

# Transcript of interview to Stuart Russell by Lex Fridman

Alfonso R. Reyes

2020-01-19

## Contents

<b>Transcript of interview to Stuart Russell by Lex Fridman</b>	<b>1</b>
Stuart Russell: Long-Term Future of Artificial Intelligence   Artificial Intelligence (AI) Podcast   Dec 9, 2018 . . . . .	1
Introduction by Lex Fridman . . . . .	2
Start of Interview . . . . .	2
AI mind games: Chess, AlphaGo, AlphaZero . . . . .	2
Meta-reasoning. What made me work on AI . . . . .	4
Reasoning, AI winters and Lisp machines . . . . .	6
How to prevent another AI winter. Snake oil. Self-driving cars . . . . .	6
The unrelenting human desire of creating super-intelligence . . . . .	8
Can we stop a Super-AI? . . . . .	8
An AI system that has already taken over the world . . . . .	10
Machines and Utilitarianism . . . . .	11
The dangers of unregulated AI. Algorithms that may be contributing to the destruction of Democracy . . . . .	11
Super-AI: Recall. Once upon a time scientists thought nuclear chain reaction wasn't possible	13
Overusing AI. We've seen that movie before: Wall-E . . . . .	15
Machines can be beneficial to humans but somebody will find loopholes . . . . .	16
Favorite Sci-Fi movie. Favorite robots . . . . .	17
End of Interview . . . . .	17
References . . . . .	17
Links . . . . .	20
Toolbox . . . . .	20

## Transcript of interview to Stuart Russell by Lex Fridman

### Stuart Russell: Long-Term Future of Artificial Intelligence | Artificial Intelligence (AI) Podcast | Dec 9, 2018

This is a quick transcript of the interview of Stuart Russell by [Lex Fridman](#). The interview has been published in [YouTube](#) as well as a [podcast](#). This is a transcript I wanted to do for a long time. Stuart Russell is an authority on AI and coauthor of one of the best books on Artificial Intelligence. I have learned so much listening to the interview; checking his references; the subjects and personalities he mentioned. parallels in history that are worth considering in AI. It is simply breathtaking! . Enjoy!

*Alfonso R. Reyes. The Woodlands, Texas. Jan 18, 2020.*



Figure 1: image-20200119130037529

## Introduction by Lex Fridman

The following is a conversation with Stuart Russell. He's a professor of Computer Science at UC Berkeley and a co-author of a book that introduced me -and millions of other people-, to the amazing world of AI called [Artificial Intelligence: A Modern Approach](#). So, it was an honor for me to have this conversation as part of [MIT course and Artificial General Intelligence \(AGI\)](#), and the [Artificial Intelligence podcast](#). If you enjoy it, please subscribe on YouTube, iTunes, or your podcast provider of choice. Or simply connect with me on Twitter at [Lex Fridman](#), spelled F.R.I.D.M.A.N. And now, here's my conversation with [Stuart Russell](#).

## Start of Interview

### AI mind games: Chess, AlphaGo, AlphaZero

[Lex] You've mentioned in 1975, in high school, you've created one year first AI programs that play chess. Were you ever able to build a program that beat you at chess or another board game?

[Stuart] My program never beat me at chess. I actually wrote the program at Imperial College. I used to take the bus every Wednesday with a box of cards this big, and shove them into the card reader and they gave us eight seconds of CPU time. It took about five seconds to read the cards in and compile the code. So, we had three seconds of CPU time which was enough to make one move. You know, with a not very deep search, and then we would print that move out, and then we'd have to go to the back of the queue and wait to feed the cards in again.

[Lex] How do you post a search?

[Stuart] Well, I think we got an eight move, you know, depth eight with alpha beta and we had some tricks of our own about move ordering and some pruning of the tree ...

[Lex] And you were still able to beat that program?

[Stuart] Yeah. I was a reasonable chess player in my youth. I did a fellow program and a backgammon program. So, when I go to Berkeley I worked a lot on what we call meta-reasoning which really means reasoning about reasoning and in the case of a game playing program you need to reason about what parts

of the search tree you're actually going to explore because the search tree is enormous, or, bigger than the number of atoms in the universe. And the way programs succeed, and the way humans succeed, is by only looking at a small fraction of the search tree, and if you look at the right fraction, you play really well. If you look at the wrong fraction -if you waste your time thinking about things that are never gonna happen, the moves that no one's ever gonna make-, then you're gonna lose because you won't be able to figure out the right decision. So, that question of how machines can manage their own computation, either how they decide what to think about is the meta-reasoning question. We developed some methods for doing that and very simply a machine should think about whatever thoughts are going to improve its decision quality. We were able to show that both, for a [unintelligible], which is a standard two play game, and for backgammon which includes dice. Also it's a two-player game with uncertainty. For both of those cases we could come up with algorithms that were actually much more efficient than the standard alpha-beta search, which chess programs at the time we're using, and that those programs could beat me. And I think you can see same basic ideas in AlphaGo and AlphaZero today. The way they explore the tree is using a former meta-reasoning to select what to think about based on how useful it is to think about it.

**[Lex]** Is there any insights you can describe without Greek symbols of how do we select which paths to go down?

**[Stuart]** There's really two kinds of learning going on. So, as you say, AlphaGo learns to evaluate board positions; so it can look at a Go board and it actually has probably a superhuman ability to instantly tell how promising that situation is. To me the amazing thing about AlphaGo is not that it can be the world champion with its hands tied behind his back, but the fact that if you stop it from searching altogether - so, you say: okay you're not allowed to do any thinking ahead, right? You can just consider each of your legal moves and then look at the resulting situation and evaluate it; so, what we call a depth-one search, so just the immediate outcome of your moves and decide if that's good or bad. That version of AlphaGo can still play at a professional level, right? And human professionals are sitting there for five, ten minutes deciding what to do, and AlphaGo, in less than a second, instantly into it what is the right move to make based on its ability to evaluate positions. And that is remarkable because we don't have that level of intuition about Go. We actually have to think about the situation. So, anyway, that capability that AlphaGo has is one big part of why it beats humans. The other big part is that it's able to look ahead 40, 50, 60 moves into the future. And you know, if it was considering all possibilities -40 or 50 or 60 moves into the future-, that would be you know 10 to the 200 ( $10^{200}$ ) possibilities, so way, way more than you know atoms in the universe, and so on. So, it's very, very selective about what it looks at. So, let me try to give you an intuition about how you decide what to think about. It's a combination of two things: one is how promising it is. If you're already convinced that a move is terrible, there's no point spending a lot more time convincing yourself that it's terrible because it's probably not gonna change your mind. So, the real reason you think is because there's some possibility of changing your mind about what to do, right? And that changing your mind that would result then in a better final action in the real world, so that's the purpose of thinking is to improve the final action in the real world. And so if you think about a move that is guaranteed to be terrible, you can convince yourself is terrible, and you're still not gonna change your mind all, right? But on the other hand you I suppose you had a choice between two moves one of them you've already figured out is guaranteed to be a draw, let's say. And then the other one looks a little bit worse like it looks fairly likely that if you make that move you're gonna lose but there's still some uncertainty about the value of that move. There's still some possibility that it will turn out to be a win. Then it's worth thinking about that so even though it's less promising on average than the other move, which is guaranteed to be a draw, there's still some purpose in thinking about it because there's a chance that you will change your mind and discover that in fact it's a better move. So, it's a combination of how good the move appears to be and how much uncertainty there is about its value. The more uncertainty, the more it's worth thinking about because there's a higher upside if you want to think of it that way.

**[Lex]** And of course in the beginning, especially in the AlphaGo Zero formulation, it's everything is shrouded in uncertainty, so you're really swimming in a sea of uncertainty so it benefits you too. I mean, actually following the same process as you described but because you're so uncertain about everything you basically have to try a lot of different directions?

**[Stuart]** Yeah. So, the early parts of the search tree are fairly bushy, that it will . . . , when looking a lot of

different possibilities but fairly quickly the degree of certainty about some of the moves, I mean if moves are really terrible you'll pretty quickly find out right; you lose half your pieces or half your territory, and then you'll say "okay, this this is not worth thinking about any more". And then so a further down the tree becomes very long and narrow and you're following various lines of play, you know, 10, 20, 30, 40, 50 moves into the future, and you know that's again it's something that human beings have a very hard time doing mainly because they just lacked the short-term memory. You just can't remember a sequence of moves that's 50 moves long, and you can't you can't imagine the board correctly for that many moves into the future.

[Lex] Of course the top players- I'm much more familiar with chess-, but the top players probably have they have echoes of the same kind of intuition instinct that in a moment's time AlphaGo applies when they see a board, I mean they've seen those patterns human beings have seen those patterns before at the top at the Grandmaster level. It seems that there is some similarities, or maybe it's our imagination, creates a vision of those similarities, but it feels like this kind of pattern recognition that the AlphaGo approaches are using is similar to what human beings at the top level, what do you think?

[Stuart] I think there's some truth to that ...

10:02

[Lex] But not entirely.

[Stuart] Yeah. I mean, I think to the extent to which a human Grandmaster can reliably wreak instantly recognize the right move instantly recognize the value of a position I think that's a little bit overrated ...

[Lex] But if you sacrifice a queen, for example, I mean, there's these beautiful games of chess with Bobby Fischer, somebody where it's seeming to make a bad move, and I'm not sure there's a a perfect degree of calculation involved, where they've calculated all the possible things that happen but there's an instinct there, right? That somehow adds up to the ...

[Stuart] Yeah. I think what happens is you get a sense that there's some possibility in the position even if you make a weird-looking move that it opens up some some lines of calculation that otherwise would be definitely bad. And that intuition that there's something here in this position that might might yield a win.

[Lex] Down the side ...

[Stuart] And then you follow that ... and in some sense when a chess player is following a line, and in his, or her mind, they're they mentally simulating what the other person is gonna do while the opponent is gonna do, and they can do that as long as the moves are kind of forced, as long as there's a fourth we call a forcing variation where the opponent doesn't really have much choice how to respond, and then you see if you can force them into a situation where you win. We see plenty of mistakes even in Grandmaster games where they just miss some simple three, four, five move combination that you know wasn't particularly apparent in the position but we're still there.

## Meta-reasoning. What made me work on AI

[Lex] That's the thing that makes us human. So, when you mentioned that in a fellow those games were after some meta-reasoning improvements and research, was able to beat you, how did that make you feel?

[Stuart] Part of the meta-reasoning capability that it had was based on learning, and you could sit down the next day and you could just feel that it had got a lot smarter. All the sudden you really felt like you sort of pressed against the wall because it was it was much more aggressive and was totally unforgiving of any minor mistake that you might make, and actually it seemed understood the game better than I did. Gary Kasparov has this quote. During his match against Deep Blue he said he suddenly felt that there was a new kind of intelligence across the board.

[Lex] Do you think that's a scary or an exciting possibility? As for Kasparov, and for yourself, in the context of chess, purely; sort of in this like that feeling whatever that is?

[Stuart] I think it's definitely an exciting feeling. This is what made me work on AI in the first place. As soon as I really understood what a computer was I wanted to make it smart. I started out with the first program I

wrote was for the Sinclair programmable calculator. And i think you could write a 21 step algorithm that was the biggest program you could write, something like that. And do little arithmetic calculations, so, I say think I implemented Newton's method for square roots and a few other things like that. But then I thought: okay, if I just had more space I could make this thing intelligent. And so I started thinking about AI. I think the thing that's scary is not the chess program because you know chess programs they're not in they're taking over the world business. But if you extrapolate ... there are things about chess that don't resemble the real world, right? We know the rules of chess. A chess board is completely visible to the programmer of course the real world is not most you most the real world is not visible from wherever you're sitting, so to speak, and to overcome those kinds of problems you need qualitatively different algorithms. Another thing about the real world is that you know we regularly plan ahead on the timescales involving billions or trillions of steps. Now we don't plan that was in detail but you know when you choose to do a PhD at Berkeley, that's a five-year commitment and that amounts to about a trillion motor-control steps that you will eventually be committed to.

[Lex] Including going up the stairs, opening doors, drinking water type

[Stuart] I mean every every finger movement while you're typing every character of every paper, and the thesis, and everything else. So you're not committing in advance to the specific motor-control steps but you're still reasoning on a timescale that will eventually reduce to trillions of motor-control actions. And for all these reasons you know AlphaGo and and Deep Blue, and so on, don't represent any kind of threat to humanity, but they are a step towards it. Progress in AI occurs by essentially removing one by one these assumptions that make problems easy like the assumption of complete observability of the situation. We remove that assumption, you need a much more complicated kind of a computing design, and you need something that actually keeps track of all the things you can't see and tries to estimate what's going on, and there's inevitable uncertainty in that; so it becomes a much more complicated problem. But we are removing those assumptions. We are starting to have algorithms that can cope with much longer timescales, they can cope with uncertainty, they can cope with partial observability, and so each of those steps sort of magnifies by a thousand the range of things that we can do with AI systems.

[Lex] The way I started wit AI: I wanted to be a psychiatrist for long time to understand the mind in High School and of course program and so on. And then I showed up University of Illinois to an AI lab, and they said okay I don't have time for you but here's a book "AI a Modern Approach" I think was the first edition at the time. Here, go learn this, and I remember the lay of the land was, well it's incredible that we solve Chess but we'll never solve Go. I mean it was pretty certain that Go in the way we thought about systems that reason was impossible to solve and now we've solved this as a very ...

[Stuart] I would have said that it's unlikely we could take the kind of algorithm that was used for chess, and just get it to scale up and work well for Go. And, at the time what we thought was that in order to solve Go we would have to do something similar to the way humans manage the complexity of Go, which is to break it down into kind of sub-games. When a human thinks about a Go board they think about different parts of the board as sort of weakly connected to each other. And they think about "okay, within this part of the board here's how things could go, and that part about this how things could go". And now you try to sort of couple those two analyses together, and deal with the interactions, and maybe revise your views of how things are going to go in each part and then you've got maybe five, six, seven, ten parts of the board. That actually resembles the real world much more than chess does because in the real world you know we have work, we have home life, we have sports, whatever different kinds of activities; shopping. These all are connected to each other but they're weakly connected. When I'm typing a paper you know I don't simultaneously have to decide which order I'm gonna get the you know the milk and the butter you know that doesn't affect the typing but I do need to realize "okay, better finish this before the shops closed", because I don't have anything you don't have any food at home. There's some weak connection but not in the way that chess works where everything is tied into a single stream of thought. So, the thought was that Go, just sort of Go we'd have to make progress on stuff that would be useful for the real world and in a way AlphaGo is a little bit disappointing because the program designed for AlphaGo was actually not that different from from Deep Blue, or even from Arthur Samuels checker playing program from the 1950s. And in fact the the two things that make AlphaGo work is: one, it's amazing ability to evaluate the positions, and the other is the meta-reasoning capability, which allows it to explore some paths in the tree very deeply and to abandon

other paths very quickly.

20:08

### Reasoning, AI winters and Lisp machines

[Lex] So, this word meta-reasoning. while technically correct, inspires perhaps the wrong degree of power that AlphaGo has. For example, the word reasoning is a powerful word. Let me ask you, you were part of the symbolic AI world for a while like whatever the AI was-, there's a lot of excellent interesting ideas there, that unfortunately met a winter, do you think it re-emerges?

[Stuart] Well, I would say ... yeah. It's not quite as simple as that ... the first AI winter, that was actually named as such, was the one in the late 80s, and that came about because in the mid 80s there was a really a concerted attempt to push AI out into the real world, using what was called expert system technology. And for the most part that technology was just not ready for prime time. They were trying, in many cases, to do a form of uncertain reasoning judgment combinations of evidence diagnosis those kinds of things which was simply invalid. And when you try to apply invalid reasoning methods to real problems -you can fudge it for small versions of the problem-, but when it starts to get larger, the thing just falls apart. Many companies found that the stuff just didn't work and they were spending tons of money on consultants to try to make it work, and there were other practical reasons like they were asking the companies to buy incredibly expensive [Lisp](#) machine workstations which were, literally, between fifty and a hundred thousand dollars, in 1980s money, which would be like between a hundred and fifty and three hundred thousand dollars per workstation in current prices.

[Lex] So, then the bottom line, they weren't seeing a profit from it?

[Stuart] Yeah, in many cases. I think there were some successes, there's no doubt about that. But people, I would say, over-invested. Every major company was starting an AI department, just like now. And I worry a bit that we might see similar disappointments, not because the technology is invalid but it's limited in its scope, and it's almost the [unintelligible] of ... the scope problems that expert systems had so ...

### How to prevent another AI winter. Snake oil. Self-driving cars

[Lex] What have you learned from that hype cycle, and what can we do to prevent another winter, for example?

[Stuart] Yeah. When I'm giving talks these days that's one of the warnings that I give. Two warning slides: one is that rather than data being the new oil, data is the new "snake oil".

[Lex] That's a good line!

[Stuart] And then, the other is that we might see a kind of very visible failure in some of the major application areas. And I think self-driving cars would be the flagship. I think, when you look at the history, so the first self-driving car was on the freeway driving itself changing lanes overtaking in 1987. So, it's more than 30 years and that kind of looks like where we are today right you know prototypes on the freeway changing lanes and overtaking. Now, I think significant progress has been made particularly on the perception side. We worked a lot on autonomous vehicles in the early mid 90s at Berkeley. We had our own big demonstrations. We we put congressmen into self-driving cars and had them zooming along the freeway. And the problem was clearly perception.

[Lex] At the time the problem that perception ...

[Stuart] Yeah. In simulation, with perfect perception, you could actually show that you can drive safely for a long time; even if the other cars are misbehaving, and so on. But simultaneously we worked on machine vision for detecting cars and tracking pedestrians, and so on. And we couldn't get the reliability of detection and tracking up to a high enough particular level particularly in bad weather conditions nighttime, rainfall ...

[Lex] Good enough for demos but perhaps not good enough to cover the general ...

[Stuart] Yeah. The thing about driving is: suppose you're a taxi driver. You drive every day eight hours a day for ten years, right? That's a hundred million seconds of driving, and any one of those seconds you can make a fatal mistake. So, you're talking about eight nines of reliability, right? Now, if your vision system only detects ninety eight point three percent of the vehicles, right? That's sort of one on a bit nines of reliability. So, you have another seven orders of magnitude to go. And this is what people don't understand. They think "oh, because I had a successful demo I'm pretty much done". But you know you're not even within seven orders of magnitude of being done. And that's the difficulty, and it's not that can I follow a white line. That's not the problem, right? We follow a white line all the way across the country but it's the the weird stuff that happens ...

[Lex] it's some of the edge cases

[Stuart] Yeah, the edge cases. Other drivers doing weird things. So, if you talk to Google, right. So, they had actually very classical architecture where you had machine vision which would detect all the other cars and pedestrians, and the white lines, and the road signs. And then basically that was fed into a logical database. And then you had a classical 1970s rule-based expert system telling you: "okay, if you're in the middle lane and there's a bicyclist in the right lane who is signaling this then, don't need to do that". Yeah, right. And what they found was that every day they go out and there'd be another situation that the rules didn't cover, so they they come to a traffic circle and there's a little girl riding a bicycle the wrong way around a traffic circle. Okay, what do you do we don't have a rule? Oh, my god, okay, stop. And then they come back and had more rules and they just found that this was not really converging. And if you think about it right, how do you deal with an unexpected situation, meaning one that you've never previously encountered and the sort of the the reasoning required to figure out the solution for that situation has never been done, it doesn't match any previous situation in terms of the kind of reasoning you have to do. Well, in chess programs this happens all the time. You're constantly coming up with situations you haven't seen before and you have to reason about them, you have to think about okay here are the possible things I could do here the outcomes here's how desirable the outcomes are and then pick the right one you know. In the 90s we were saying: "okay, this is how you're gonna have to do automated vehicles: they're gonna have to have a look ahead capability". But the look ahead for driving is more difficult than it is for chess because. Humans are less predictable.

[Lex] than chess ...

[Stuart] You have an opponent in chess, who's also somewhat unpredictable. But for example in chess you always know the opponent's intention: they're trying to beat you. Whereas in driving you don't know is this guy trying to turn left, or has he just forgotten to turn off his turn signal, or is he drunk, or is he you know changing the channel on his radio, or whatever it might be you got to try and figure out the mental state the intent of the other drivers to forecast the possible evolutions of their trajectories and then you've got to figure out okay which is the trajectory for me that's going to be safest, and those all interact with each other because the other drivers going to react to your trajectory, and so on. So, they've got the classic merging onto the freeway, a problem where you're kind of racing a vehicle that's already on the freeway, and you are you gonna pull ahead of them, or you're gonna let them go first and pull in behind. And you get this sort of uncertainty about who's going first. So all those kinds of things mean that you need decision-making architecture that's very different from either a rule-based system, or it seems to me a kind of an end-to-end neural network system you know. So just as AlphaGo is pretty good when it doesn't do any look ahead but it's way, way, way, way better when it does, I think the same is going to be true for driving. You can have a driving system that's pretty good when it doesn't do any look ahead, but that's not good enough. We've already seen multiple deaths caused by poorly designed machine learning algorithms that don't really understand what they're doing.

30:10

[Lex] Yeah. On several levels, I think it's on the perception side there's mistakes being made by those algorithms where the perception is very shallow. On the planning side, to look ahead like you said, and the thing that we come come up against. That's really interesting when you try to deploy systems in the real world is you can't think of an artificial intelligence system as a thing that responds to the world always. You have to realize that it's an agent that others will respond to as well so in order to drive successfully you can't



just try to do obstacle avoidance.

[Stuart] You can't pretend that you're invisible car, right.

[Lex] I mean ... but you have to assert yet others have to be scared of you. Just we're all there's this tension there's this game so if we studied a lot of work with pedestrians. If you approach pedestrians as purely an obstacle avoidance so you either doing look ahead isn't modeling the intent that you're ... they're going to take advantage of you they're not going to respect you at all. There has to be a tension, a fear, some amount of uncertainty that's how we have create, we ...

[Stuart] Or at least, just a kind of a resoluteness. So you have to display a certain amount of resoluteness, you can't you can't be too tentative. The solutions then become pretty complicated. You get into game theoretic, analyses. We're at Berkeley now we're working a lot on this kind of interaction between machines and humans.

[Lex] And that's exciting

[Stuart] And so my colleague Anca Dragan, actually, if you formulate the problem game theoretically and you just let the system figure out the solution, it does interesting, unexpected things like sometimes at a stop sign. If no one is going first right, the car will actually back up a little all right and just to indicate to the other cars that they should go, and that's something it invented entirely by itself.

[Lex] That's interesting

[Stuart] We didn't say this is the language of communication at stop signs; it figured it out.

### **The unrelenting human desire of creating super-intelligence**

[Lex] That's really interesting stuff. So let me one just step back for a second. Just this beautiful philosophical notion. So, [Pamela McCorduck](#) in 1979 wrote "AI began with the ancient wish to forge the gods". So, when you think about the history of our civilization do you think that there is an inherent desire to create -let's not say gods-, but to create super intelligence. Is it inherent to us? Is it in our genes, that the natural arc of human civilization, is to create things that are of greater and greater power, and perhaps no echoes of ourselves so to create the gods as Pamela said?

[Stuart] Maybe. We're all individuals. Certainly we see over and over again in history individuals who thought about this possibility.

[Lex] Hopefully, when I'm not being too philosophical here but if you look at the arc of this where this is going, and we'll talk about AI safety, we'll talk about greater and greater intelligence, do you see that there in when you created the [unintelligible] program and you felt this excitement. What was that excitement? Was it excitement of a tinkerer who created something, cool like a clock? Or was there a magic, or was it more like a child being born?

[Stuart] Yeah. I certainly understand that viewpoint. And if you look at the [Lighthill Report](#), which was .... In the 70s there was a lot of controversy in the UK about AI and whether it was for real, and how much the money the government should invest. There was a lot long story but the government commissioned a report by Lighthill, who was a physicist, and he wrote a very damning report about AI which I think was the point. And he said that these are frustrated men who unable to have children would like to create, and you know create life, as a kind of replacement, which I think is really pretty unfair. But there there is a kind of magic I would say you when you you build something, and what you're building in is really, just you're building in some understanding of the principles of learning, and decision-making, and to see those principles actually then turn into intelligent behavior in specific situations it's an incredible thing. And that is naturally going to make you think: "okay. where does this end?"

### **Can we stop a Super-AI?**

[Lex] And so there's a there's magical optimistic views of word and whatever your view of optimism is, whatever your view of utopia is, it's probably different for everybody. But you've often talked about concerns you have of how things might go wrong. So I've talked to [Max Tegmark](#), there's a lot of interesting ways to



think about AI safety. You're one of the seminal people thinking about this problem among sort of being in the weeds of actually solving specific AI problems, you also think about the big picture of where we're going. So, can you talk about several elements of it let's just talk about, maybe the control problem, so this idea of losing ability to control the behavior of an AI system. So how do you see that? How do you see that coming about? What do you think we can do to manage it?

[Stuart] It doesn't take a genius to realize that if you make something that's smarter than you might have a problem. Turing, [Alan Turing](#), wrote about this and gave lectures about this. In 1951 painted a lecture on the radio and he basically said "once the machine thinking method stops, very quickly they'll outstrip humanity. If we're lucky we might be able to turn off the power at strategic moments but even so a species would be humbled". I think he was wrong about that. If it's a sufficiently intelligent machine is not gonna let you switch it off, so it's actually in competition with you.

[Lex] What do you think he's meant -just for a quick tangent-, if we shut off this super intelligent machine that our species will be humbled?

[Stuart] I think he means that we would realize that we are inferior. That we only survive by the skin of our teeth because we happen to get to the off switch, just in time, and if we hadn't then we would have lost control over the earth.

[Lex] Are you more worried when you think about this stuff about super intelligent AI, or are you more worried about super powerful AI that's not aligned with our values? So the paper clip scenario is kind of ...

[Stuart] The main problem I'm working on is is the control problem of machines pursuing objectives that are as you say not aligned with human objectives. And this has been the way we've thought about AI since the beginning. You build a machine for optimizing, and then you put in some objective, and it optimizes. And we can think of this as the the [King Midas](#) problem. King Midas put in this objective: everything I touch you turned to gold and the gods you know that's like the machine they said okay done. You now have this power and of course his food, and his drink, and his family all turned to gold and then he's dies of misery and starvation. And this is a warning, it's it's a failure mode that pretty much every culture in history has had some story along the same lines. There's the the genie that gives you three wishes and you know third wish is always you know please undo the first two wishes because I messed up. When [Arthur Samuel](#) wrote his checker playing program which learned to play checkers considerably better than Arthur Samuel could play and actually reached a pretty decent standard. [Norbert Wiener](#), who was a one of the major mathematicians of the 20th century, sort of a father of modern automation control systems. He saw this and he basically extrapolated you know as Turing did, and said okay this is how we could lose control. And specifically that we have to be certain that the purpose we put into the machine as the purpose which we really desire, and the problem is we can't do that.

40:59

[Lex] You mean we're not ... it's a very difficult to encode so to put our values on paper is really difficult, or you're just saying it's impossible?

[Stuart] Theoretically it's possible but in practice it's extremely unlikely that we could specify correctly in advance the full range of concerns of humanity ...

[Lex] You talked about cultural transmission of values I think is how humans to human transmission of values happens, right?

[Stuart] What we learned, ... as we grow up we learn about the values that matter; how things should go, what is reasonable to pursue, and what isn't reasonable to pursue.

[Lex] Machines can learn in the same kind of way?

[Stuart] Yeah. I think that what we need to do is to get away from this idea that you build an optimizing machine, and you put the objective into it, because if it's possible that you might put in a wrong objective. And we already know this is possible because it's happened lots of times, right? That means that the machine should never take an objective that's given as gospel truth because once it takes them the objective is gospel truth, then it believes that whatever actions it's taking in pursuit of that objective are the correct things to

do. So you could be jumping up and down and saying: “No, no, no; you’re gonna destroy the world”. But the machine knows what the true objective is and it’s pursuing it and tough luck to you. And this is not restricted to AI right this is you know I think many of the 20th century technologies right. In statistics you minimize a loss function. The loss function is exogenously specified in control theory. You minimize a cost function, in operations research, you maximize a reward function, and so on. So, in all these disciplines this is how we conceive of the problem, and it’s the wrong problem because we cannot specify with certainty the correct objective. We need uncertainty, we need the machine to be uncertain about what it is that it’s supposed to be maximizing.

[Lex] it’s my favorite idea of yours I’ve heard you say somewhere -well, I shouldn’t pick favorites- but it just sounds beautiful “we need to teach machines humility”, It’s a beautiful way to put it. I love it.

[Stuart] They know that they don’t know what it is they’re supposed to be doing and that those objectives I mean they exist. They are within us. But we may not be able to explicate them we may not even know you know how we want our future to go so

[Lex] Exactly

[Stuart] And the machine, ... a machine that’s uncertain he’s going to be deferential to us. So if we say don’t do that. Well, now the machines learn something a bit more about our true objectives because something that it thought was reasonable in pursuit of our objectives turns out not to be so now it’s learn something. So it’s going to defer because it wants to be doing what we really want and you know. That point, I think, is absolutely central to solving the control problem, and it’s a different kind of AI when you take away this idea that the objective is known. Then, in fact a lot of the theoretical frameworks that we’re so familiar with: [Markov decision](#) processes, goal based planning, standard games research, all of these techniques actually become inapplicable. And you get a more complicated problem because now the interaction with the human becomes part of the problem because the human, by making choices, is giving you more information about the true objective and that information helps you achieve the objective better. And that really means that you’re mostly dealing with game theoretic problems where you’ve got the machine and the human, and they’re coupled together rather than a machine going off by itself with a fixed objective.

### **An AI system that has already taken over the world**

[Lex] Which is fascinating on the machine and the human level that we when you don’t have an objective means you’re together coming up with an objective I mean there’s a lot of philosophy that you know you could argue that life doesn’t really have meaning we together agree on what gives it meaning and we kind of culturally create things that give why the heck we are in this earth anyway. We together as a society create that meaning and you have to learn that objective and one of the biggest I thought that’s what you were gonna go for a second. One of the biggest troubles we’ve run into outside of statistics and machine learning and AI and just human civilization is when you look at I came from the south was born in the Soviet Union and the history of the 20th century we ran into the most trouble as humans when there was a certainty about the objective and you do whatever it takes to achieve that objective whether you talking about in Germany, or communist Russia, ... the trouble ...

[Stuart] I would say with corporations. In fact some people argue that we don’t have to look forward to a time when AI systems take over the world they already have, and they are called corporations. That corporations happen to be using people as components right now but they are effectively algorithmic machines and they’re optimizing an objective which is quarterly profit that isn’t aligned with overall well-being of the human race and they are destroying the world. They are primarily responsible for our inability to tackle climate change. I think that’s one way of thinking about what’s going on with with cooperations but I think the point you’re making you is valid that there are there are many systems in the real world where we’ve sort of prematurely fixed on the objective and then decoupled the machine from those that’s supposed to be serving and I think you see this with government. Government is supposed to be a machine that serves people but instead it tends to be taken over by people who have their own objective and use government to optimize that objective regardless of what people want.

## Machines and Utilitarianism

[Lex] Do you have, do you find appealing the idea of almost arguing machines where you have multiple AI systems with a clear fixed objective. We have in government the red team and the blue team that are very fixed on their objectives and they argue and it kind of maybe it would disagree but it kind of seems to make it work somewhat that the the duality of it. Let's go a hundred years back when there was still was going on, or at the founding of this country, there was disagreement, and that disagreement is where so there's a balance between certainty and forced humility because the power was distributed.

[Stuart] I think that the nature of debate and disagreement argument takes as a premise the idea that you could be wrong, which means that you're not necessarily absolutely convinced that your objective is the correct one. If you were, absolutely there'll be no point in having any discussion or argument because you would never change your mind, and there wouldn't be any sort of synthesis, or anything like that. So, I think you can think of argumentation as an implementation of a form of uncertain reasoning. I've been reading recently about [utilitarianism](#) in the history of efforts to define, in a sort of clear mathematical way, I feel like a formula for moral or political decision-making. And it's really interesting that the parallels between the philosophical discussions going back 200 years, and what you see now in discussions about existential risk, because it's almost exactly the same. So someone would say: "okay, well here's a formula for how we should make decisions". So, in utilitarianism roughly each person has a utility function and then we make decisions to maximize the sum of everybody's utility. And then people point out: well, in that case the best policy is one that leads to the enormously vast population, all of whom are living a life that's barely worth living. This is called the repugnant conclusion. Another version is you know that we should maximize pleasure and that's what we mean by utility and then you'll get people effectively saying well in that case you know we might as well just have everyone hooked up to a heroin drip. They didn't use those words but that debate, what's happening in the 19th century, as it is now about AI, that if we get the formula wrong you know we're going to have AI systems working towards an outcome, that in retrospect, would be exactly wrong.

51:22

## The dangers of unregulated AI. Algorithms that may be contributing to the destruction of Democracy

[Lex] Do you think there's ... it has beautifully put, so, the echoes are there, but do you think -I mean if you look at [Sam Harris](#)-, is our imagination worries about the AI version of that because of the speed at which the things going wrong in the utilitarian context could happen. Is that a worry for you?

[Stuart] Yeah, I think that in most cases, not in all, if we have a wrong political idea, we see it starting to go wrong, and we're not completely stupid. And so we said "okay, that was maybe a mistake, let's try something different". And also we're very slow and inefficient about implementing these things, and so on. So you have to worry when you have corporations or political systems that are extremely efficient. But when we look at AI systems, or even just computers in general, right? They have this different characteristic from ordinary human activity in the past. So let's say you were a surgeon you had some idea about how to do some operation, right? Well, and let's say you were wrong that that way of doing the operation would mostly kill the patient. Well, you'd find out pretty quickly like after three, maybe three or four tries, right? But that isn't true for pharmaceutical companies because they don't do three or four operations; they manufacture three or four billion pills, and they sell them, and then they find out, maybe six months, or a year later, that "oh, people are dying of heart attacks or getting cancer from this drug". That's why we have the [FDA](#), right? Because of the scalability of pharmaceutical production. There have been some unbelievably bad episodes in the history of pharmaceuticals, and adulteration of products, and so on, that have killed tens of thousands, paralyzed hundreds of thousands of people. Now, with computers we have that same scalability problem. That you can sit there and type for "I equals 1 to 5 billion", doo! And all of a sudden you're having an impact on a global scale. And yet we have no FDA. There's absolutely no controls at all over what a bunch of undergraduates with too much caffeine can do to the world. And, you know, we look at what happened with Facebook. Well, social media in general, and click-through optimization. So you have a simple feedback algorithm that's trying to just optimize click-through. That sounds reasonable because you don't want to be feeding people ads that they don't care about -I'm not interested in. And you might even think of that

process as simply adjusting the feeding of ads, or news, articles, or whatever it might be, to match people's preferences, which sounds like a good idea. But in fact that isn't how the algorithm works. You make more money, the algorithm makes more money if it could better predict what people are going to click on because then it can feed them exactly that. So, the way to maximize click-through is actually to modify the people; to make them more predictable. And one way to do that is to feed them information which will change their behavior and preferences towards extremes that make them predictable. Now, whatever is the nearest extreme, or the nearest predictable point, that's where you're going to end up, and the machines will force you there. Now, I think there's a reasonable argument to say that this, among other things, is contributing to the destruction of democracy in the world. And where was the oversight of this process? Where were the people saying "okay, you would like to apply this algorithm to five billion people on the face of the earth", can you show me that it's safe? Can you show me that it won't have various kinds of negative effects? No, there was no one asking that question; there was no one placed between the undergrads with too much caffeine and the human race. Well, it's just, they just did it.

[Lex] And but some -way outside the scope of my knowledge-, economists would argue that it's the invisible hand, so the capitalist system, it was the oversight. So, if you're going to corrupt society, with whatever decision you make as a company, then that's going to be reflected in people not using your product. Sort of one model of oversight ...

[Stuart] We shall see. But in the meantime you might even have broken the political system that enables capitalism to function.

[Lex] Well, you've changed it, and so ...

[Stuart] We shall see.

[Lex] Change is often painful. So, my question is ... Absolutely, it's fascinating. You're absolutely right that there is zero oversight on algorithms that can have a profound civilization changing effect. Do you think it's possible -have you seen in government- to create regulatory bodies of oversight over AI algorithms, which are inherently such cutting edge set of ideas and technologies?

[Stuart] Yeah. But I think it takes time to figure out what kind of oversight, what kinds of controls. It took time to design the FDA regime, and some people still don't like it, and they want to fix it. And I think there are clear ways that it could be improved. But the whole notion that you have stage 1, stage 2, stage 3, and here are the criteria for what you have to do to pass a stage 1 trial. We haven't even thought about what those would be for algorithms. I think there are things we could do right now with regard to bias, for example. We have a pretty good technical handle on how to detect algorithms that are propagating bias that exists in datasets, how to debias those algorithms, and even what it's going to cost you to do that. So, I think we could start having some standards on that. I think there are things to do with impersonation of falsification that we could work on. A very simple point. Impersonation is a machine acting as if it was a person. I can't see a real justification for why we shouldn't insist that machines self-identify as machines. Where is the social benefit in fooling people into thinking that this is really a person when it isn't? I don't mind if it uses a human-like voice that's easy to understand; that's fine. But it should just say "I'm a machine", in some some form.

[Lex] Many people are speaking to that. I would think ... relatively obvious factors ...

59:28

[Stuart] There is actually a [law in California](#) that bans impersonation, but only in certain restricted circumstances. So, for the purpose of engaging in a [unintelligible] transaction, and for the purpose of modifying someone's voting behavior. Those are the circumstances where machines have to self-identify. But I think this, arguably, it should be in all circumstances. And then, when you talk about deep fakes, we're just beginning. But already it's possible to make a movie of anybody saying anything in ways that are pretty hard to detect.

[Lex] Including yourself, because you're on camera now and your voice is coming through with high resolution ...

[Stuart] So, you could take what I'm saying and replace it with it pretty much anything else you wanted me to be saying. And even it will change my lips and expression expressions to fit. There's actually not much in the way of real legal protection against that. I think in the commercial area you could say "yeah, you're using my brand", and so on. There are rules about that. But in the political sphere, I think, at the moment it's anything goes. That could be really, really damaging.

### **Super-AI: Recall. Once upon a time scientists thought nuclear chain reaction wasn't possible**

[Lex] Let me just try to make, not an argument, but try to look back at history, and say something dark in essence. While regulation seems to be . . . , oversight seems to be exactly the right thing to do here. It seems that human beings, what they naturally do is: they wait for something to go wrong. If you're talking about nuclear weapons, you can't talk about nuclear weapons being dangerous until somebody, actually, like the United States [drops the bomb](#), or [Chernobyl melting](#). Do you think we will have to wait for things going wrong, in a way that's obviously damaging to society, not an existential risk, but obviously damaging?

[Stuart] I hope not. But I think we do have to look at history. The two examples you gave, nuclear weapons and nuclear power, are very, very interesting because nuclear weapons we knew in the early years of the 20th century that atoms contained a huge amount of energy. We had  $e = mc^2$ ; we knew the mass differences between the different atoms and their components, and we knew that you might be able to make an incredibly powerful explosive. So, [H. G. Wells](#) wrote a science fiction book, I think, in 1912. [Frederick Soddy](#) who was the guy who discovered [isotopes](#), Nobel Prize winner, he gave a speech in 1915 saying that one pound of this new explosive would be the equivalent of 150 tons of dynamite, which turns out to be about right. And this was in World War I. So, he was imagining how much worse the world would be if we were using that kind of explosive. But the physics establishment simply refused to believe that these things could be made.

[Lex] Including the people who were making it

[Stuart] Well, so they were doing the nuclear physics . . .

[Lex] I mean, eventually were the ones who made it . . .

[Stuart] Well, up to the development it was mostly theoretical. It was people using sort of primitive kinds of particle acceleration, and doing experiments at the level of single particles, or collections of particles. They weren't yet thinking about how to actually make a bomb, or anything like that. But they knew the energy was there, and they figured if they understood it better it might be possible. But the physics establishment -their view-, and I think because they did not want it to be true, their view was that it could not be true; that this could not provide a way to make a super weapon. And there was this famous speech given by [Rutherford](#), who was the sort of leader of nuclear physics, and it was on September 11th 1933. And he said, "anyone who talks about the possibility of obtaining energy from transformation of atoms is talking complete moonshine". And the next morning [Leo Szilard](#) read about that speech, and then invented the nuclear chain reaction. And, as soon as he invented, as soon as he had that idea, that you could make a chain reaction with neutrons, because neutrons were not repelled by the nucleus, so they could enter the nucleus, and then continue the reaction. As soon as he had that idea, he instantly realized that the world was in deep doo-doo. Because this is 1933. Hitler had recently come to power in Germany. Szilard was in London, and eventually became a refugee and came to the US. And in the process of having the idea about the chain reaction, he figured out, basically, how to make a bomb, and also how to make a reactor. And he patented the reactor in 1934. But because of the situation, the great power conflict situation, that he could see happening, he kept that a secret. And, so, between then, and the beginning of World War II, people were working, including the Germans, on how to actually create neutron sources; what specific fission reactions would produce [neutrons](#) of the right energy to continue the reaction. And that was demonstrated in Germany, I think in 1938, if I remember correctly. The first nuclear weapon patent was 1939 by the French. So, this was actually you know this was actually going on well before World War II really got going. And then the British probably had the most advanced capability in this area but for safety reasons, among others, and which is sort of just resources, they moved the program from Britain to the US, and then that became [Manhattan Project](#). So, the reason why we couldn't have any kind of oversight of nuclear weapons and nuclear technology, was because we were basically already in an arms race in a war.



[Lex] But you've mentioned then in the 20s and 30s. So, what are the echoes . . . , the way you've described this story, there's clearly echoes. Why do you think most AI researchers, folks who are really close to the metal, they really are not concerned about it, and they don't think about it, whether they don't want to think about it. Why do you think that is? What are the echoes of the nuclear situation to the current situation? And what can we do about it?

[Stuart] I think there is a kind of a motivated cognition, which is a term in psychology, which means that you believe what you would like to be true rather than what is true. And, it's unsettling to think that what you're working on might be the end of the human race, obviously. So, you would rather instantly deny it, and come up with some reason why it couldn't be true. I have I collected a long list of reasons, that extremely intelligent competent AI scientists have come up with, for why we shouldn't worry about this. For example, calculators are super human at arithmetic and they haven't taken over the world. So, there's nothing to worry about. Well, okay my five-year-old could have figured out why that was an unreasonable, and it's really quite weak argument. Another one was: while it's theoretically possible that you could have superhuman AI destroy the world, it's also theoretically possible that a black hole could materialize right next to the earth and destroy humanity. Yes, it's theoretically possible; quantum theoretically extremely unlikely that it would just materialize right there. But that's a completely bogus analogy. Because if the whole physics community on earth was working to materialize a black hole in near Earth orbit, wouldn't you ask them: is that a good idea? Is that is gonna be safe? What if you succeed?

[Lex] Right.

[Stuart] And that's the thing. The AI community sort of refused to ask itself: what if you succeed? And initially, I think, that was because it was too hard. But Alan Turing asked himself that, and he said we'd be toast. If we were lucky we might be able to switch off the power but probably we'd be toast.

[Lex] But there's also an aspect, that because we're not exactly sure what the future holds, it's not clear, exactly. So, technically what to worry about. Sort of how things go wrong, and so there is something it feels like, . . . maybe you can correct me if I'm wrong, but there's something paralyzing about worrying about something that's logically is inevitable but you don't really know what that will look like.

70:11

[Stuart] Yeah. I think that it's a reasonable point. And it's certainly in terms of existential risks it's different from asteroid colliding with the earth, which again is quite possible. It's happened in the past; it'll probably happen again. We don't know right now, but if we did detect an asteroid that was going to hit the earth in 75 years time, we'd certainly be doing something about it.

[Lex] Well, it's clear there's got big rocks. We'll probably have a meeting: what do we do about the big rock. With AI?

[Stuart] With AI, there are very few people who think it's not gonna happen within the next 75 years. I know, [Rod Brooks](#) doesn't think it's gonna happen, maybe Andrew Ng doesn't think it's gonna happen. But a lot of the people who work day-to-day, as you say at the rock face, they think it's gonna happen. I think the median estimate from AI researchers is somewhere in forty to fifty years from now, or maybe a little more. I think in Asia they think it's gonna be even faster than that. I am a little bit more conservative. I think probably take longer than that. But I think it's as happened with nuclear weapons.

[Lex] Well, it went overnight.

[Stuart] it can happen overnight. That you have these breakthroughs. And we need more than one breakthrough. But it's on the order of half a dozen. This is a very rough scale but so half a dozen breakthroughs of that nature, it would have to happen for us to reach the superhuman AI. But the AI research community is vast now. The massive investments from governments, from corporations, tons of really, really smart people. You just have to look at the rate of progress in different areas of AI to see that things are moving pretty fast. To say "oh, it's just gonna be thousands of years", I don't see any basis for that. I see for example the [Stanford Hundred Year AI project](#), which is supposed to be sort of the serious establishment view. Their most recent report actually said it's probably not even possible.

[Lex] Oh, wow.

[Stuart] If you want a perfect example of people in denial, that's it. Because for the whole history of AI we've been saying to philosophers who said it wasn't possible. Well, you have no idea what you're talking about. Of course it's possible. Give me an argument for why it couldn't happen, and there isn't one. And now because people are worried that maybe, oh it might get a bad name, or I just don't want to think about this, they're saying okay well of course it's not really possible. Imagine if the leaders of the cancer biology community got up and said "well, of course curing cancer it's not really possible". Complete outrage and dismay. I find this really a strange phenomenon. So, okay, so if you accept it as possible, and if you accept that it's probably going to happen the point that you're making that how does it go wrong, a valid question. Without an answer to that question then you're stuck with what I call the [Gorilla Problem](#), which is the problem that the gorillas face. They made something more intelligent than them namely us a few million years ago, and now they're in deep doo-doo. Yeah, so there's really nothing they can do; they've lost the control theater. They failed to solve the control problem of controlling humans. And, so they've lost. So, we don't want to be in that situation. And if the gorillas problem is the only formulation you have, there's not a lot you can do, other than to say "okay, we should try to stop, we should just not make the humans", or in this case, not make the AI. And, I think that's really hard to do. I'm not actually proposing that that's a feasible course of action. I also think that if properly controlled, AI could be incredibly beneficial. So, it seems to me that there's a consensus that one of the major failure modes is this loss of control: that we create AI systems that are pursuing incorrect objectives. And, because the AI system believes it knows what the objective is, it has no incentive to listen to us anymore, so to speak. It's just carrying out the strategy that it has computed as being the optimal solution. And it may be that in the process it needs to acquire more resources to increase the possibility of success, or prevent various failure modes by defending itself against interference. And so that collection of problems I think is something we can address. The other problems are, roughly speaking, misuse. So, even if we solve the control problem, we make perfectly safe [controllable AI](#) systems, well, why does Dr. Evil going to use those. He wants to just take over the world and he'll make unsafe AI systems that then get out of control. So, that's one problem, which is sort of a partly a policing problem, partly a sort of a cultural problem for the profession, of how we teach people what kinds of AI systems are safe.

[Lex] You talk about [autonomous weapon system](#), and how pretty much everybody agrees, there's too many ways that that can go horribly wrong if this great slaughter-bots movie that kind of illustrates that beautifully.

### Overusing AI. We've seen that movie before: Wall-E

[Stuart] That's another another topic I'm happy talking about. I just want to mention that what I see is the third major failure mode, which is overuse -not so much misuse- but overuse of AI. That we become overly dependent. So, I call this the [Wall-E](#) problem. If you've seen Wall-E the movie, all the humans are on the spaceship, and the machines look after everything for them, and they just watch TV and drink big gulps. And they're all sort of obese, and stupid, and they sort of totally lost any notion of human autonomy. In effect this would happen like the slow boiling frog; we would gradually turn over more and more of the management of our civilization to machines, as we are already doing. If this process continues, we sort of gradually switch from sort of being the Masters of Technology to just being the guests. So, we become guests on a cruise ship which is fine for a week but not not further the rest of eternity. And it's almost irreversible. Once you lose the incentive to, for example, learn to be an engineer, or a doctor, or a sanitation operative, or any other of the the infinitely many ways that we maintain and propagate our civilization, if you don't have the incentive to do any of that, you won't. And then it's really hard to recover.

[Lex] And of course AI is just one of the technologies that could that third failure mode result in that. There's probably other technologies in general detaches us from ...

[Stuart] It does a bit but the difference is that in terms of the knowledge to run our civilization, up to now we've had no alternative but to put it into people's heads ...

[Lex] Oh, it's not it were Google, I mean, so software in general ...

[Stuart] Computers in general. But the knowledge of how a sanitation system works, the AI has to understand



that. It's no good putting it into Google. We've always put knowledge on paper but paper doesn't run our civilization. It only runs when it goes from the paper into people's heads again. We've always propagated civilization through human minds and we've spent about a trillion person-years doing that.

[Lex] literature, right?

[Stuart] You can work it out. It is about just over a hundred billion people who've ever lived, and each of them has spent about ten years learning stuff, to keep their civilization going. That's a trillion person years we put into this effort.

[Lex] Beautiful way to describe all of civilization.

79:34

[Stuart] And now we're danger of throwing that away. So this is a problem that AI control it's not a technical problem. If we do our job right, the AI systems will say "you know, the human race doesn't, in the long run, want to be passengers in a cruise ship, the human race wants autonomy, this is part of human preferences. So, we, the AI systems, are not going to do this stuff for you, you've got to do it for yourself. I'm not going to carry you to the top of Everest in an autonomous helicopter, you have to climb it if you want to get the benefit". And so on. But I'm afraid that, because we are short-sighted, and lazy, we're gonna override the AI systems. There's an amazing short story that I recommend to everyone that I talk to about; it is called "[The Machine stops](#)", written in 1909 by [E. M. Forster](#), who wrote novels about the British Empire, and sort of things that became costume dramas on the BBC. He wrote this one science fiction story, which is an amazing vision of the future. It has basically iPads, it has video conferencing, it has [MOOCs](#), it has computer induced obesity. I mean, literally, the whole thing. it's what people spend their time doing is giving online courses, or listening to online courses, and talking about ideas, but they never get out there in the real world. That they don't really have a lot of face-to-face contact; everything is done online. So, all the things we're worrying about now were described in this story. And then the human race becomes more and more dependent on the machine, loses knowledge of how things really run, and then becomes vulnerable to collapse. It's a pretty unbelievably amazing story for someone writing in 1909 to imagine all this loss.

### **Machines can be beneficial to humans but somebody will find loopholes**

[Lex] There's very few people that represent artificial intelligence more than you Stuart Russell ...

[Stuart] So, it's all my fault right?

[Lex] You're often brought up as the person. Well, Stuart Russell, the AI person is worried about this; that's why you should be worried about it. Do you feel the burden of that? I don't know if you feel that at all, but when I talk to people, ... people outside of computer science, when they think about this. Stuart Russell is worried about [AI safety](#), you should be worried too? Do you feel the burden of that?

[Stuart] In a practical sense, yeah. Because I'd get a dozen sometimes 25 invitations a day to talk about it, to give interviews, to write press articles, and so on. So, in that very practical sense I'm seeing that people are concerned and really interested about this.

[Lex] Are you worried that you could be wrong as all good scientists are?

[Stuart] Of course. I worry about that all the time. That's always been the way that I I've worked. I have an argument in my head with myself. So, I have some idea and then I think "okay, how could that be wrong, or did someone else already have that idea". So I'll go and search as much literature as I can to see whether someone else already thought of that, or even refuted it. Right now, I'm reading a lot of philosophy because in the form of the debate over utilitarianism, and other kinds of moral formulas, shall we say, people have already thought through some of these issues. One of the things I'm not seeing in a lot of these debates is this specific idea about the importance of uncertainty in the objective, that this is the way we should think about machines that are beneficial to humans. So, this idea of provably beneficial machines based on explicit uncertainty in the objective. It seems to be ... my gut feeling is, this is the core of it. It's gonna have to be elaborated in a lot of different directions.

[Lex] ... to be beneficial

[Stuart] Yeah, it has to be, right? We can't afford you know hand-wavy beneficial. Because whenever we do hand wavy stuff there are loopholes, and the thing about super intelligent machines is they find the loopholes. Just like tax evaders. If you don't write your tax law properly then people will find the loopholes and end up paying no taxes. And so you should think of it this way. And in getting those definitions right. It is really a long process. So, you can define mathematical frameworks and within that framework you can prove mathematical theorems that yes, this theoretical entity will be proven beneficial to that theoretical entity but that framework may not match the real world in some crucial way.

85:23

### **Favorite Sci-Fi movie. Favorite robots**

[Lex] A long process, thinking through it, of iterating and so on. The last question. you have ten seconds to answer. What is your favorite [Sci-Fi](#) movie about AI?

[Stuart] I would say "[Interstellar](#)" has my favorite robots ...

[Lex] Oh, beats [HAL-9000](#)

[Stuart] Yeah. [TARS](#), one of the robots in [Interstellar](#), is the way a robot should behave. And, I would say [Ex-Machina](#) is in some ways the one like the one that makes you think in a nervous kind of way about a lot where we're going.

[Lex] Stuart thank you so much for talking today.

[Stuart] A pleasure














### **End of Interview**

### **References**

All references in the web: <https://www.one-tab.com/page/qRzU2-NxQHWThMTZSNTG7g>

-  [Stuart Russell: Long-Term Future of Artificial Intelligence | Artificial Intelligence \(AI\) Podcast - YouTube](#)
-  [Artificial Intelligence Podcast | AI Podcast | Lex Fridman](#)
-  [Stuart Russell: Long-Term Future of AI | MIT | Artificial Intelligence Podcast](#)
-  [Stuart Russell](#)
-  [Lex Fridman | MIT | Human-Centered AI & Autonomous Vehicles](#)
-  [\[Lex Fridman \(@lexfridman\) / Twitter\]\(https://twitter.com/lexfridman?ref\\_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor\)](#)
-  [MIT Deep Learning Lectures: Introduction, Tutorials, Videos, and Course Materials](#)
-  [MIT 6.S099: Artificial General Intelligence](#)
-  [Artificial Intelligence: A Modern Approach](#)
-  [AlphaGo - Wikipedia](#)
-  [AlphaGo Zero: Starting from scratch | DeepMind](#)

-  [AlphaZero - Wikipedia](#)
-  [Anca Dragan](#)
-  [Max Tegmark: Life 3.0 | MIT | Artificial Intelligence Podcast](#)
-  [\[1810.10525\] Toward an AI Physicist for Unsupervised Learning](#)
-  [Pamela McCorduck - Wikipedia](#)
-  [Arthur Samuel - Wikipedia](#)
-  [Lighthill report - Wikipedia](#)
-  [Sam Harris | Home of the Making Sense Podcast](#)
-  [Debiasing - Wikipedia](#)
-  [Rodney Brooks - Wikipedia](#)
-  [One Hundred Year Study on Artificial Intelligence \(AI100\) |](#)
-  [E. M. Forster - Wikipedia](#)
-  [Interstellar \(film\) - Wikipedia](#)
-  [Ex Machina \(film\) - Wikipedia](#)
-  [Lisp \(programming language\) - Wikipedia](#)
-  [Garry Kasparov - Wikipedia](#)
-  [Lighthill report - Wikipedia](#)
-  [Alan Turing | Biography, Facts, & Education | Britannica](#)
-  [King Midas](#)
-  [Norbert Wiener | American mathematician | Britannica](#)
-  [Markov decision process - Wikipedia](#)
-  [The History of Utilitarianism \(Stanford Encyclopedia of Philosophy\)](#)
-  [Food and Drug Administration - Wikipedia](#)
-  [california law impresonation - Google Search](#)
-  [Atomic bombings of Hiroshima and Nagasaki - Wikipedia](#)
-  [Chernobyl disaster - Wikipedia](#)

-  [H. G. Wells - Wikipedia](#)
-  [Frederick Soddy | British chemist | Britannica](#)
-  [Isotope - Wikipedia](#)
-  [Ernest Rutherford | Accomplishments, Atomic Theory, & Facts | Britannica](#)
-  [Leo Szilard | American physicist | Britannica](#)
-  [Neutron - Wikipedia](#)
-  [Manhattan Project - Wikipedia](#)
-  [Rodney Brooks Home](#)
-  [The gorilla problem | The Enlightened Economist](#)
-  [AI control problem - Wikipedia](#)
-  [Lethal autonomous weapon - Wikipedia](#)
-  [WALL-E - Wikipedia](#)
-  [Massive open online course - Wikipedia](#)
-  [Stanford AI Safety](#)
-  [Provable Synonyms, Provable Antonyms | Merriam-Webster Thesaurus](#)
-  [Science fiction - Wikipedia](#)
-  [Interstellar \(film\) - Wikipedia](#)
-  [HAL 9000 - Wikipedia](#)
-  [TARS | Interstellar Wiki | Fandom](#)
-  [The Machine Stops - Wikipedia](#)
-  [nuclear weapon | History, Facts, Types, & Effects | Britannica](#)
-  [Artificial Intelligence Network | Research groups | Imperial College London](#)
-  [Meta-Reasoning: Monitoring and Control of Thinking and Reasoning. - PubMed - NCBI](#)
-  [Meta-Reasoning: Monitoring and Control of Thinking and Reasoning - ScienceDirect](#)
-  [Metareasoning | The MIT Press](#)
-  [Backgammon | Play it online](#)

-  [Short Term Memory | Simply Psychology](#)
-  [Grandmaster \(chess\) - Wikipedia](#)
-  [Bobby Fischer - Chess Player, Author - Biography](#)
-  [20 Years after Deep Blue: How AI Has Advanced Since Conquering Chess - Scientific American](#)
-  [CS221](#)
-  [IBM100 - Deep Blue](#)
-  [Sinclair ZX-81 computer](#)
-  [Computing Square Roots with Newton's Method](#)
-  [Artificial Intelligence | Illinois Computer Science](#)
-  [AI University of Illinois - Google Search](#)
-  [Go \(game\) - Wikipedia](#)
-  [HOW TO PLAY GO](#)
-  [What Is Go? | American Go Association](#)

## Links

Article in LinkedIn: <https://www.linkedin.com/pulse/transcript-interview-stuart-russell-lex-fridman-alfonso-r-reyes>

Tweet: <https://twitter.com/OilGains/status/1218723904681922560?s=20>

Transcript in GitHub: [https://github.com/f0nzie/transcript\\_interview\\_stuart\\_russell\\_by\\_lex\\_fridman](https://github.com/f0nzie/transcript_interview_stuart_russell_by_lex_fridman)

YouTube video: <https://youtu.be/KsZI5oXBC0k>

Podcast: <https://lexfridman.com/stuart-russell/>

## Toolbox

Software utilities used for this transcript.

- Typora: <https://www.typora.io/>
- GitKraken: <https://www.gitkraken.com/>
- GitHub: <https://github.com/>
- TextFixer: <https://www.textfixer.com/tools/remove-line-breaks.php>
- 4K Video Downloader: <https://www.4kdownload.com/products/product-videodownloader>
- SubTitle Edit: <https://www.videohelp.com/software/Subtitle-Edit>