







# AI

Fri, Apr 26, 2024 11:56AM 37:46

## SUMMARY KEYWORDS

ai, model, explainable, black box, glenn, researchers, systems, algorithm, parole, ais, question, users, human, build, decision, trust, output, answer, gpt, good

---

-  00:00  
I was incarcerated as a 16 year old, tried as an adult sentenced to 26 and a half years to life in prison for a robbery and murder.
-  00:12  
This is Glenn Rodriguez, a former inmate at the eastern Correctional Facility in yolk. And when he first went to prison, he was far from a model inmate,
-  00:23  
I had been in quite a bit of trouble earlier on, obviously, there is an adjustment period when you go into a carceral setting. You know, you have people there that are never going home people who pretty much survive off of pretty much preying on others. So you have to pretty much survive in there. But
-  00:40  
as his sentence went on, he managed to turn things around. But
-  00:44  
as I got older, when I got to my mid 20s, I started telling myself well, if I've come this far, I can certainly do the rest of this and I can see myself I can see that little light at the end of the tunnel, right. So pretty much and I managed to turn things around very, very drastically.
-  01:09  
Such a drastic turnaround could mean an early release. But this wasn't quite how Glenn story worked out.



01:25

Now for all intents and purposes, Glenn became a model prisoner. He went to college, he got certifications and counseled at risk schools. And most importantly, he had a completely clean record for over a decade.



01:40

And so by the time I went to parole, I had been out of trouble for at least 11 years.



01:46

In fact, due to his good behavior, Glenn was given a parole hearing six months early. The outcome would be determined by free parole commissioners, who would be presented with information to assess the risk of him reoffending if released. But that decisive information included the output of a bit of software known as compass, which used an algorithm to assess Glen's chances of reoffending. Compass and software like it are used in an attempt to weed out known human biases. But according to Glenn, the opposite was true. The algorithm suggested he was at high risk of reoffending,



02:30

and I was denied parole.



02:36

Glenn, though, wasn't content to leave it at that. He thought that his decade of model behavior should mean that the algorithm would give him a low score. So he did some investigating. Compass uses a questionnaire filled in by an assessor to come up with its fiscal. So we found out how other inmates forms were filled in,



03:01

I think it was 14 of them that were like identical, identical every answer if was answered in the same way, with the exception of question 19. According



03:08

to Glenn, question, 19 was unusual, was most focused on quantitative answers, like how many misdemeanors do you have in the past 24 months? Question. 19 was, in his view, subjective



03:23

and the question is, does this person appear to have notable disciplinary issue?



03:33

And so I don't know what someone appears what that appearance is, but apparently this person who was conducting this assessment for me, like Oh, yes.



03:47

Glen and the other inmates were able to reverse engineer the answers to the questionnaire and the resultant Riscal. And this question appeared to have a lot of weight. According to Glenn, a yes to this question would get you a high score around eight on a 10 point scale, whereas a no would get you a low score around one to three, regardless of the answers to the other questions. Because of this, Glenn and his lawyers attempted to challenge the school and the effect on his parole decision. But here they run into a dead end.



04:23

The weights for the different input factors were claimed as a trade secret.



04:28

This is Rebecca Wexler, a lawyer and researcher Berklee School of Law who works in technology in the criminal justice system and dealt with Glenn's case. According to her when Glenn's legal team challenged the compass score, they were told that the specific weights attributed to different parts of the questionnaire like that question 19 were trade secrets. They could not be shared with Glenn or his legal team, hobbling them from challenging the parole decision.



04:57

Economic prove it But that one answer was significant to the output high risk score that he got.



05:10

Part of the algorithm was in what is called a black box. They kind of shroud over what's happening between inputs and outputs. And from Rebecca and Glen's point of view, this is pretty worrying. The legal team argued that this was against Glenn's constitutional right to confrontation, part of the Sixth Amendment that gives someone the right to confront witnesses or evidence against them. This argument was ultimately unsuccessful. Now, we aren't here to make a judgment of whether the parole board's decision to deny Glenn's parole was right or wrong. That isn't something I or even nature are really qualified to do. But we are here to look

at the role of algorithms and AI in decisions like this. Because this isn't an isolated case. Automation algorithms and AI are becoming increasingly used in criminal justice and other high stakes applications. And that has put a lot of researchers on edge. AIS have suggested to people that heavily polluted air was fine to breathe during a wildfire. Millions of black Americans were affected by racial bias in a health algorithm. People have been wrongfully accused of crimes based on faulty AI facial recognition. And these are just some examples that we know of.



06:41

The criminal justice system is becoming automated across the board and AI is a big part of that from every stage of a criminal proceeding investigations, pretrial release decisions or bail decisions. And also, AI systems are increasingly being used to analyze evidence of guilt that's introduced at the trial of an individual.



07:10

Glen was granted parole at his next hearing six months later. According to him, that was in part because he was able to get the committee to look past his high comfort score. But had it not been for that proprietary algorithm. It's conceivable he may have been able to prison much sooner. And proprietary black boxes like the one in Glens case, are just the tip of the iceberg.



07:42

In the world of more and more complex AIS like Chachi VT, the specter of the black box is looming ever larger, as AI agents like these are so complicated, that proprietary data is no longer the biggest problem. The nature of the machine means that even if we're able to look at the models innards, we may not be able to figure out how it comes to its decisions.



08:14

In this special episode of the nature podcast, we'll be discussing how to deal with black boxes, is it possible to break open and explain them? And if not, what should we do?



08:51

The Black Box in AI has a few different definitions. But broadly, it means that part of the AI or algorithm is obscured, you cannot see exactly what it's doing, either intentionally, or due to the sheer complexity of it being too much for a human to comprehend, or in some cases, both. To give you an example, we can think about large language models like Chachi Beatty, they build associations between words using a neural network, kind of like how your brain builds connections by ingesting a whole lot of words and symbols like billions is starting to learn patterns, which words have stronger connections than others, a bit like how your neurons strengthen connections between concepts or memories. So for example, it might learn that certain kinds of words like king, queen and throne have stronger connections to one another

than perhaps the words of panda and royalty. Now imagine how many words and symbols that are and how many different connections there could possibly be between them. From concepts, and you can start to see just how enormous this model can get. Getting to a point where there's such complexity that a human would never be able to fully unpick it all. The LLM becomes a black box. Someone who's been thinking about this a lot is chief news and features editor here at Nature Selects Beaver,

 10:20

much of what they learn and what they know has not been specified by a programmer. It's emerged from a training process. And so what that means is, even the people that made them don't know necessarily what they're capable of, they don't know why many of the things they do, why the AI is doing it, and they don't know all the processes that lead the AI to say a certain thing in a certain situation. So this process of training is extremely powerful, and leads to a lot of impressive achievements. But it also leads to a complete mystery, or a large amount of mystery as to what the AI is actually capable of.

 10:57

This black box not only means that it's hard to interpret how decisions are made, it is also hard to predict what it might be able to do. And that is especially important when we take into account that these chatbot-style machines can answer in a way, which feels very real and smart to a human user. One

 11:19

researcher described it as studying an alien intelligence, I think we have to be very, very careful of assuming that a capability because it means something and a person means that same thing in a machine. It's

 11:31

easy to feel that an AI is thinking like we do, but really, they aren't at all.

 11:42

So, there was a really good example with a researcher there was an announcement that an AI could answer some questions from a business school exam.

 11:54

World Antique is a new venture under development by two graduates of a New England Business School. The value proposition of the venture is simple. Purchase antique paintings at yard sales and auctions and then sell them and



12:07

the researcher took one of the questions that I answered correctly and reworded the question a little bit.



12:13

Bowl is a new company founded by three recent MBAs. The company plans to buy used automobiles at auctions and then sell them at four times the purchase price direct to consumers. I call



12:26

any human being who could answer the first version of the question, they would certainly be able to answer the second question because the modifications are quite trivial. But the large language model totally couldn't answer version two and could answer version one, which is, as a person is completely mystifying.



12:42

dollars should the company expect its operation to require



12:47

paintings become automobiles, New England Business School graduates become MBAs, but the basis of the business presented the question by cheap goods and sell them for four times the price that stays the same. And that is a relatively easy adjustment for a human to follow. Ultimately, the specifics don't actually matter to the question asked, which was a mathematical one about turnover and revenue calculations. The answer was the same. But the AI couldn't handle the changes. It gave a totally different answers. And researchers just don't know why.



13:22

We've never before built machines where even the creators don't know how they will behave, or why.



13:28

This is Jessica Newman, the director of the AI Security Initiative at UC Berkeley.



13:34



13:34

And we might not be so concerned about this fact, if the models were reliable, but they're really not.



13:40

The fact is that many AIs really aren't very reliable. The business school example is just one of many, and yet, they're still so impressive that their use is exploding. In the past year, the number of AI based startups in the US alone has increased by 1000s. And investment in AI is reaching hundreds of billions of dollars. And more and more of those models are becoming inherently linked to black boxes. So what to do with all these black box models? Well, one solution that many researchers propose is Explainable AI. Here's Jessica again, Explainable



14:19

AI is broadly defined as machine learning techniques that help people understand and manage AI. And it is often given as the answer to the black box problem. Explainable AI is referenced in most principles and guidelines for Responsible AI all around the world. So there's widespread consensus about its importance. And the promise of Explainable AI is significant. The promise is that we can really pull back the curtain and get a meaningful view into why a model is giving a particular output and what we would need to change for it to give a different one. So when that's properly achieved, it opens up so many avenues for using a technology in higher stakes or higher risk domains.



15:03

For example, if an AI is incredibly complex, perhaps there are tools or even other AIs that can pick them. There are machine learning approaches that subtly change the inputs of a black box model, and then look at the outputs. They use this comparison to build another model of what's happening inside the black box, a model of the model, and specifically a simple interpretable model with explanatory power. Then there are the human led methods such as red teaming, this is where researchers will essentially probe blackbox models repeatedly using different prompts or attacks to try and build a picture of its underlying structure. Some of the safety rails you might be familiar with in ChatGPT, btw, were created through this approach, like how it won't most of the time produce results that could be used for violence or harmful activities. It's important to note here, though, that neither of these approaches are perfect. Human led safeguards can be bypassed, and many people have done so. And interpretable models of models, well, they're never going to be able to explain everything. It's a simplification of how it thinks the model is working. It's not actually breaking open the black box. But despite imperfections, this idea of explainability has really caught on, and it appears in various legislation around the world. In the EU's AI act. For example, there is a right to explainability baked in. If an algorithm affects you, you have the right to know how it reached its output. But it's kind of unclear exactly how this can be achieved. Plus, writing a set of principles is very different to actually acting on them. In practice, this has all been much more complicated. Jessica has been investigating why this is the case. And



17:04

we found that part of the problem was that different communities actually have quite distinct goals and objectives for Explainable AI. So we explored how three different domains broadly described as engineering, deployment and governance, how they all articulate the goals of Explainable AI, and we found that there was shockingly little overlap between them, there



17:26

isn't really a good consensus about what Explainable AI is, for example, the engineers may be interested in what parts of the AI they can tweak to make it perform better. Whereas lawyers or ethicist might be interested in the underlying data, the model was trained on to assess privacy risks. Meanwhile, users might just be more interested in whether they can actually trust what the AI is telling them. And so people are doing different things for desperate goals. And often those making AI tools to figure out explanations are the ones making the blackbox AI in the first place, which doesn't really help you if you want to know if you could trust an output from chat GPT for, say, a recipe you're making.



18:12

And it's not necessarily a problem that different communities have different needs and goals. It's really to be expected. But it does highlight why Explainable AI is hard, because different communities are using the same term to mean different things. And the problem in particular is that by and large, only engineering objectives are being met, which are typically in line with the objectives that the technology companies, while the objectives of users and other stakeholders are often an afterthought.



18:41

Some in the field have even argued that we shouldn't focus on Explainable AI at all. Instead, we should only use algorithms that are interpretable in the first place. For example, in Glenn's case, there were simple open interpretable algorithms that could have been used, but they weren't. If they were, it could have allowed Glenn and his legal team to more easily challenge unexpected outcomes, as they could clearly show where things might have gone arrive.



19:19

Because of this, Jessica thinks we need to go beyond explainability explainability is



19:24

not enough to enable trust. Even if we improve explainability methods and more people are able to better understand and AI systems, capabilities, limitations and reasoning, it still doesn't mean that the outputs will be good enough, or that there will be accountability for harmful



outputs. So for people to really trust AI, we need to have robust systems of governance for the people and organizations building and using the AI to ensure there is accountability for any harm that's caused.

 19:56

For example, this could mean that the company behind The Compass algorithm, which Glenn argued prevented his first parole hearing going his way would have had to take responsibility for preventing his early release. This is in contrast to what actually happened when a court ruled in favor of the company to protect that intellectual property. Now, this is starting to change, and courts are starting more often to order companies to disclose information, such as what's in a proprietary black box to defendants. Additionally, the executive order from President Biden said that quote, it is necessary to hold those developing and deploying AI accountable to standards that protect against a more full discrimination and abuse on quotes. Such accountability is vital to ensure trust, according to Jessica,

 20:53

and to ask for trust from users organizations also need to make it clear that they will only use AI models where their use is appropriate, and the risks are manageable. So that means that given the current challenges with effective explainability for deep learning models and generative AI, that they should only be used in cases where people can accept that they won't fully know what it will do and why.

 21:22

But there is another wrinkle in explainability. One concept which makes this whole discussion quite a bit trickier. And it might not be what you expect.

 21:33

In some cases, AI systems for legitimate reasons need to be blackbox.

 21:49

So there are going to be several contexts where keeping it opaque is good is good for society. Overall,

 21:56

this is our Bochy, a computer scientist from Purdue University, he points out that there are sometimes reasons to keep hold of a black box. The first



22:06

one is probably the most obvious one, that is to protect the intellectual property of the vendor, the software development organization that came up with this, and therefore to maximize the commercial benefits that accrue from these AI systems. And these AI systems are by no means cheap to build, maintain, and therefore for very understandable reasons, the commercial entities behind them would like to maximize the commercial benefits and maximize these benefits over as long a period of time as possible, the



22:40

AI market is likely worth hundreds of billions of dollars. And that is growing at an incredibly rapid rate being at the head of this particular race would certainly be lucrative.



22:52

Reason number two is to increase the security of these kinds of systems. So the idea is that if I keep parts or all of the system obfuscated from the end user, then this improves the security of the overall system. Since the thinking goes, people cannot find out what are the vulnerabilities in it and cannot therefore exploit these vulnerabilities. This in. In classical the non AI security systems is a principle that has been debunked, where security by obfuscation for the large measure does not work. But in this notion of AI system security. This is one of the reasons why organizations tend to keep this black box



23:33

there is an ever present worry that hackers and bad actors could harness AIS for ill. For example, if they were able to see the underlying model behind Chuck GPT, perhaps they would be able to use it to generate the sorts of harmful material that it normally prevents. Or what if AI is are used in more high stakes scenarios like warfare. And



23:55

the third of three reasons is that these models ingest various kinds of data, often enormous amounts of data. And some of this data may be sensitive, maybe privacy sensitive, and therefore there is a legitimate reason to protect the privacy of this data that has been used to create the model.



24:12

This is especially true of AI is working with things like medical data, hospital records, and things like that. Now, so does also think that these black boxes can be a problem, especially in places like healthcare or policing. Also, if AI is open, it could make it easier for altruistic hackers to find

and patch up vulnerabilities. And like Jessica, he thinks that we may need to go further than just trying to open the black box. For it to be really useful to users. We need to focus on fairness, reliability, and security. Fairness

 24:49

is another very important property of AI systems. So if two people come to an AI system with roughly equivalent background circumstance then does the AI system treat them in a similar equivalent manner, reliability is another one. And third one a big one, which there's been a lot of public discourse on is the security of these AI systems. So can malicious actors will fully support these AI systems by feeding it wrong data or by making changes to the code. And for societal acceptance to be widespread, they have to believe that AI systems are built with a higher level of security than maybe traditionally our systems have been built with.

 25:34

To achieve this, he believes that researchers should try and develop theories and tools that can help make the black boxes a little less black, trying to maintain their opaqueness when necessary for security, for example, but generally trying to make them more open. So we can focus on this fairness and reliability.

 26:01

To get to that point, though, he thinks that researchers and the technology companies need to do a better job of working together.

 26:08

It seems like there are two parties which are arrayed against each other. So there is a party which is in this commercial art, which are in a man race to win this race. So where we have seen even in our dealings with some of the software companies, which are at the leading edge of this AI race, is there is a decision to step back from openness, openly publishing datasets, openly publishing algorithms openly publishing results, to step back from that in order to hide these kinds of crown jewels in order to win the race. Again, and then there is another sector of the community, which is very loudly proclaiming the dangers of AI systems that are completely blackbox. And they're saying, we need to make AI systems more open to end users, not just to the researchers and experts in the field. And we seem to be talking over each other in this conversation. And what I'm hoping is this podcast that you're doing is going to allow us to understand that there are nuances on both sides and have these two sectors talk more constructively with each other. Because ultimately, what is a win for everyone is AI systems that are trusted, and therefore they're adopted more widely by society.

 27:39

But perhaps the way to achieve some of these goals actually doesn't need researchers to open black boxes. In fact, what if instead of investigating what is happening in the black box, you

black boxes. In fact, what if instead of investigating what is happening in the black box, you instead look harder at the fleshy bits in front of it, the human using the AI



28:06

we call it human centered Explainable AI because what we're really doing is we're backtracking from the user's needs. So the first thing you have to do is figure out what the user is trying to do. And then figure out the sort of information they're actually going to need to be able to do that. This



28:24

is Mark ryedale, an artificial intelligence researcher from the Georgia Institute of Technology, who is a proponent of this human centered approach. Key to this is action ability. What is the actual information that a user requires in order to use the AI in a way that addresses their needs? For example, say I'm asking chat GPT for a recipe for a cake. And in this case, it might not really be important to me how exactly it's coming up with its answer. But in order to use it in a way that's going to be best for me, I still want to know a few things. Like I might want to know if the recipe is tasty. Normally, when I browse the web for recipes, I can be reasonably sure that some human has eaten the cake and can vouch for his electability. There are likely to be reviews that will help guide me to with that information, I might be able to reliably think that this recipe will produce a tasty cake. But with chat GPT I don't have that. I can't ask it. Is this recipe tasty?



29:39

So sometimes in question answering systems, people are asking a question because they need that particular answer. And they don't know whether the answer is right or wrong or not. And the AI system really can't tell you yes, my answer is right or no, my answer is not right. But they're trying to calibrate their trust to the system. Right. Is this a suit situation where chatty PT is really reliable. Is this a situation where GPT is not reliable? And how do I know if I have kind of walked from an area where DBT is really reliable to one in which it's not reliable? What information would GPT need to give me to know that? One example is, it can kind of not just tell you the answer, but it can tell you other facts that are around the area. Right. So if this is true, then this other thing should also be true. And this other thing should also be true.



30:30

So for my cake, instead of asking Chuck GPT, whether or not the recipe is tasty, something it can't really know, I could instead ask it about other recipes that I'm familiar with. To get a sense of how reliable I think it is in recipe production. If I have a recipe seems similar to ones I know are good, then I might decide to make the cake. The actual information is that recipe is probably going to make a tasty cake. The idea behind human centered Explainable AI is that this would be an automatic part of the process, I'll be given information that would allow me to use my own judgment to take the actions I want, and allow my own brain to counteract potential weirdness that has happened in the black box. We started seeing some of this in

existing API's, for example, Bard, Microsoft's chat bot, give sauces when it answers a question, which you could then interrogate to see if you trust the answer, effectively allowing you to take the answer with a pinch of salt. Exactly how to wrap this into AI is still a matter of research. What are the best ways to meet a user's needs? And how do we do that when many different users have many different needs? For Mark, he would like to see that these technologies are not so one sided. Instead, with them being able to communicate with the users,

 31:56

you know, I think that these technologies, they should be able to enter into a dialogue with us about both the user's needs and their wants and their desires, and what they want to do with it, and to provide actionable advice about the consequences of using their information. So we just did a study with doctors and teachers, and they said, like, what they really care about is, you know, they wanted to calibrate their trust. But ultimately, they want to know, kind of the hypotheticals, if I use this information, how will that affect whatever it is I'm trying to do with it. And I think AI systems are approaching the level where they can enter into a dialogue about those sorts of conditionals. And to the extent that they can kind of zero in on what the users who are trying to do, they should be able to, I guess in a broad future sense, kind of adjust how they communicate, the level of detail, the external and extra justificatory pieces of information that need to come along with that. And that can be anything from how it was trained to know how different factors are impacting how the model works, to, you know, things that don't have anything to do with the model at all, but kind of how this information affects the real world in the future. And that would be kind of an exciting new way of thinking about explainability. Whilst

 33:16

this could be an interesting avenue for explainability. One thing that everyone I spoke to emphasize was the research isn't really there yet. But at the same time, we're living in a world where people like Glenn are being affected by black boxes, and AIs seem to be sprouting like daisies. So what do we do right now?

 33:43

More research is one avenue. But another is to critically ask when would AI is for useful? Do we need an AI to tell us whether someone should stay in prison or not? Are there better ways to deal with our human biases than implementing an algorithm with a black box? Perhaps there are simpler ways. But if we do decide to use API's, if we believe that they will be helpful, then collectively as a society, we need to think about what we want from them, and how to implement them in a way to best protect everyone. And for that, maybe it's worth remembering that these AIs aren't just synthetic. There are human fingerprints all over them.

 34:32

It's shaped by people at every stage, including for large language models painstakingly curated and filtered by human reviewers. So there are many opportunities for people to shape the system. It doesn't just appear out of nowhere

system. It doesn't just appear out of nowhere.



34:50

How much of those pieces they need to be opened up is also going to depend upon the context the domain and that's a conversation that we need to have and we need to start on Back failures.



35:04

I think there's value to be achieved in augmenting humans and human decision making with artificial intelligence. But the higher the stakes, I think the more we have to broadly question whether we need to kind of wait until we really understand the implications of using AI.



35:25

I love the analogy with like testing a drug or medicine. So a clinical trial is based on just looking at does it work and making sure you've kind of controlled for the right things and done a systematic test, but you're basically just looking does it work? So we want to do that with the AI is and then you can imagine the AI sort of being approved for use in certain situations, but there being a statement about uncertainties as you might have with the medicine, but then also, you know, scientists have mechanistic understandings of why drugs work, and that can build a huge amount of confidence. And so I think that is definitely it's very interesting to do what kind of people have called it like doing neuroscience on the AI so you're trying to understand the underlying processes and connections that the AI has built in order to deliver the results it delivers. And like so I think both are important being worked on, can help break open the black box.



36:38

This was a special episode from the nature podcast. It was narrated and produced by me Nick Patrick, how, with editing help from Noah Baker. Thanks to everyone who spoke to me for this episode. And to all of you for listening