When AI Goes Wrong

https://mindmatters.ai/podcast/ep160/

Robert J. Marks:

Welcome to Mind Matters News. I'm your gregarious host. Robert J. Marks 2.0, We're here to talk about AI. AI design ethics requires AI to do what is designed to do and no more, but problems pop up in complex systems, including any attempts at generating artificial general intelligence or AGI. AGI, whether you think it'll be achieved or not will by necessity be complex. And the more complex the system, the more that it can go wrong. To talk about this today is our PhD student, Samuel Haug, and freshly minted PhD, Dr. Justin Bui, both are members of my research group and are really smart. I really feel fortunate to have worked with them and to continue to work with them. So Sam welcome.

Samuel Haug:

Thank you. Happy to be here.

Robert J. Marks:

Okay. And you too, Justin.

Justin Bui:

Yes. Thank you very much for having me on.

Robert J. Marks:

Okay. I want to start out with something that rings of Paul Harvey's The Rest of the Story, either Sam or Justin, have you ever heard of Paul Harvey?

Justin Bui:

I have not.

Samuel Haug:

No, I have not.

Robert J. Marks:

Okay. That shows, that shows I'm a senior citizen here. Paul Harvey had a series on the radio a very popular series. In fact, wrote a couple of books too, where he recounted a sometimes familiar story and then added a little twist at the end, kind of an Alfred Hitchcock twist at the end, which few or few if anyone had ever heard about, and the twist at the end was the rest of the story. It was a little elaboration on the that nobody expected. We're going to do this today with some popular AI stories. And the twist is going to be something not well known about how AI failed these failures called unexpected contingencies are our rest of the story. And so it will illustrate some of shortcomings of AI and illustrate this idea of unattended contingencies, which we want to talk about in the podcast.

Robert J. Marks:

We'll start out with some simple examples and then we'll get more serious cases involving human life. The list is from a peer reviewed paper that Sam and I wrote with Bill Dembski and it's a peer reviewed in the IEEE Transactions on Systems, man and cybernetics. And we'll make a link to that available in the podcast notes.

Robert J. Marks:

So let's do the following. I'll tell a story. And Sam I'd like you to give the rest of the story, give the stories twist at the end. Is that okay?

Samuel Haug:

Yes. Yes it is.

Robert J. Marks:

Okay. Okay. First one number one, Jeopardy is one of the most popular quiz shows in the history of television. Could AI win at Jeopardy? Well, it made big news. The answer is yes. In 2011, the world champions in Jeopardy took on an IBM computer program named Watson. Watson didn't respond to every answer correctly. It wasn't designed to do so, but in the end playing the game, Jeopardy, Watson recorded a resounding win over both of these other Jeopardy champions and that made headlines, but people left out maybe a little quirk in Watson. So Sam what's the next, what's the rest of the story on This?

Samuel Haug:

Yes. So in this particular contest, there was quite a funny occurrence where Alex Trebeck asked one of the contestants a question, and the answer that the human contestant gave was what are the 20s as the answer to that question, which was noted as incorrect by Alex Trebeck and immediately afterwards Watson buzzed in and gave the exact same response. What are the twenties? And obviously this answer was incorrect because it was just revealed to be incorrect. And this was something that the programmers of Watson did not foresee.

Robert J. Marks:

Yeah, this was an unintended contingency. I imagine when the Watson programmers heard this duplicate response, they face palmed and they go, oh my gosh, that was such an obvious thing. We could have put into the software, but chose not to do just fascinating. You know, Watson had great plans for itself in the field of medicine after it premiered on Jeopardy, but it failed. The idea was this, there are just hundreds and thousands of different papers published in the medical field. And wouldn't it be wonderful if Watson could mine all of this data, which was published in the medical field. And then based on a query from a physician who gave symptoms and details about the case they were dealing with was able to respond with a list of papers relevant to what was happening. This would save the doctor from wading through thousands of papers in the literature.

Robert J. Marks:

Watson contracted with medical research group and hospital MD Anderson. But after a while, MD Anderson just fired Watson. It just wasn't doing the job. And in fact, we listed this as the number one in the top 10 AI exaggerations, hyperbole and failures in the year 2018, we list to this on Mind Matters

News since then, IBM Watson's application expectations have even fallen further. And so we're not sure what's in the future for Watson, but we can see that even though it was working well, it did have this unintended contingencies, okay. Example number two for the rest of the story. And this has to do with also another IBM piece of software in 1997, IBM's Deep Blue software beat world champion, Gary Kasparov at chess. This made world headlines. One of Deep Blue's moves was particularly curious, the unexpected move psychologically threw Kasparov off his game. And he lost. Kasparov, looked at the move and said, I can see no reason for why, IBM Deep Blue made this particular move and it blew him all off his game. Psychologically.

Robert J. Marks:

One of the chess experts who were commenting about the game said, quote, It was an incredibly refined move of defending. While ahead to cut off any hint of counter moves. Well, you know, I guess skill in a game is like interpreting art in a painting. Some people will look at a painting and, and some will think this is great art. And others will say, this looks like a kid's finger painting. And that was indeed the case for this incredibly educated commentator. So the interesting thing is what is the rest of the story? So Sam, could you, could you kind of finish this out? What little twist on this Deep Blue move?

Samuel Haug:

Yes. So this one's also a little humorous here. It turns out that over a decade after this match, one of the computer scientists who designed deep blue, Murray Campbell, he confessed that the move that Deep Blue made that threw Kasparov, off his game was a random move that Deep Blue had chosen because Deep Blue was unable to choose a good move. And so he just chose one at random.

Robert J. Marks:

Yeah, that's fascinating. I think one of the quotes from Murray as Kasparov had concluded the counter-intuitive play must be a sign of superior intelligence. He had never considered, that it was simply a bug in the code. So that was just a fascinating sideline of the rest of the story about Kasparov being beaten by IBM Deep Blue.

Robert J. Marks:

Okay. Third story for the rest of the story, a deep convolutional neural network was trained to detect wolves. Now, deep convolutional neural networks. They do make mistakes in their classification. That's just the way that it works, but this one incorrectly classified a Husky dog as a Wolf. And so the designers of the code went in and did some forensics and they found out that this was a fluke of the Neural Network. What happened here, Sam? What was the rest of the story?

Samuel Haug:

Yes. So this seems to be a theme of some humor in these stories, the neural network in this particular instance had not been training on the features of the animals that it was classifying, but it had picking up on the fact that all of the Wolf pictures that it was fed as training data had snow in the background and all of the dog pictures that it was given as training data did not have snow. And so the Neural Network had not learned anything about the features of these animals, but had just learned to detect the presence of snow.

Robert J. Marks:

That is really incredible. Justin, have, have you found out that this is something that which can happen in deep convolutional neural networks? Have you ever bumped across it?

Justin Bui:

Yeah. It it's pretty, it is pretty comical when you see things like that, it kind of goes hand in hand with how the network's developed and, and how it's trained. You know, it takes some very careful thought and preparation to not only design a neural network, but to train it. In fact, you know, it's, it's often said that the 90% of a systems' value is in it training and input data, right? Data is everything garbage in equals garbage out. And it's funny, I chuckle when I hear about that story.

Justin Bui:

But you know, just for, for kicks a couple weeks ago, I built a simple little convolution neural network to classify cats and dogs. It did so with, with, with really good accuracy of the, was in the order of like 97, 98%. And then for, for, for last I fed it a piece of fruit. I, I fed it an image of a, of a Kumquat and yeah, it turns out Kumquats are a lot like dogs, apparently. So there's some, just some oddities, some peculiarities that go into developing these systems. And again, it's garbage in is garbage out. And if you're not thinking about some of these contingencies, you may never come across them.

Robert J. Marks:

That's incredible. Okay. Thank you. Okay. Story number four for the rest of the story that we're talking about. So driving cars are still under development and despite promises are still very far away from a level five, which a level five is a self-driving car doing what a human can do. So self-driving cars in early development. Were trained to watch out for things like pedestrians, deer, and road debris. You don't want to hit a pedestrian. You don't want to hit a deer. You don't want to run over road debris. This worked out most of the time, but there were some serious flaws, at least in this early development. So Sam, what's the rest of the story here?

Samuel Haug:

Yes. So this one, there's a very serious side effect that happened in 2018, an Uber self-driving car in Tempe, Arizona actually struck and killed a pedestrian because it was unable to correctly classify this pedestrian as a pedestrian and as such did nothing to avoid the collision. One of the engineers that worked on this self-driving car thinks that the vehicle was able to see the pedestrian, but that it was not able to correctly identify it and avoid it. And it's just a, very, very sad occurrence of a, an unexpected contingency.

Robert J. Marks:

So I think the bottom line here is when AI involves human life and the potential death of a human being, you have to be very, very careful about unintended contingencies. I also think that early in the early in the development of self-driving cars, that when blown plastic bags were often interpreted as deer and stationary plastic bags were sometimes considered road debris. And so these are things which can be fixed, and we still have hope that this artificial intelligence that caused this death of this pedestrian in the Uber self-driving car and be corrected, but still, this was a terrible unintended contingency. And they remained a, a, a major obstacle in the development of level five self-driving cars. Justin, do you have any comments on this?

Justin Bui:

Yeah, it's one of those things that I think the self-driving nature of, of cars is still quite a ways away. There's a lot, a lot of systems out there that can reasonably identify pretty much every road hazard, you know, with a high level of confidence, but when it comes to human life, it's one of those things that, you know, even a three, 4% chance of misclassification is catastrophic, you know? So I think a lot more due diligence needs to be paid to classification and detection systems. And, and it's something that I think it's just going to take some time to tackle.

Robert J. Marks:

Yeah. And, you know, Tesla keeps coming out with all these press releases that are doing great things and they clearly are doing great things. One of our writers at Mind Matters News, Jonathan Bartlet comments extensively on Tesla's update. And I've talked to some people with some Tesla self-driving cars. They can take their hands off the steering wheel for a while, but Tesla will warn them after a while. It says, you know, your hands haven't been on the steering wheel for a while. Let's see them. And so they're not ready to go to totally autonomous self-driving cars as of yet.

Robert J. Marks:

Okay. Here is the fifth story and the stories are getting more and more serious. Now we started out with little things like jeopardy having IBM Watson, repeat an answer. That was, that was a little curious thing we just got done with talking about how Uber the self-driving car would kill people.

Robert J. Marks:

And now we're going to get to something which is very serious. It's a complex system that could have caused millions of deaths. Let me give you an example, or let me give you the story. I should say, during the height of the cold war, the U.S and the, the Soviet Union were existing on the political knife edge of something called mutually assured destruction or M.A.D. The idea was is that if the United States blew up Soviet Russia, then Soviet Russia would blow up the United States. And both of the countries would be flat and glow in the dark in order to play this terrible game a little bit more intelligently. The Soviets deployed a satellite early warning system called O.K.O, O-K-O. And O.K.O's job was to watch for in missiles fired from the United States on September 26th, 1983, O.K.O detected incoming missiles.

Robert J. Marks:

At a military base outside of Moscow sirens, Blared that the Soviet brass was told by O.K.O to launch a Thermo-nuclear Counterstrike against the United States. Doing so would result in millions being killed the officer in charge, Lieutenant Colonel Stanislav Petrov, looked at these incoming missiles, and he felt that something was fishy. It just didn't feel right. The United States would not launch a preemptive strike doing this sort of strategy.

Robert J. Marks:

So after informing his superiors of his hunch, that O.K.O was not operating correctly. Petrov did not obey the O.K.O order. Upon further investigation. O.K.O was found to have mistakenly interpreted sun reflecting off of clouds as incoming U.S. missiles. In other words, these signals were simply the sun reflecting off of clouds. There was no U.S. missile attack and Petrov's skepticism of O.K.O's alarm may have saved millions of lives.

Robert J. Marks:

So we've gone from the very innocent to the very serious of what happens with AI unintended contingencies. Unexpected contingencies from complex. AI can become more and more serious as we've seen. I don't know about you guys, but I play Alexa. And when you can't get Alexa to play a song you want it's annoying, but it doesn't cost any human lives on the other end of the spectrum, killer self-driving cars and detectors of thermo-nuclear strikes. can't be allowed to make mistakes if they do lives will be lost.

Robert J. Marks:

In the examples that Sam and I have gone through. We, we have run the gambit from the very innocent to the very serious, the name of the paper, which this is outlined in is called Exponential Contingency Explosion (Implications for Artificial General Intelligence). It's by Sam, William Demsky and me. And it appears in the peer reviewed AI journal. IEEE Transactions on Systems span and Cybernetics, and Sam is the first author. Now in that paper, we also do a bit of math. We show that the number of contingencies can increase exponentially with respect to the system complexity. The number of contingencies can become so numerous that they cannot all be looked at individually. This is troubling. This is not good news for AGI, which by its very nature must be very complex. We'll explore this exponential explosion of contingency increases as complexity increases linearly. Next time on mine matters news. I want to thank my guests, Sam Haug and Dr. Justin Bui for their really interesting insights until next time on Mind Matters News, be of good Cheer.

Announcer:

This has been Mind Matters News with your host. Robert J. Marks explore more at mindmatters.ai that's mindmatters.ai. Mind Matters News is directed and edited by Austin Egbert the opinions expressed on this program are solely those of the speakers. Mind Matters News is produced and copyrighted by the Walter Bradley Center for Natural and Artificial Intelligence at Discovery Institute.