4/30 書報

# 逐字稿

###### 00:04 A vertical rate of the aircraft after an initial advisory.

###### 00:05 His arm. This is the fast invited Talk in the Americas session of the each Guide. Twenty Twenty, I am dumb. V of speaker is Michael Khokand. Also an unfair to introduce him on Michael Coconut for is an Associate professor of Est Aeronautics and Aesthetics at Stanford, and he works on robust decision making systems On is also the Director of A Sail, Pyotr Center for the Research at Stanford, and a Co. Director of the Sand Prefer the I safety Arm.

###### 00:11 However, we see that the number of reversals are significantly decreased. So overall, verses, I, a great story. Ah. But if we wanted to reduce the rate of strengthens men, what would we do while we win, and perhaps increase the cost of strengthening and re optimize in this will be at the expensive some these other metrics. But the key innovation here, it is that it moves the burden on the engineers from trying to design or pseudocode, to directing much more of air. Focus on the balancing of these different metrics.

###### 00:42 Coconut Flood is also an author of textbooks, including Audition Make, another uncertainty theory and Application, a copy of which I how welcome, and algorithms for optimization, bought from a mighty press on. He has deceived of his Phd, alms from my University of Edinburgh and Emerson, the of bachelor's degrees from Stanford, he's going to be talking to us today about a very important topic that high up on our minds, which is automated decision making further safety critical applications, how to get agents to our show, a robust behaviors are, I should point out, the item, are Michael actually has recorded his talk, but he's also available live, so he will be monitoring the are the zoom chat, so you few have questions during the dark, you can put them on the zoom charred, and he will are be. Nor, I'm answering them as they talked progress and, of course, and of.

###### 00:53 The Monte Carlo analysis can go a long way. Invalidating the system were interested of courts in estimating the probability of failure. Missus, difficult for a number of reasons. Run is that we need to run ritson simulation and we need a good model of the environment.

###### 01:11 An under appreciated fact is that this model absolutely must be independent of the model used for optimizing the logic of for a cassettes, we used a relatively simple Pompey P. Model probably used a much higher fidelity model for the validation of the resulting policy, recollected. Nine months of raider data from Oliver F. A. And Department of Defence radar isn't built up the Dijon network to capture all the various statistical correlations between the various variables. Then we can sample from that many times and arrived at the probability of collision, since our systems are very safe, we naturally need many simulations in order to arrive at a good failure, probability estimate, we can take advantage of important sampling and the cross entropy method to reduce the variance of are estimates, but we still need quite a bit of computation during the validation.

###### 01:42 Attack is over. We will get him back. Live in front of you. And then Dick would be time for questions, too, so that swerve Aviat on saw or let's start the talk.

###### 01:54 My name is Michael cooking. Their film be talking about automated decision making for a safety, critical systems know much of the focus, san advances of a. I have involved applications where it has been located. Two occasionally make mistakes, like accidentally recommending the runaway be, were miss classifying a dog as a cat.

###### 02:10 Of a Cossacks. We ran tens of billions of simulations, and we were wearing the primary users of the Lincoln Laboratory Super computing grid, a design time. We also want to be able to identify the most likely failure modes, we of course need a problem ballistic model of the environment in order to measured likelihood, searching the space of possible scenarios is also challenging, because there are many variables, and the scenarios extend across in time. I'll dig more deeply into the shortly when we throw massive compute power at simulating many simulations, we will of course, fine failures, Elise for a aircraft collision voting systems, and this goes for automated driving as well, no system will be perfectly safe. If our simulations discover hundreds of failure scenarios, we'd like to categorize them into failure modes, the are interpreted, will, then we can decide what to do about those failure mints.

###### 02:18 For many years, our work on Ai has focused on algorithms for making these A. I. Systems better, which is great, and there is still awaits to go, not to Loango, are the focus turned to March data sets with the hope that large data sets will continue to improve the performance of our systems now refocused, starting to turn to the challenge of validation. There is a growing realisation that may be the hard part in Ian's. Not really, the building of the algorithms are a collection of the data, but rather is making true that this system will behave the way we want when deployed in the real world.

###### 02:57 After a series of environment embarrassments, massive research efforts are underway to validate the fairness of a. Assistance and reducing various forms of bias and discrimination that may be difficult to predict during the design phase of these A. I. Systems, and the only really become apparent when deployed in the real world. These are costly mistakes, costly in terms of Costa Society, cost to suicidal trust, a costly in terms of the reputations of the companies that use and deployed these systems.

###### 03:09 Identifying the most likely failure can actually be frame doesn't optimization problem, where we tried to maximize the probability of a trajectory subject to failure, the key insight here is that this optimization problem can be framed as a Pompey P. M. We can use different Pompey p solution techniques to arrive at a failure.

###### 03:31 We call this methodology adopted stress testing, and it works like this. We have a simulator, melican step forward in time. It encapsulated an environment.

###### 03:31 I mean, I, aerospace professor. In this industry, we tried to build highly reliable systems that take people into outer space and fly the four point. Five Billion passengers in twenty, Nineteen Around the world.

###### 03:42 And the system under test, so that system under tests, we just need to be able to execute it, so are we don't need the source code, we just need to be able to execute it. So it makes it very easy to trial different systems, like a Cossacks and Tea Cass, and a automated driving system for systems and flight management systems, and so forth.

###### 03:48 A. That number, of course, went down in. Twenty Twenty I aerospace systems are incredibly complex, yet an incredibly safe place to be. Now, a days is in a metal tube at thirty thousand feet.

###### 04:02 As all the faculty in my department can tell you, the future of aerospace's autonomy were greater levels of autonomy can enable greater flexibility in the exploration of distant planets or safer, more efficient, a more sustainable aviation to transport, passengers and goods in this talk, I'll discuss some of her work on validating complex decision making systems, as most of you know, there are different ways to model decision making problems, others, the Markov decision process, where recapture outcome, uncertainty, and then there are the partially observable Markov decision processes or poverty piece were be capture state uncertainty, this model is especially relevant in aerospace contacts, where we need to be robust, to censor airs inclusions, and so forth, there is also reinforcement, learning where we don't know the state transition probabilities are rewards ahead of time and have to interact with environment to understand.

###### 04:04 The output of the simulator is an observation and reward, the reward is related to the log probability of the transition in the environment.

###### 04:15 Now I in upon the peak. Of course, we want to maximize the sum of rewards, when were maximizing the son of log probabilities were actually maximizing the likelihood of a trajectory kept, so this gets fed into a reinforcement learner. Their output's actions, their actions represent disturbances to the environment.

###### 04:40 It turns out we can use any reinforcement learner, we won't, we can use Monte Carlo Tree surge are deep reinforcement, learning where whatever and other aim is to just output the most likely failure path.

###### 04:54 This is an example, failure from a very early prototype of a Cossacks and its. It's really nice we get these trajectories on the left, we see a plot of the a lateral motion of the aircraft, with the different advisories indicated, an on the right hand side, we see the vertical profile of this encounter with a different advisories. This one actually led to a collision. No understanding these kinds of failures at design time are it's extremely useful.

###### 05:01 The it works.

###### 05:03 Palm, the piece, or especially relevant to safety critical systems, were we need to provide robustness to various sources of uncertainty. Unfortunately, many of the algorithms for assaulting Pompey peas are extremely complicated.

###### 05:17 Are there were some tools available at the time they're allowed you to specify problems using text files containing transition probabilities, reward parameters and so forth, but for many of the problems of interest, even just the transition probabilities, would be too large to fit in a file, we wanted to create a library that when allow non expert users to specify their problems, programmatic lay. So one of the first activities of my graduate students at Stanford was to create a library that would make it easy for people from different disciplines to specify their problems and then try out different solution algorithms. We wanted something that would be fast and easy to use, as well as usable for teaching this resulted in poverty, penstock jail, which is written in the divine language, a that is, Julia, its interactive, its fast, its typed, and we can write for loops and you branching without a huge, been performance penalty.

###### 05:29 So with the success of the application of adaptive stress testing to a cassettes, my students were interested in seen how well It skills to automated driving problems, and this is a simulation in the open source, Carlist simulator were simulating a driving policy, as well as a perceptual system. You can see the bounding box around the pedestrian.

###### 05:55 But we find, is that the most likely failure is not very relevant to us.

###### 06:02 The pedestrian is basically running out our vehicle and flinging themselves on the hood.

###### 06:09 So this is not very relevant because we are not outfall, so it turns out that we can actually incorporate into the reward function the notion of fault. And to do this, my student Anthony Corso implemented the are assessed system from mobilise. I just based on a white paper that was posted on, archive.

###### 06:17 To make many of these ideas accessible, I wrote decision making under uncertainty that came out from Mit Press in twenty Fifteen, it was targeted at a general engineering audience, many of the ideas came from the a community, of course, the many of those ideas originated in the operations research community a few decades previously, many of the algorithms were decision making under uncertainties, fundamentally, a form of optimization, and it turns out that the validation of decision making and other forms of A. I. Systems can also be framed as optimization problems. I found that many people on, both within the a community and outside, were limited in terms of their breadth of exposure to optimization algorithms sewed together with my former Phd student, Tim Wheeler, we wrote a textbook that came out in twenty nineteen on Optimization and serve suit.

###### 06:33 What we get, as the most likely at fault, failure is much more interesting. We see the the bounding boxes bouncing around. This is due to a perceptual airs, and we also see that eard car is driving a little bit more aggressively than perhaps we would like to see wenn, a pedestrian is starting to wander into the road.

###### 06:54 So understanding this at design time is, is very important because it may lead us to decide to add additional censers are changed parameters in our present sexual system, Mimi. Changing some the parameters in or particle filter, or Kalman filter, and we may also want to play with the tuning of the aggressiveness of the vehicle. Adapted stress testing is a sampling base method and, regardless of whether we use Monte Carlo Tree search or deep reinforcement learning in our optimization loop, we do not arrive at any formal guarantees. We have also explored the use of formal methods for verification which do allow us to make formal guarantees, develop performance out these guarantees with respect to some model. A collaboration with others in the Stanford Computer Science department with few years ago led to the creation of the Really plex Algorithm, which is a sound, incomplete algorithm for verifying properties of neural networks. The properties are defined over a continuum of inputs and so there.

###### 07:16 We used actual executable Julia code to elaborate on the algorithms, while still preserving human readability. I'm really excited about a new book that we're wrapping operate now, it's titled Algorithms for a Decision making a, together with Tim Wheeler and Kyle Re. A. You can find the pdf on algorithms. Book dot com and its seven hundred pages. We discuss the various sources of uncertainty and the challenges they present to making robust decisions, including outcome, uncertainty, model, uncertainty, state uncertainty and interaction, uncertainty, that is, the uncertainty the arises from multiple agents interacting with each other again, we provide implementations of the algorithms, and Julia, as you can see, we were inspired by the format of Edward Tufte, these books. A major success story of a Palm tepee approach to modeling decision problems is a Cossacks, the next generation Aircraft collision, avoiding System. It grew out of our research back when I was her, Mit Lincoln.

###### 07:54 Is not a way to exhaustively enumerate aldo's inputs and manually check that the outputs adhered to the specified property, a actually involves formerly reasoning about the individual weights in the neural network and arriving at a proof that the property holds were if it doesn't, it will produce a counterexample, we can use verification algorithms to support reachability, analysis as old discuss shortly. We may want to prove that if we follow a policy represented by a neural network, that it is impossible to enter an unsafe set. We have also done work on probable Isak model checking were, instead of determining whether a property holds her, not we can prove that are property holds without leaves, some probability were initiated our work on really plex with some worked on by my former student, Kyle Julian, and we were interested in seen. Whether we could compress policies were amend aircraft, collision avoidance, represented as numerical look up tables, what I'm showing here is just a national policy that was generated.

###### 08:16 Laboratory, it started out as a concept on a wakeboard in two Thousand Seven, and we were able to sketch out much of the details of the decision making magic over a couple years, resulting in a technical report in twenty eleven, you can find a pdf of it online, but it wasn't until twenty eighteen. After tremendous effort by folks at the a Lincoln Laboratory, Johns Hopkins Applied Physics laboratory, the Mitre Corporation and others, that it became an international standard, Archie idea three, eighty Five, that's the version for manned aircraft, the version for unmanned aircraft was accepted as the standard, just in twenty twenty, so you might be wondering why it took so long to have it accepted as an international standard on despite teams working on its founding multiple organizations will add. This is a system that we don't want to mess up, and it is the system that everyone in the world needs to agree upon.

###### 08:53 Using an open source, implementation of the toy collision avoidance problem, not the official a cassocks system, of course, this table occupied one point to to Gigabytes, and he was able to fit in neural network to the action value function, which resulted in a policy that looks like what you see on the right in serve, one point to to Gigabytes, the weights in the neural network occupy only hundred three kilobytes, which is quite an impressive compression ratio. He didn't spend two long training the network. And as he can see, the decision, boundaries are not perfect by looks roughly sensible, and but what was surprising about this research, which you can read about in the journal article Blow, that it performs super well when evaluated on a variety of metrics using Monte Carlo simulation of course, when should be very wary of using a neural network for a safety critical system and the reason, for those that neural networks can produce unreasonable outputs for what appeared to be.

###### 09:15 There was quite a bit of interest in collision, avoiding flung before two thousand, seven, in fact, in Nineteen, Fifty Six, there was a mirror collision over the Grand Canyon involving United Airlines in t. W. A. There were a hundred, twenty eight fatalities of press frenzy, congressional hearings, and, ultimately, the establishment of the essay in Nineteen Fifty Eight. Even with the ground based radars, an airtraffic control, there were still these midair collisions.

###### 09:42 At the one on the left was actually captured on film, as it happened in nineteen seventy Eight, there were a hundred forty four fatalities, and it spurred additional interest in developing a system that was later called teaches the traffic alert and collision avoided system, it was until after the nineteen eighty six collision over cerito California, few miles from Disneyland, reconquered, decided to put laws into effect. Redwood mandate, the equipage of tea couch. In the Us. By around the year, two thousand wrestler world has fallen to kick. Ass is composed of a surveillance system that detects and tracks intruders, it passes the sensor measurements onto the advisory logic that decides whether to provide an alert to the pilots, and which maneuver to advise these resolution advisory surpassed onto the pilots in the cockpit, the advisory logic contains a bunch of, if then rules, different parameters and so forth.

###### 09:53 Benign inputs, as we have seen in the literature on adversarial examples in image classification, so we wanted to see whether there was something formal that we could say about the properties of these learn networks, one property that we might want to proof, is that regions where we expect one type of advisory, provide the advisory consistently over the infinitely many points that compose that region year. In our plot, it looks like we consistently produce a descend, fifteen hundred advisory in that top rectangle, but of course the image we see is produced only by a finite sampling of those points in the rectangle, but really plexus, able to verify that this property holds, there are plenty of other kinds of properties that we might also like to prove, we found that proving properties like these can take on the order of minutes that is, for networks of the size relevant to collision avoidance, unfortunately, rarely plucks doesn't scale up to networks of the size used for assaulting image clay.

###### 10:35 So I'm, maybe you can guess as to how many pages of pseudocode is required to specify this logic are, before I tell you on audio, say, I'm a third generation pilot, my dad, rana, flight school in his dad round and flight school by even so, I didn't properly appreciate the complexity of getting this logic right.

###### 10:52 Suffocation tests, however, there has been quite a bit of follow up work on improving the efficiency of railly plaques and leading to the open source release of the marabou framework and skills better, to larger networks, but we still can't prove property is about networks with hundreds of thousands of units. Many other labs around the world have also made progress in the past few years in developing verification algorithms. My a former Postdoc, John Lulu, who is now a professor at Sea Amu, when through and surveyed all these different neural verification algorithms and reproduced, a Julia library, and a rather long survey article, is algorithms combine insights from reachability analysis, a primal and dual optimization, and search algorithms, some of these algorithms sacrifice completeness in order to achieve greater levels of scalability.

###### 10:56 I want to show you something that I found a couple years ago on Instagram.

###### 11:01 To someone on Instagram posted as something that says, are they use coding an algorithm? So the drones Jane crashing into each other, and then someone wrote, if going to crashing into each other down, and then a someone whose line about Alice, one thing says. As a robotics major, I can confirm this is one hundred percent how coding works.

###### 11:24 And I. So there is a general lack of appreciation by the general public, and perhaps also essence engineers, in the complexity of getting this right.

###### 11:34 This is a screenshots, a volume, one of our tc Idea, when Eighty Five be the tea Cast standard, the first volume has One Thousand, Seven Hundred, ninety Nine pages in Four Hundred forty of it is pseudocode. I got to know many of the brilliant people who worked on the system are, some of them had worked on it from before I was born I. It placed a huge burden on the designers to anticipate all the various edge cases and decide on away to resolve them. It required a massive engineering effort. A. Consequently, Tickuss has been a major success story.

###### 11:45 We can use these neural network verification algorithms to do some reachability analysis to prove, for an example, though if we follow the advisories given by the neural network, that will always be safe. The first up here is to discretized the state space, musing a grin.

###### 12:07 To illustrate this process, let's say we are in the selborne, a highlighted in Red Hear. It looks like the two options that are available here are either descend or climb fifteen hundred feet permanent, but we can call out to one of these neural network. Firefighters to Ashley proved that the is indeed the case.

###### 12:13 Since teaches, or was mounted in the Us on large aircraft, there have been exactly zero midair collisions involving teaches.

###### 12:22 A smaller general aviation aircraft, like the ones are used to fly, are not equipped with teaches and are. There have been several midair collisions every year on adjacent the United States.

###### 12:28 As I've indicated here, this is indeed the case.

###### 12:32 What we can do is simulate the dynamics forward using an over approximation, conditioned on the various actions.

###### 12:35 So when we were designing the next generation of this system, I spent awhile thinking about what are the major challenges in building a reliable system, and it really boils down to these three challenges, the first, when estate uncertainty we have imperfect censers are, we also have dynamic uncertainty most of the time aircraft fly straight, but they might turn left or right with some probability.

###### 12:42 In this case, were simulating descend fifteen hundred forward, and we are guaranteed to end up somewhere within these two cells.

###### 12:54 We can also simulate a climb and we end up in those hills.

###### 13:00 The Marcelle's we can look out where we might end up. We've applied this technique to study reachability for both of vertical avoidance maneuvers as well as horizontal avoidance maneuvers, this is an example of a horizontal avoidance maneuver. What we want to show is that are the intruder aircraft cannot come within.

###### 13:03 The other challenge is that there are multiple completing objectives, the system has a very carefully balance both safety and operational considerations in getting this. Balanced. Raising is extremely difficult, and it turns out that, of course, these three challenges can be framed in terms of the Poverty P framework, and we can use dynamic programming to find an approximate solution. Hello to show you. A plot of the policy that resulted from optimizing one of her first models when we first started working on this line of research. The horizontal axis here is the time to potential collision, and the vertical axis is our altitude relative to the intruder, this intruder states, fixed at zero zero.

###### 13:22 A certain threshold are thus, if we follow our advisories.

###### 13:28 Soon, Chelsea Sidran has been leading somewhere. Can. Symbolic methods for closed loop reachability analysis. The does not required discretized ation, and may be more scalable to larger problems, together with the former Postdoc of mind, may created a tool that provides relational piece wise, linear over approximations of multi dimensional functions, this over approximation can be used to represent the dynamics of the system exemple here, is of the inverted pendulum. We start in some initial region designated by that square at the bottom left of the state space, or we would like to determine, is what region we might end up in after ten times steps. One approach, of course, is to run a bunch of simulations with Iron rail network policy, and the Lupin A. We see that we end up in that orange region. Of course, we don't have any guarantees, but that's roughly where we will end up. Another approaches to use a mixed in injure problem formulation and using off the shelf salver.

###### 13:50 Ah, this is a slice, through this speak space, a higher dimensional state, space under the assumption that we have not yet issued an advisory, and both aircraft and the intruder aircraft are currently level what you'll notice here, that there is a little notch on the left hand side, so why is that it actually our first? We thought this could be a bug in our implementation, and I, it turns out there, if you fill in that gap, you actually do worse, so the reason for this is your able to lower the alert rate a bit when you delay your alert when you're roughly co altitude is possible, that the intruder, my just move out of the way, and we can lower year alert rate, which is a very important operational metric.

###### 14:27 You determined the rectangle, we will be annette each step, and then just iterate a few times. This approach is called concrete reachability, because we make the reachable set concrete after each step, another approach is to use what is called symbolic reachability, which results in a larger mixin Injure program that takes longer to solve. Their can result in a tighter bound for where the pendulum might be after. Ten steps elective briefly highlight some of the work that my student, Kyle did on Taxi Net. So Taxi Net is a neural network that has been trained to follow the centreline of a runway or a taxiway, and what we want to do is figure out whether image airs can make the aircraft depart, the runway were the taxiway. This is done in the explain simulator, which is a photo realistic aircraft simulator. The camera is mounted on.

###### 14:39 Mother metric thet it helps with is the reversal rate to Teak Allison, a Cossacks in allows you to reverse the direction of the aircraft, so let's a initially you issue a climb, I, you can later reversed the climbed to a descend, so if you'd deleah, alert a little bit if you roughly co altitude it.

###### 15:04 Reduces the likelihood that you commit to a direction that you later have to reverse, and is brings us to another reason, or perhaps where the most important metrics witches safety. Now, because aircraft are physically constrained, it takes a while for the aircraft to change direction, and if you duo reversal it eats up quite a bit of time as the aircraft are converging upon each other. Another interesting feature here is that you don't alert at all when you're within five seconds of collision, and this is because we carried over an assumption inherent in teaches read, it takes the pilots exactly five seconds to respond to their advisory, since were modelling Miss problem ants of Pomme de P. And there's a cost for alerting, it doesn't want to a crew, the cost of alerting you to a potential collision when it makes no difference in terms of the trajectories of the aircraft.

###### 15:26 Side of the aircraft end, we want to translate those images into steering canet.

###### 15:35 So we take as input, the image we down, sample it, not image the down sampled images than passed into the mule network, the neural network outputs predictions of cross truck error, as well as heading air men from nose to quantities this are fed into a simple raynier controller, the outputs of Rudder Newsreel commands. And this wren gets updated in the simulator Soviet before we started using our neural network verification tools, we serve plane with different lighting conditions are ranging from a clear days to overcast, isn't. What we saw was read the neural networks actually seem to perform extremely well. We wanted to see what would happen if we allowed up to three percent perturbation of the input image to the system, and see if those perturbed images could get the aircraft off the left side of the runway. So we actually made a series of calls out tomorrow.

###### 16:35 The to find the perturbation that would lead to the most extreme rough steering command as a run. This video, we can see the trajectory evolve on the bottom left, together with the true image, the down sampled image and the perturbed image, the perturbed image looks relatively benign as you can see. We can get the aircraft pretty far to the left side of the runway. However, are there comes a point where no allowed perturbation prevents it from correcting itself, so this approach failed to find a failure trajectory.

###### 17:12 The problem with the previous approaches, that we were only optimizing for extreme left steering, commends the key insight that Kyle had was to not constrain ourselves to extreme left steering, but rather allow for both extreme left and right steering, and then use adaptive stress testing to find the right sequence of image perturbations that would result in a failure.

###### 17:36 In these series of plots, you can see the results of are Monte Carlo's tree Search Tree with a nodes shown in red, with the a final output of the adopted stress testing, shown in blue with two percent perturbation adapted, stress testing didn't find a failure.

###### 17:55 Albert, with three percent perturbation, we were able to find a very fascinating failure here, we have a series of perturbations, the lead to buff steering commands that bring the aircraft close to the left side of the runway, then we produce a series of perturbations, the lead to write steering commends the neural network, finds itself in a situation that is quite different from the examples it was trained on, and it does not know how to recover. Here are some plots were three point, five percent perturbation and four percent perturbations, when we go up to four percent, then we can actually get the aircraft to just go off the side of the runway.

###### 18:37 They should emphasize, here that adaptive stress testing on its own were not have been able to find these failures if we just you. Adoptive stress testing with random image perturbations, new failures result. The success of this verification method is in the ability of Barboux to find adversarial, a adversarial image examples combined with the ability of adoptive stress testing to search through a sequence of actions, bringing it closer to a failure.

###### 19:09 We can actually fine more likely failures if we allow Mara Boo. To search. With different amounts of image perturbation, adaptive stress testing can find the right sequence of image perturbation levels that not only leads to or failure by also maximizes the likelihood of failure.

###### 19:26 We don't have a good model of perception likelihood, but we can say that less image perturbation is more likely than were image perturbation, so we played with different perturbation levels and looked out the main perturbation about discerned to notice the adoptive stress testing when time, the crossing of the runway centreline, to correspond to gaps between the dashed centreline, the idea being that it is easier to fool the neural network if there is less guidance from the centreline.

###### 19:56 It also like to quickly mentioned some work that my former student, Maxime Baton I did on probabilistic model, checking the idea in probabilistic model checking is that you check whether some property holds with some probability, this probability can be specified in some language, like multicolour, linear, temperate logic, there are many great tools for a probabilistic model checking. But they generally assumed full state observability, there was one exception, elaborated in a Twenty Seventeen paper that involve partial observability, but the approach they took was to discretized the belief space and turn it into a fully observed problem now, of course, this will not scale very well at all, so what Maxime did was to interface with a library that converts L. T. L. Formulas into finite state machines, than he wrote a little bit of code that combines that firing state machine with the poverty p representation of the problem. This allows us to use pompey peace dot, jail to solve, for a policy.

###### 20:55 That maximize the probability of these Althea properties, and it also tells us the probability of satisfaction. Or you can find her implementation online.

###### 21:06 Led by my students, Sydney cats, we have also started to look out using probabilistic muddled checking to provide probabilistic safety guarantees when using a neural networks controller, instead of using a uniform discotheque station of the state space, like were killed. In for his reachability analysis, she developed an adaptive discretized Asian method that minimises the volume of the input space covered by cells with multiple different actions recommended by the neural network policy. The number of actions of available from a cell can be determined by calls to Morrow, Bu. We can then use this of over approximation to arrive at a failure probability estimate from different points in the state space, we have to be careful about how we do this, since we want to preserve the formal guarantee that we want from a probabilistic model, checker, that that probability of failure is guaranteed to not accede some threshold anyway, Sydney came up with some methods to provide.

###### 22:06 Variable levels of refinement of these probabilistic guarantees, inspired by some state abstraction methods discussed in the reinforcement learning literature with additional computational time, we can afford to prove that we are safe in certain areas of the state space that we would not have been able to prove before, when we are building safety, critical systems were actually any kind of decision making or a decision support system, we want to make sure that the system is robust to the various sources of uncertainty inherent in the real world upon the p is one way to frame this kind of uncertainty, and we can use a variety of different optimization algorithms for solving them were approximating their solutions. This optimization can improve robustness in ways that are sometimes counter intuitive. We saw this with aircraft collision avoidance, and this is where these kinds of techniques can bring quite a bit of value, however, proper validation of the system is absolutely critical, establishing trust in safety critical systems.

###### 23:05 Requires a variety of different approaches, including identifying and categorizing failure moods, that there is no silver bullet here. Formal methods of difficulty scaling to problems of interest, and sampling based approaches on their own can have difficulty finding relevant failures. There are some very exciting applications of formal methods that can support the validation of controllers represented by neural networks will have to see how far we can take them are. These are still a very early days.

###### 23:37 You can find my email at the bottom, the site, as well as a link to the Web page of the Stanford Intelligent Systems Laboratory.

###### 24:06 Okay, so we're here.

###### 24:23 So, I guess you're talk just ended in the other side so we can start the question on sufficient arm.

###### 24:32 I can Michael, you may want to of a new girl, never get beg alms, tanks for greater Michael, arm and dub attendees, please arm. If you are questions either raised your hand, are of post your question on the Kewaunee, I'm that me start a with a question that I have arm, which is so I, I think you started, if I understood correctly, are you started with term? Our safety than the models are specified by us, essentially armed and and then lighting.

###### 25:11 Towards the end, you also looked at what you can do in terms are found on guarantees for learned models, presumably, so I am assuming I'm ashamed to understand, or what is the maximum we can expect in terms of guaranteeing safety of learned models other than may be checking that second failures don't occur, whereas for had no specified models, you can do lot more things like reachability analysis and all sorts of things that in alarm at known before answer, as wondering what your cake, and that is, yasser didn't. Checking safety of learned model, so I guess we should distinguish between two different things here, one is learned models, where are you then freeze what you learned are of wine and invalidated of wine, and and so forth.

###### 26:02 I've versus I verifying lika a learning system, and then you continuously learn in real time and and so forth on. I think the latter is super tricky.

###### 26:15 I thought, so I've what we did for this. A cause X, I. I relieved. Prototype will be basically take a look of people and then trained it and you'll network on it.

###### 26:28 I didn't we prose that neural networks can, and then I approve, property is about it.

###### 26:33 Now, I, I. I think the the use case for learning in real time is a little bit questionable, especially for the aerospace i, a. Sector Maria make sense in in other, in other Countdown, I'mi'm am. I'm very happy that Indiana spirits were advent into Reviler ever gets the weather in a. But I thought at that, the election trusting that you also talk about what you could do this other kinds of modest Nelson, couple of questions that me are, read them out and baby can also of them armed, so.

###### 27:09 Any thoughts on how the promotional Boston safety with lower speak system, such as recommending the wrong video, as you mention, but where the cumulative consequences of Montreville'sthat's a good point. So I did it, start a failure vedere, but over a period of time, you, I guess, may be somebody's suggesting this Facebook style recommendation thing that garters into the mess we out and write up, yeah, does the utter ethical question, so, um.

###### 27:38 I am, I think trust generally has to be built incrementally. Iso, so it's generally, but generally what you want to do, and engineering is take Beaton's baby steps, I towards the the anvil and don't comfort at as you go along on. So with a Cossacks, I for the Virgin for a manned aircraft. The the human is in the route, by its making recommendations to to the human, and then the human, can I intervene, were all of the the the guidance or not, and and thus the can be a stepping stone to greater levels of autonomy.

###### 28:19 However, even when you have recommendation systems, I you need to think very carefully about the human factors involved in the also yet you don't wanta system that the I know completely stops everything on, and then handed over to the human in in a state that the human can I recovered from are what we saw, the salute a little bit with the and caste system from elbowing and, sir, so studying that care place is super important, so this is one last question that actions have almost follows into what you're alluding to, that is the have you considered the human factor elements of such safety critical systems. How does one figure out vent to alert the humans who that such a great question, sue, I mean, is that the trade off between the the different operational performance metrics for a Cossacks, it's really trading of the the timing right, so.

###### 29:19 If you alert super early right then on the venue, you have a pretty good shot at a avoiding collision, but you'll have to tolerate a lot of no unnecessary alerts and and so forth and Gideon getting all of this right.

###### 29:39 Some is really tricky at a. M. In the sides, I only highlighted it, a few matrix, like I probably, of mere bitter collision, rehearsal rate strengthening Rick, but in reality, there a whole bunch of other metrics are just relevant to the the aircraft village importance problem and not required a lot of study. It it took on the order of, you know, two or three years trying to get the operational as suitability are ripe for a cassettes that involve bringin of a bunch of different stakeholders to inflate tests are human loop studies in in sober. So actually, I'm desperate as one more question, because it is a follow up on the first question, and envy will have to, actually Ann, is very interesting discussion for this armed, so connected to the question of this is our imports. Sturdier connected to the question on learned models, even with the Upland Learning onto Limited by the car.

###### 30:38 Last of the simulator, how do you certify these simulators, all that graze Zoe? In addition to validating the the system under test, we also need to validate the models that we used to invalidate the system under test. And that's absolutely tricky business. I can outline i. A. Couple strategies. Irene coming from a a machine learning community will be no about things like in a cross validation, measuring more likely, verdun and so forth.

###### 31:09 I'd bet I've one of the most useful approaches is, do I do? The the terrain text is conflict, it's gonna like the the fake data during the Soviet we did this with with a Cossacks, are we needed on a model that created synthetic radar tracks, and so what we did was we are, we've got a big and dynamic these in network i an incredible bunch of synthetic tracks. And then we asked I. Aviation experts and reader experts, to try to distinguish between the big radar attracts and the real radio tracks, and when we were able to get it to the point where they couldn't tell the difference between them mimic, than we felt an additional level of our comfort in our hearts, that we were on the right track, in addition to Eunoe, looking at Cale, Divergence and and Log, Likelihood, and and support, a thank you, thank you again, by a great talk, that actually, several other questions, by compassionately be done, upper.

###### 32:08 Maybe you would actually answer them on on the anthem of are, using the answering abdul are the opening trade tax again for your back on. And then I guess we will, I'm the guy.

###### 32:22 Take too much.