systems will likely always opt for the highest level or performance even in the face of total opacity.

The other special case is law enforcement. While there is a substantial amount of concern about AI systems that racially profile, within the broader law enforcement world that issue doesn't really present itself as often. Whether it be money laundering, child pornography, or human trafficking, explainability might also be deemed a secondary luxury compared to system performance.

The larger issues of using AI to do facial recognition to find criminals or scan all email traffic to uncover terrorist plots are beyond the scope of this issue, but will be addressed in an upcoming issue on privacy.

Methods to achieve explainability

So, how do you do achieve explainability given all the difficulties explored earlier? Let's look at nine choices.

Possibility 1 – "It's a black box, trust us."

The first option is simply to say that there is no explainability. This may sound like a dodge, but in many arenas, this is more than adequate. I don't require explainability from my spam filter or my GPS for reasons we have already discussed. So in many applications, explainability is superfluous.

Possibility 2 – "We can't explain it, but here are stats about how well it works."

Another shortcut to explainability is to give consumers statistics about how effective the AI has been in the past. If consumers can be shown a baseball card of statistics about the model, like how accurately it predicts default levels or how much better it is than the older system, then perhaps that can be a substitute for actually understanding the system. If the system can be shown not to discriminate against certain groups, that's icing on the cake.

Naveen Rao

On **Episode79 of** **Voices in AI** I interviewed Naveen G. Rao, Corporate VP and the GM of the Artificial Intelligence Products Group at Intel, and he expressed ideas along these lines:

"I think that's where we need to get to, more bounds around decisions, and "does this system tend to make positive and good decisions vs. not making good decisions? I think that's where we really need to get, and we are getting there in certain areas. Like the visual tools they have in Google already, where you can do a photo search and stuff like that… Yes, there are biases and things like that, which they try to fix as quickly as they can. But we don't necessarily have to ask, 'Why do you categorize this image this way?' It's OK. It doesn't matter. It is what it is. And we'll say, 'Hey, that's probably not right. Go back and fix it.' I don't need a full level of explainability that I had with a regression-based system."

Possibility 3 – Surrogate models

A surrogate is an interpretable model derived from the inputs and outputs of a non-interpretable model. For example, the training set for a neural network that creates a credit score may have ten thousand records with 1000 features each. The resulting AI would be completely uninterpretable.

You could determine the 10 most important features and train a new, interpretable model such as a decision tree, with those 10,000 records with ten features each. You can then run the same test data set through the black box and the decision tree to judge how close the surrogate performs compared to the original.

In other words, run your black box, learn what you can, and then build a simpler model that is explainable that tries to replicate those same results.

Possibility 4 – "Here are the inputs we use."

Another possibility is to worry less about the logic of the model and just reveal the data that is used in it. While not an explanation of the recommendation, this does provide a great deal of transparency. If a college says, "These are all the factors we look at when considering a candidate," you at least have the sense of what is powering the system.

One problem with this approach is that AI systems are likely to use inputs for which there is not an intuitive link between the input and a conclusion. A major credit card company in Canada uses the items that people purchase with the card to increase and decrease their credit line. People who spend money on marriage counseling or debt counseling see their credit line shrink. This is somewhat understandable. But people who buy off brand motor oil see theirs shrink as well compared to those who buy name brand. And people who purchase bird seed see their credit lines expand. Why? The link between bird seed and likelihood to pay isn't evident to a consumer and the knowledge that this is a factor used to assess their creditworthiness may actually give them less confidence in the system.

The kinds of factors that go into an AI model often seem to have little to do with what the model is purporting to measure. In China, for instance, where half a billion people are unbanked, potential lenders can only use data they have on hand to make a credit decision. As such, the factors that go into their models include: the words used in text messages sent and received by the person, the identity of people they are connected with on social media, the type of phone they have, and even how often they plug their phone in. Browsing history is used to determine creditworthiness, as well as an applicant's search history on Baidu. Does the knowledge that these are factors give people more confidence in the system?

For this approach to be useful, the source of the data is important. Selecting data is an editorial decision, and all kinds of biases are implicit in it. There is a famous example demonstrating implicit bias. If you do a Google image search for "unprofessional women's hairstyles," most of the results are images of African Americans. This likely happens because the data Google is likely using is the words around those images (perhaps on social media), such as "Can you believe my boss said this was an unprofessional hairstyle?" thus reinforcing the view that these are unprofessional. Thus, if you simply reveal "We trained the model with examples of unprofessional hairstyles," this actually doesn't tell you much because it only begs the question of how the hairstyle was determined to be unprofessional to begin with.

Possibility 5 – Partial explanation

Perhaps we should give up on the notion of a full explanation and just offer people a partial explanation.

Qualcomm's Max Welling, quoted earlier, suggested this in an interview with me:

> "When we build very, very complex algorithms that look at this sort of complicated set of patterns, then we may have to give up on trying to completely understand how a decision was reached. What we can do is try to come up with a proxy for it. So we could say, 'well we tried to explain in human language the most important reasons why you made that decision.' If we can ask the algorithms to do that, I think that would be quite successful. That would be quite similar to asking a human being or a doctor when the doctor makes a diagnosis. You can tell me why I made that diagnosis. And then we'll come up with reasons, some explanations, but it might not be the whole picture."

Is a partial explanation useful? It would be a mistake to somehow argue that because you cannot understand everything about an AI model, then a better option is to understand nothing. Partial understanding is helpful, in theory. But in practice, if partial understanding appears to be a full explanation, then the model itself takes on a kind of authority that is unwarranted.

PARC's Raj Minhas, quoted earlier, and I discussed partial explanations at length:

Byron Reese: **"If an answer is not understandable by people, explainability can't exist. So if I said, "Why did the hurricane hit Raleigh versus Tampa?" There is an answer to that. It's just physics, but it may be beyond our ability to answer. It may be butterfly wings in South America; it may be beyond our cognitive ability to understand that, and therefore you can't get explainability."**

Raj Minhas: "Can I pull on that thread a little bit? It might be that it says, 'Oh there was a low-pressure area here that caused the change in the direction and caused the weather pattern to move this way rather than that way.' That explanation may be sufficient for a lot of purposes, for making plans, for doing things, but it's not sufficient in the sense that: why did the low pressure exist there and not somewhere else? It wasn't because some butterfly flapped its wings in Argentina, and so it doesn't answer all the way down. It is still useful, so that it says: if there is low pressure formations there, the pressure's like this, the likely impact will be this and so we should plan to be out of its way."

"It's a partial explanation, but it's a useful explanation that you can use to plan to avoid injuries, loss of life, whatever the case might be. To me, a reason for explanation is not simply understanding, it's also ability to do something, and maybe you get partial understanding and you get the ability to do something, and we do that all the time. We have heuristics, we have rules of thumb where we don't understand what's going on, but we can use those to make decisions about the world, and achieve a better outcome than would be achieved without those heuristics that we don't completely understand.

Possibility 6 – Sensitivity analysis

Sensitivity analysis explores how changing certain inputs changes the output of the AI model. It isn't XAI, but it is a popular way to get to something like XAI, for two reasons. First, it is actionable in that it gives you tools to understand how moving certain variables changes the model; and second, it can be arrived at relatively easily through technology. In other words, tools to do it can be productized and sold.

Several AI practitioners I spoke with suggested variants of this. One of them was PARC's Raj Minhas, quoted above. I posed to him my question about how would you explain to someone why they ranked number five on some Google search instead of number four. He replied:

"And that's a good example, so let's discuss that. Again, at least at the level of AI we have, and the level of explanation of intelligence right there, our approach is to sort of narrow the scope of what we're trying to do. One explanation there that may be sufficient for these purposes, and that we may be able to generate using some of the ideas we're coming up with is to explain it in terms of a counterfactual. We may not be able to explain exactly why this came about, but we should be able to sort of talk about some decision boundaries, and so to be able to say, 'What would have to change for you to go to number four? What does that decision boundary look like? What is the smallest increment that you would have to effect for that to change?'"

"That still doesn't give you a broad notion of explainability, but it gives you explainability in the sense that you can act upon that information to make a change in the world, right? So you can say, 'We don't know exactly how this came up, but these are the changes you can make.' For example, in your case, if three more influential websites like CNN were pointing to you, our systems rating would put you at number four, right? So that doesn't give you much insight into how some things will generate, but gives you some insight into where the boundaries are, where decisions change, and what you would need to do is to affect those changes, and so when you think about laws that may acquire explainability, that may be the kinds of explainability you get in the beginning.

On **Episode 62 of Voices in AI**, I spoke on this topic with Atif Kureishy, Global VP for Emerging Practices - AI & Deep Learning at Teradata Corporation. His explains sensitivity further:

Atif Kureishy

"So there's a certain class of approaches and pros and cons with doing that. Another approach, which is something that we've done, is: 'How do you drive interpretability into the deep neural network itself?'

And so there's a whole science of this space, and so we've used different types of open source frameworks out there that allow us to do interpretation, and perturbative techniques that essentially say, 'If I include noise into this model, and be able to express the amount of variants and the amount of contribution that these features have, ultimately to an output, let's say a classification system, then I can understand which features of the model are influencing that output the most.' And that's the technique to then say, 'Which features in a probabilistic way, are the ones that are contributing the most?' And so that's important. Let's say if you're making a determination of fraud, going back to our example, that you can cite that [a decision was made] because you're associated with this actor or this certain amount, or coming from this geography or other types of attributes about the transaction. That's the level of expression you need to pass GDPR or some other sort of audit and policy definition."

That last bit is front and center in many people's minds, including Kureishy's. How much explainability is enough for audit purposes and does sensitivity analysis get you there?

The problem with sensitivity is that the more complete it is, the less explainable it itself becomes, at least to a lay person. Imagine a model that scores people on some scale. It uses a number of factors, which we can call X, Y, and Z. In reality, it might be 1000 factors. But in any case, a sensitivity analysis might say that your score would go up tremendously if X increased and go down a little bit if Y increased, and altering Z has very little effect on your score. That would be useful right? Not necessarily. It might turn out that increasing X and Z together plummets your score. Or when Y is high, lowering X actually increases your score. At some point, the explanation is as complex as the model, especially if you really do have 1000 variables that all interplay.

Think of it like a recipe for cookies. If you double the sugar, you get a worse cookie. If you double the flour, you get a worse cookie, and if you double the eggs, you get a worse cookie. So a sensitivity analysis would suggest that increasing ingredients as a rule is bad. However, if you double everything, you are just fine – you just made a double batch of cookies. Sensitivity analysis would seem to tell you something about that recipe, but it is incomplete at best and misleading at worse.

Sensitivity, like many of these methods, is highly useful to programmers optimizing their systems, but in pursuit of XAI, that is, plain-language explanations of why an AI does something, these methods give the illusion of explanation, but do not necessarily deliver on the substance of it.

Possibility 7 – Interpretive explainability

This is a really interesting one … It doesn't try to explain the model per se, but to construct a plausible explanation for what the model does. Let me explain:

Say you have a program that diagnoses illnesses, and it gets really good. But it is built using all kinds of data sets, even ones that don't seem to have any relationship whatsoever to the illness at hand. However, given enough training data, the diagnostic program becomes quite accurate.

A patient comes in and presents a set of symptoms and the AI outputs a diagnosis. Then the patient says, "Why do you think that?" and we are back to the problem of explainability.

But let's say you got a human doctor and recorded him or her as patients streamed in. Patients present their illnesses; the doctor calls out a diagnosis. A patient asked, "Why do you think that?" and then that's where our AI researcher turns on the tape recorder.

Repeat that a thousand times. Then, connect that tape recorder to the AI so that when the AI hears a set of symptoms and offers a diagnosis, then the bit of audio from when a doctor gave that diagnosis plays out loud. Everyone's happy.

But … the AI came to the conclusion in a completely different manner than the doctor. The "Why?" for the computer is a printout of 1s and 0s, which is quite unsatisfactory to your average patient. On the other hand, the doctor's explanation may not be a sound medical explanation at all. You may come in with aches, a cough and a fever, and the doctor and the AI both conclude you have the flu, and when asked why, the doctor could reply, "Cause you got demons in your blood." That would become the AI's explanation as well.

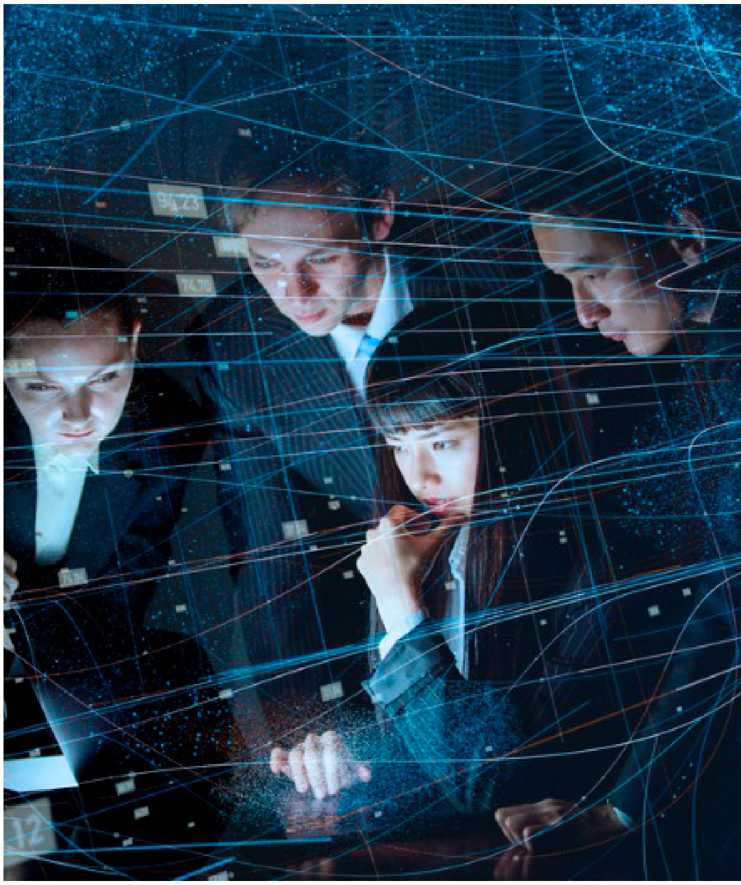
Well, it is an explanation that matches the AI's conclusion. Does it matter that it is a human explanation and not the AI's explanation?

Possibility 8 – Certification of models

What if we gave up on explanations, and instead certified that the machine learning techniques used to create the model are known to be sound. What if – and this is an imperfect analogy – instead of certifying the cake isn't poisonous, instead certifying that none of the ingredients that went into the cake is poisonous? Can AI develop a set of best practices and techniques that if used are deemed to produce results that are accurate and unbiased?

Teradata's Atif Kureishy, quoted earlier, suggested something similar to this to me:

"A lot of times what you'll see, I mean very rarely are we talking about one model, right? These are usually ensembles or stacks of models, in some instances that could go into the hundreds, okay, so being able to coordinate all this is the key critical aspect, but there's different techniques for explainability. So for instance, some enterprises will use traditional machine learning, decision trees, boosted decision trees, to express certain concepts and that are well understood by, let's say, regulators. And that they have techniques for documenting and expressing how certain decisions are made, and then they'll leverage deep learning to essentially do latent feature extraction, so derive new features from these neural networks, that will then be put into a traditional machine learning model, so that they don't have to worry about how to make the neural network itself, explainable."



Possibility 9 – Or some partial combination of the factors above

It might be that there is no silver bullet of explainability, but there could be silver buckshot. It could be that parts of these techniques can be combined to come up with something that provides a level of explainability that we collectively agree is enough. Time will tell us the answer on this. It is important to remember Manoj Saxena's suggestion early in this report that we remember this is like the web in 1996. If you had asked then, "How will we achieve security for online commerce?" you might have been given a number of ideas, and the reality was a blending of them to build the world (and the web) we have today.

Alternatives to explainability

Are there alternatives to explainability? Yes, a few…

Alternative 1 – Keep models simple

If we want models to be explainable, then we will simply have to limit their complexity. It is simple math.

Qualcomm's Max Welling, quoted earlier, suggested this to me:
"The other option would be… and this actually might become quite necessary because of new legislation in Europe on privacy and explainability that you say, 'Okay well maybe I'm just trading in a little bit of performance or quite a bit of performance in favor of full explainability.' So you then have a model that is quite simple. It doesn't look at all these complex patterns, but in fact it is completely explainable.

Alternative 2 – Only use AI in unimportant arenas

If explainability is so essential to us in the big areas, and if we conclude that we can't really get it, we can just decide to use AI only in areas where no one is really all that concerned about explainability, that is, all the less important areas.

Alternative 3 – Give up

Finally, we might simply stop caring about XAI in most cases. I actually think this is the case, as I will explain shortly. For us old-timers who have been online since before the web, we've seen all kinds of fads come and go. It used to matter what your home page was or how many hits your website got. Then we went through a period where "no one in their right mind would enter their credit card online" because of cybercrime. And who would order products they had never seen anyway? All of these and a hundred more objections like them came and went.

Intel's Naveen Rao, cited earlier, seems to think this will happen with explainability as well. He told me:

"I think this whole notion of explainability will diminish with time. Systems are getting just more and more complicated. It's like you say, it's almost impossible already, and you know we don't do this with a human. When I ask a neurosurgeon who has been doing it for 20 years, 'Hey, why did you use that stitch there?' or 'Why did you make this decision in a split second?' do I really care that I can get into their head and pull out the weights of their visual system and this, that and the other thing? No, I don't because I trust that system. That human has been trained sufficiently to make the right decision."

Analysis: Will consumers demand XAI?

So how does all of this net out? Which possibility will come to pass? I suggest there will be three main buckets.

First will be entirely opaque systems such as Google search that people use without any expectation of explainability.

Second will be the highly regulated fields in which XAI will be engineered into the system in an auditable way that's sufficient to persuade regulators and juries that the parties creating these AIs acted responsibly.

Finally, for the bulk of AI use cases, I suspect that faux XAI will predominate. When it comes to systems which are inherently unexplainable, things that look like explanations will be offered. They will be completely true insofar as they go, but they will also be incomplete and misleading in the sense that they won't really explain how the AI came to the conclusion that it did.

While this final choice may seem disappointing, I don't think we will perceive it that way in the future. We already live in a world where we accept major decisions with little understanding of how they come about. To truly understand a medical diagnosis and treatment almost requires we go to medical school ourselves. The division of labor, that marvel which gives us modernity and prosperity, is based on the idea that we don't have know how to do everything or how everything works. The modern world, and by extension much of our lives, is run by experts of all kinds.

In our own fields, we are experts for other people. When a doctor suggests someone take a certain medicine, we don't usually require the doctor to explain all the reasons why they came to prescribe that medicine. Of course, it can be argued that people trust their doctor, but have no reason at all to trust an AI. This is true, of course, but is explainability the only pathway to trust? Perhaps not.

I don't know how it is that airplanes stay in the air given how heavy they are, nor do I understand how the water that comes out of my tap became safe to drink. But when I go to a new city, I don't need to be convinced the water is safe to drink, nor if I fly a new airline do I need to understand how their planes are maintained.

> *The best we generally have in this world and on our journey through it is a façade of understanding, for, like Thomas Edison observed, "We don't know a millionth of one percent about anything." People would go crazy trying to understand the "why" of every decision that affects them. So, while we don't generally talk about how we live our lives with only an illusion of understanding, it is the case, and we seem to be fine with it.*

There is no reason to think that AI will somehow be a new exception to this, that suddenly we will demand a deep and nuanced understanding of the recommendations that ever-more-complicated computers make. Just like in our larger lives, we will draw comfort in the decisions based on the fact that by and large, they work.

In all fairness, my view on this is a minority position in the AI world, or at least a minority position among people's publicly stated public opinions. Many AI experts simply regard XAI as a requirement. But the two questions I would pose to them are: "Over time, will people ask for explanations more often or less often? And, is the younger generation that is coming up going to require explanations the same way that a 50-something person today will?" I think in the answers to these questions, we can get a glimpse of the world to come.

In support of my position, I offer the data point of credit scores. Most people accept that they have a credit score, that it is used to determine many things in their financial world, and that they have no real idea how it is created. We are told a number of factors that go into it, but it is a black box if there ever was one. Yet most people don't lose sleep over this fact.

In fact, credit scores might be a good proxy here. In the event of an adverse action, companies must give you reasons. While the underlying algorithm isn't known, we are given a glimpse of the data that composes it and the relative weightings of that data: Amount owed: 30 percent, length of credit history: 15 percent and so on. But there isn't any explanation beyond this. You are, in the parlance we discussed earlier, given a global post-hoc explanation about how the system works with only a slight peek into your own reason for denial. Is this system fair? Does it discriminate? How was the model formed? Is it audited? What was the training data used on it? Is it updated? What data of mine was used? These are questions we cannot get answers to and for the most part, no longer ask. So I think it will be with XAI.

Regulation of AI

Consumers aside, will governments require explanations? The answer will likely vary among the US, the EU and China. Let's look at each in turn.

In the US, will regulation happen? I put this question to CognitiveScale's CEO Akshay Sabhikhi, cited earlier, who believes it is both good and inevitable.

"So there is absolutely a level of regulation that's required. There's no doubt, because you have systems and the ability today, to go absolutely nuts with the amount of data that exists and frankly you can go deep into social networks and extract information, which is downright creepy, right? So when you talk about regulation and especially what's happening right now in Europe, I think you're going to start seeing some of that here with some of the recent events that have happened. I think they're not a bad thing because we may be crossing a line, and frankly getting the appropriate consent from the end consumer… and frankly I think it's a lot easier to get consent when you're delivering value to people. Things are going to start becoming more important, so we'll see how this plays out frankly, and we're really going to be impacted because we do business in Europe as well and we're going to see how that affects our business."

But will the US actually provide meaningful regulation? Personally, I would be surprised. Our government is not made up of people with a subtle and nuanced view of the intricacies of XAI, and I think there will be enough caution around stifling innovation to discourage regulation. Perhaps Sabhikhi is right, that regulations will be put into place around privacy, but this is an entirely different thing.

I will add an important caveat to that, and that is that such regulation could come from California. The legislature of that state has shown a willingness to regulate in this area with the California Consumer Privacy Act (CCPA), which doesn't mandate XAI, but makes hints in that direction. If California passes an XAI measure, it would effectively be a US-wide policy since most vendors aren't willing to write off the one seventh of the US economy that is found in that state.

What about Europe? There the situation is a little different. The European Union's General Data Protection Regulation (GDPR) is a wide-ranging law which attempts to bring the entire EU under a single set of rules related to various digital issues including privacy. There are also sections of the law that address the right to get an explanation and the requirement of explainable artificial intelligence.

Does the GDPR require explainability? That is a simple question, just four words long, but it has a very long answer. The short version of the answer is that no one knows, because regulatory law and case law haven't clarified some of the ambiguous language in the provision.

If in fact GDPR does have an explainability requirement, it only applies to completely automated decisions; and thus by extension, if human judgement is present anywhere in the process, then the requirement doesn't apply. Also, it doesn't seem to be a post hoc requirement, and individuals don't have the personal right to demand an explanation about their own personal status. That being said, the extent of the explainability provisions is hotly debated, with many people maintaining that there are meaningful explainability requirements scattered throughout the legislation. The main points of contention relate to the language used in the law. When it provides for "meaningful information" to be provided, what does that entail? If it explicitly requires an explanation, what exactly constitutes an explanation?

With regard to the rest of the world, including China, legislation around explainability hasn't become widespread nor even entered into the public discourse in a meaningful way. If it does emerge outside of the US and Europe, the most likely place would be a place like Canada, which is an AI powerhouse, or perhaps from Israel, another bastion of AI innovation.

In addition, AI restrictions might also emerge through self-regulatory action by industry. The insurance industry, for instance, has some rudimentary explanation requirements the US, and it is a safe bet that other industries will conduct forays into such regulations.

DEEPDIVE INTO AI

About Byron Reese

As the CEO and Publisher of GigaOm, Byron leads the company in their mission to help business leaders understand the implications of emerging technologies and their impacts on business, media, and society. Byron produces and hosts GigaOm's podcasts Voices in AI and The AI Minute and has published two books: "Infinite Progress: How Technology and the Internet Will End Ignorance, Disease, Hunger, Poverty, and War" and "The Fourth Age: Smart Robots, Conscious Computers, and the Future of Humanity."

Byron possesses a diverse body of patented work, and enjoys exploring the intersection of technology, history and the future with both technical and non-technical audiences around the world.

LEARN MORE:

bryonreese.com
@byronreese
voicesinai.com
@voicesinai
gigaom.com/shows/

SUBSCRIBE TO **DEEPDIVE INTO AI**

Monthly in-depth reports on AI authored by Byron Reese.

Try free for 60-Days.
SUBSCRIBE HERE

GIGAOM

To learn more about how we help transform enterprises in AI-enriched data-driven world, visit gigaom.com

GigaOm provides technical, operational, and business advice for IT's strategic digital enterprise and business initiatives. Enterprise business leaders, CIOs, and technology organizations partner with GigaOm for practical, actionable, strategic, and visionary advice for modernizing and transforming their business. GigaOm's advice empowers enterprises to successfully compete in an increasingly complicated business atmosphere that requires a solid understanding of constantly changing customer demands.

GigaOm works directly with enterprises both inside and outside of the IT organization. To apply proven research and methodologies designed to avoid pitfalls and roadblocks while balancing risk and innovation. Research methodologies include but are not limited to adoption and benchmarking surveys, use cases, interviews, ROI/TCO, market landscapes, strategic trends, and technical benchmarks. Our analysts possess 20+ years of experience advising a spectrum of clients from early adopters to mainstream enterprises.

GigaOm's perspective is that of the unbiased enterprise practitioner. Through this perspective, GigaOm connects with engaged and loyal subscribers on a deep and meaningful level.