Tesla CEO Elon Musk, Professor of Computer Science at UC Berkeley Stuart Russell, Professor of Computer Science at Cornell University Bart Selman, Futurist Ray Kurzweil, Philosopher at New York University David Chalmers, Director of the Future of Humanity Institute at Oxford University Nick Bostrom, DeepMind CEO Demis Hassabis, Philosopher and neuroscientist Sam Harris, and Co-founder of Skype Jaan Tallinn discuss with Max Tegmark(moderator)

Max Tegmark: I'm going to ask a question, but you can only answer by saying either 'yes,' 'no,' or 'it's complicated.' Alright? So, let's start over here. Is some form of superintelligence possible Jaan? Yes, no, or it's complicated
Bart Selman, Ray Kurzweil, David Chalmers, Nick Bostrom, Demis Hassabis, Sam Harris: Yes.
Stuart Russell: Definitely.
Elon Musk: No.
All: (Laughter)

Max Tegmark: Well, this was disappointing, we didn't find any disagreement. Let's try harder. Just because it's possible doesn't mean that it's actually going to happen. So, before I asked if superintelligence was possible at all according to the laws of physics. Now I'm asking, will it actually happen?
Jaan Tallinn: A little bit complicated but yes.
Sam Harris Yes and if it doesn't then something terrible has happened to prevent it.
Demis Hassabis: Yes.
Nick Bostrom: Probably.
Sturat Russell, Bart Selman, Ray Kurzweil, David Chalmers: Yes.
Elon Musk: No.
All: (Laughter)

Max Tegmark: Sucks, still haven't found any interesting disagreements. We need to try harder still. So, you think it is going to happen but would you actually like it to happen at some point? Yes, no, or it's complicated
Jaan Tallinn: Complicated, leaning towards yes.
Sam Harris It's complicated.
Nick Bostrom, Demis Hassabis: Yes.
David Chalmers: It's really complicated.
Ray Kurzweil: Yes.
Bart Selman: It's complicated.
Stuart Russell: Very complicated.
Elon Musk: Well, heck, I don't know. It depends on which kind.

Max Tegmark: Alright, so it's getting a little bit more interesting. When I think, we had a really fascinating... When is it going to happen? Well, we had a really fascinating discussion already in this morning's panel about when we might get to human level AI. So, that would sort of put a lower bound. In the interest of time, I think we don't need to rehash the question of when going beyond it might start. But, let's ask a very related question to the one that just came up here. Mainly, the question of well if something starts to happen, if you get some sort of recursive

self-improvement or some other process whereby intelligence and machines start to take off very very rapidly, there is always a timescale associated with this. And there I hope we can finally find some real serious disagreements to argue about here. Some people have been envisioning this scenario where it goes PHOOM and things happen in days or hours or less. Whereas, others envision that it will happen but it might take thousands of years or decades. So, if you think of some sort of doubling time, some sort of rough timescale on which things get dramatically better, what time scale would you guess at, Jaan?
Jaan Tallinn: Start now or starting at human level?

Max Tegmark: No no, so once we get human level AI or whatever point beyond there or a little bit before there where things actually start taking off, what is the sort of time scale? Any explosion, as a nerdy physicist, has some sort of time scale, right, on which it happens. Are we talking about seconds, or years, or millennia?
Jaan Tallinn: I'm thinking of years, but it is important to act as if this timeline was shorter.
Sam Harris Yeah, I actually don't really trust my intuitions here. I have intuitions that we are thinking of years, but I also think human level AI is a mirage. It is suddenly going to be better than human, but whether that is going to be a full intelligence explosion quickly, I don't know.
Demis Hassabis: I think it partly depends on the architecture that ends up delivering human level AI. So, this kind of neuroscience inspired AI that we seem to be building at the moment that needs to be trained and have experience in order for it to gain knowledge that may be, you know, on the order of a few years so possible even a decade.
Nick Bostrom: Some numbers of years, but it could also be much less. But, I wouldn't be surprised if it was much more.
David Chalmers: Potentially days or shorter, especially if it's AI researchers designing AI researchers
Ray Kurzweil: Every time there is an advance in AI, we dismiss it as 'oh, well that's not really AI' chess, go, self-driving cars. An AI, as you know, is the field of things we haven't done yet. That will continue when we actually reach AGI. There will be lots of controversy. By the time the controversy settles down, we will realize that it's been around for a few years.
Bart Selman: Yeah, so I think we will go beyond human level capabilities in many different areas, but not in all at the same time. So, it will be an uneven process where some areas will be far advanced very soon, already to some extent and other might take much longer.
Stuart Russell: What Bart said.
Elon Musk: But, I think if it reaches a threshold where it's as smart as the smartest most inventive human then, I mean, it really could be only a matter of days before it's smarter than the sum of humanity.

Max Tegmark: So, here we saw quite an interesting range of answers. And this, I find is a very interesting question because for reasons that people here have published a lot of interesting papers about the time scale makes a huge difference. Right, if it's something that happening on the time scale of the industrial revolution, for example, that's a lot longer than the time scale on which society can adapt and take measures to steer development, borrowing your nice rocket metaphor, Jaan. Whereas, if things happen much quicker than society can respond, it's much harder to steer and you kind of have to hope that you've built in a good steering in

advance. So, for example in nuclear reactors, we nerdy physicists like to stick graphite sticks in a moderator to slow things down maybe prevent it from going critical. I'm curious if anyone of you feels that it would be nice if this growth of intelligence which you are generally excited about, with some caveats, if any of you would like to have it happen a bit slower so that it becomes easier for society to keep shaping it the way we want it. And, if so, and here's a tough question, is there anything we can do now or later on when it gets closer that might make this intelligence explosion or rapid growth of intelligence simply proceed slower so we can have more influence over it. Does anyone want to take a swing at this? It's not for the whole panel, but anyone who...

Jaan Tallinn: I'm reminded of the conversation we had with Rich Sutton in Puerto Rico. Like, we had a lot of disagreements, but definitely could agree about paths slow being better than faster.

Max Tegmark: Any thoughts on how one could make it a little bit slower?
Ray Kurzweil: I mean, the strategy I suggested in my talk was somewhat tongue and cheek. But, it was also serious. I think this conference is great and as technologists we should do everything we can to keep the technology safe and beneficial. Certainly, as we do each specific application, like self-driving cars, there's a whole host of ethical issues to address. But, I don't think we can solve the problem just technologically. Imagine that we've done our job perfectly and we've created the most safe, beneficial AI possible, but we've let the political system become totalitarian and evil, either an evil world government or even just a portion of the globe that is that way, it's not going to work out well. And so, part of the struggle is in the area of politics and social policy to have the world reflect the values want to achieve because we are talking about human AI. Human AI is by definition at human levels and therefore is human. And so, the issue of how we make humans ethical is the same issue as how we make AIs that are human level ethical.

Max Tegmark: So, what I'm hearing you say is that before we reach the point of getting close to human level AI, a very good way to prepare for that is for us humans in our human societies to try and get our act together as much as possible and have the world run with more reason than it, perhaps, is today. Is that fair?
Ray Kurzweil: That's exactly what I'm saying.
Max Tegmark: Nick? Also, I just want to clarify again that when I asked about what you would do to slow things down I'm not talking at all about slowing down AI research now. We're simply talking about if we get to the point where we are getting very near human level AI and think we might get some very fast development, how could one slow that part down?
Nick Bostrom: So, one method would be to make faster progress now, so you get to that point sooner when hardware is less developed, you get less hardware overhang. However, the current speed of AI progress is a fairly hard variable to change very much because there are very big forces pushing on it, so perhaps the higher elasticity option is what I suggested in the talk to ensure that whoever gets there first has enough of a lead that they are able to slow down for a few months, let us say, to go slow during the transition.

Stuart Russell: So, I think one thing you can do, I mean this is almost in the verification area, is to make systems that approvably will not recruit additional hardware or resigned their hardware, so that their resources remain fixed. And I'm quite happy to sit there for several years thinking hard about what the next step would be to take.
Ray Kurzweil: But it's trivial to copy software. Software is self-replicating and always has been and I don't see how you can possibly stop that.
David Chalmers: I mean, I think it would be great if it went slow, but it's hard to see how it does go slow given the huge first mover advantages and getting to superintelligence. The only scenario that I see where it might go slow is where there is only one potential first mover that can then stop and think. So, maybe that speaks to creating a society where, you know, AI is restrictive and unified, but without moldable movers.

Max Tegmark: Yeah, Demis, so your colleague Sean Legg mentioned that the one thing that could help a lot here is if there's things like this industry partnership and a sense of trust and openness between the leaders, so that if there is a point where one wants to...
Demis Hassabis: Yeah, I do worry about, you know, that sort of scenario where, you know, I think, I've got quite high belief in human ingenuity to solve these problems given enough time. The control problem and other issues. They're very difficult, but I think we can solve them. The problem is will there, you know, the coordination problem of making sure there is enough time to slow down at the end and, you know, let Stuart think about this for 5 years. But, what about he may do that but what about all the other teams that are reading the papers and not going to do that while you're thinking. Yeah, this is what I worry about a lot. It seems like that coordination problem is quite difficult. But, I think as the first step, may be coordinating things like the Partnership on AI, you know, the most capable teams together to agree, at least agree on a set of protocols or safety procedures, or things, you know, agree that, maybe, you know, you should verify these systems and that is going to take a few years and you should think about that. I think that would be a good thing. I just want to caveat one thing about slowing versus fast progresses, you know, it could be that, imagine there was a moratorium on AI research for 50 years, but hardware continued to accelerate as it does now. We could, you know, this is sort of what Nick's point was is that there could be a massive hardware overhang or something where an AI actually many, many, many different approaches to AI including seed AI, self-improving AI, all these things could be possible. And, you know, maybe one person in their garage could do it. And I think that would be a lot more difficult to coordinate that kind of situation, whereas, so, I think there is some argument to be made where you want to make fast progress when we are at the very hard point of the 's' curve. Where actually, you know, you need quite a large team, you have to be quite visible, you know who the other people are, and, you know, in a sense society can keep tabs on who the major players are and what they're up to. Whereas, opposed to a scenario where in say 50 or 100 years time when, you know, someone, a kid in their garage could create a seed AI or something like that.

Max Tegmark: Yeah Bart one last comment on this topic.
Bart Selman: Yeah I think that this process will be a very irregular process and sometime we will be far advanced and other times we will be going quite slow. Yeah, I'm sort of hoping that when society sees something like fake video creation where you create a video where you have

somebody say made up things and that society will actually realize that there are these new capabilities for the machines and we should start taking the problem as a society more seriously before we have full and general AI.
Ray Kurzweil: We'll use AI to detect that.

Max Tegmark: So, you mentioned the word 'worry' there, and you Nick went a bit farther, you had the word 'doom' written on your slides three times. No wonder there was one star on Amazon on that rating and that it was even in red color. I think it's just as important to talk about existential hope and the fantastic upside as downside and I want to do a lot of that here. So, let's just get some of those worries out of the way now and then return to the positive things. I just want to go through quickly and give each one of you a chance to just pick one thing that you feel is a challenge that we should overcome and then say something about what you feel is the best thing we can do, right now, to try to mitigate it. Do you want to start Jaan?
Jaan Tallinn: To mitigate what?

Max Tegmark: Mention one thing that you're worried could go wrong and tell us about something constructive that we can do now that will reduce that risk.
Jaan Tallinn: I do think that AI arms races, I see like a lot of, like, good. I'm really heartened to see kind of great contacts between OpenAI and DeepMind, but I think this is just like a sort of toy model of what the world at large might come up with in terms of arms races. And for myself I have been spending increasing amount of time in Asia recently just to kind of try to kind of pull in more people elsewhere, what has been so far, just been, kind of like, an Anglo American discussion mostly. So, like this is, I think, this is one thing that should be done and I'm going to do it.
Sam Harris Well, as someone who is outside this field, I think the challenge I'm really in touch with is how hard it is to take the safety concerns emotionally seriously. And how hard it is for people in the field to do that as well. I can't tell you have many people outside this room who purport to be experts think the control problem is a total non-issue. I mean, it's just flabbergasting to meet these people and just therefore not worth thinking about. And one of the reasons I think is that in one case there is this illusion that the time horizon matters. If you feel that this is 50 or a 100 years away that is totally consoling, but there is an implicit assumption there, the assumption is that you know how long it will take to build this safely. And that 50 or a 100 years is enough time. The other issue is, I think, most people feel like intelligence is an intrinsic good and of course we want more of it and it's very easy to be in touch with that assumption because right now there is a cure for cancer, which we have not discovered. Right, how galling is that? But for more intelligence, but for knowing which experiments to run, or how to integrate the data we already have in hand, we would have a cure for cancer that was actionable now unless there was some physical law of the universe that prevented us from curing cancer, which seems unlikely. So, the pain of not having enough intelligence is really excruciating when you focus on it, but, and I think to your previous question of doing this quickly becomes an intrinsic good provided we have solved the alignment problems and the political problems and the chaos that would follow if we were just, if we did it quickly without solving those problems. So, I think, it's the thing that is alarming is how ethereal these concerns are even to those who have no rational argument against them.

Max Tegmark: So, Sam it sounds to me like you're agreeing very strongly with what Shane Legg that there is, in some circles, still this strong taboo that, you know, don't even consider the possibility that we might get AGI because it's just absolutely ridiculous. And he was arguing that the sooner we can get rid of this taboo the sooner people can get to work and find all these really helpful solutions and answers that we need. So, suppose for a moment that I came up to you and said to you 'this idea of super human intelligence just sounds absolutely ridiculous, sounds completely nuts. And by the way I've never seen your ted talk.' And we're in an elevator and you have only 30 seconds to persuade me to take this more seriously, what would you say? A lot of people who are here will have this exact conversation with colleagues and others in the future.

Sam Harris Well there are very few assumptions you need to make to take this seriously, intellectually. Again, the emotional part is a separate piece. But, if you assume that intelligence is just, on some level, the product of information processing in a physical system and there are very few people who dispute that who are scientifically literate at this point and you assume that we will continue to improve our information processing systems, unless something terrible happens to us to prevent that, and that seems like a very safe assumption, then it is just a matter of time before we instantiate something that is human level and beyond in our computers. And, again, the time horizon is only consoling on the assumption that we know we have enough time to solve the alignment problems and the political problems. The other thing that is humbling here that Ray brought up at one point is that even if we were handed a perfectly benign, well behaved AI just from god, you know, we are given a perfect oracle we are given a perfect inventor of good technology, given our current political and economic atmosphere that would produce total chaos. We just have not... we don't have the ethical or political will to share the wealth, we don't have the political integration to deal with this thing being given to Silicon Valley and not being given at the same moment to China or Iran. So, there is just, it's alarming that the best case scenario currently, basically just ripping out 80% of Nick's book because we've solved all those problems, is still a terrifying one. And so, clearly, that's a near term thing that we have to solve.

Max Tegmark: Thank you Sam. So, Demis do you want to tell us about one thing that you feel is a challenge and say something about what we should focus on now to tackle it.

Demis Hassabis: Yeah, I mean I think it's, you know I agree with both the statements already said that, so I think the coordination problem is one thing where you know we want to avoid this sort of harmful race to the finish where corner cutting starts happening and things like safety are easy things to, you know, will get cut because obviously they don't necessarily contribute to AI capability, in fact they may hold it back a bit by making a safe AI. So, I think that's going to be a big issue on a global scale and that seems like it's going to be a hard problem when we are talking about national governments and things. And I think also, you know, we haven't thought enough about the whole governance scenario
of how do we want those AIs to be out in the world? How many of them? Who will set their goals? All these kinds of things, I think, need a lot more thought. You know, once we've already solved the technical problems.

Max Tegmark: I think it's wonderful that you're not just saying these things, but actually doing these things since you played a leading role in setting up the Partnership on AI here which goes exactly in the direction of what you're advocating here. So, do you want to pass it off to Nick? I'm sure there is nothing at all you're worried about, right? So, tell us about one concrete useful thing you would to see us focus on.

Nick Bostrom: So, I agree with that, I mean, so fun technical work, bring in top technical talent to work on these technical issues, build these collaborations, build a community, build trust, work some more on figuring out attractive solutions to the governance solutions that could work, but don't rush to implement the first idea you have, but first trial them out a little bit more.

David Chalmers: I think a lot about consciousness, so I was really struck by the 'sentience caution' on the list of principles here that said 'avoid overly, avoid strong assumptions about the distribution of consciousness in AIs' which I take it entails avoid assuming that any human level or super human level AGI is going to be conscious. For me, that raising the possibility of a massive failure mode in the future, the possibility that we create human or super human level AGI and we've got a whole world populated by super human level AGIs, none of whom is conscious. And that would be a world, could potentially be a world of great intelligence, no consciousness no subjective experience at all. Now, I think many many people, with a wide variety of views, take the view that basically subjective experience or consciousness is required in order to have any meaning or value in your life at all. So therefore, a world without consciousness could not possibly a positive outcome. maybe it wouldn't be a terribly negative outcome, it would just be a 0 outcome, and among the worst possible outcomes. So, I worry about avoiding that outcome. Now, as a matter of fact, I'm fairly optimistic about the possibilities that AIs of various kinds will be conscious. But, in so far as this community is making this assumption, I think it's important to actually think about the question of 'in creating AGIs, are we actually creating conscious beings?' I mean, one thing we ought to at least consider doing there is making, given that we don't understand consciousness, we don't have a complete theory of consciousness, maybe we can be most confident about consciousness when it's similar to the case that we know about the best, namely human, human consciousness··· So, therefore maybe there is an imperative to create human-like AGI in order that we can be maximally confident that there is going to be consciousness.

Max Tegmark: So, what I hear you say is that when you have a nightmare about the zombie apocalypse you're not thinking of some terminator movie, but you're thinking about this problem. We create... we upload ourselves and do all these wonderful things, but there's no one home. Is that fair to say?

David Chalmers: I mean this is a different kind of existential risk. One kind of existential risk is there's no humans, there's AIs, but some people might say well that's OK they are our successors. A much worse existential risk is there are no conscious beings in our future.

Max Tegmark: So, I'll make a confession, so Shane Legg mentioned that there has been this strong taboo about talking about the possibility of intelligence getting very advanced. It's clearly also been a strong taboo for a long time to mention the C-word. In fact, before the conference when we got all these responses on the first round of the principles, guess which

one was ranked last? It got huge amounts of minus 1 ratings, that was the one with consciousness, so we changed it to..
David Chalmers: it was terribly stated

Max Tegmark: ..sentience and stated it better and then it got stated still better at lunch and it's still rated last. Even though I personally share your interests in this a lot.
David Chalmers: 88% of people agreed to the sentient caution.
Max Tegmark: But, not 90%, so that one also fell off the list here. So, maybe that is another taboo you can personally help us shatter so that people think about that question more. Ray, anything you're concerned about?
Ray Kurzweil: This isn't what I was going to say, but just to respond... a converse concern is we create AGIs, everybody assumes that of course it's just a machine and therefore it's not conscious, but actually it is suffering but we don't look out for it's conscious subjective experience because we are making the wrong assumption. But, what I did want to say was, there are three overlapping revolutions that people talk about, GNR, genetics, bio-tech, nano-technology, and robotics, which is AI. And there are proposals, there was the Asilomar conferences done here decades ago for bio-tech that have worked fairly well. There are similar proposals for nano-technology. There is a difference with AI in that there really isn't a full proof technical solution to this. You can have technical controls on, say, nano-technology. One of the guidelines is it shouldn't be self-replicating. That's not really realistic because it can't scale to meaningful quantities without being self-replicating, but you can imagine technical protections. If you have an AI that is more intelligent than you and it's out for your destruction and it's out for the world's destruction and there is no other AI that is superior to it, that's a bad situation. So, that's the specter. And partly this is amplified by our observation of what we as humans, the most intelligent species on the planet, have done to another species. If we look at how we treat animals, people, you know, are very friendly, like their dogs and pets, but if you look at factory farming we're not very benign to species that are less intelligent than us. That engenders a lot of the concern we see that if we there's a new type of entity that's more intelligent than us it's going to treat us like we've treated another species. So, that's the concern. I do think that what we are doing at this conference is appropriate. I wanted to mention that I think we should publish these guidelines the way the Asilomar guidelines in bio-tech were published decades ago. And then people can and people can, you can have an opt-in, opt-out, but I think we should actually say we had this conference and the AI leadership/community has come up with these guidelines and people can respond to them and debate them and then maybe at the next conference we'll revise them. The Asilomar bio-tech guidelines have been revised many times. But, I would advocate that we actually take a stand and put forth these guidelines and then let the whole community at large debate them. And have them be, have them guide our research. It's actually worked quite well in bio-tech.
Max Tegmark: Bart?
Bart Selman: OK, yeah so let me give a little different perspective. So, one concern I have at the high level is these machines become really smart or even in certain areas, can humans still understand, what they, decisions that they suggested, that they make. And I work in the field of automated reasoning where we have significant advance last two decades going from perhaps a few hundred variables to perhaps millions of variables being solved quite routinely. And there

was a sense in the community, well we are getting answers from these reasoning engines, mostly hardware/software verification problems, but we cannot, humans can no longer understand these answers. In the last few years, people have actually discovered that you can use the machine to generate explanations for their answers that are, again, human understandable. So, I see sort of a glimmer of hope that maybe even if we have much less intelligence we may be able to understand solutions that machines find for us and we could not find these solutions, but they may be able to provide explanations that are accessible to us. So that's a little positive note.

Max Tegmark: Thank you. Stuart?

Stuart Russell: So there are two things that keep me awake at night, other than email. So, one is the problem of misuse and bad actors. To take an analogy, it's as if we were building nuclear weapons and then delivering them by email to everybody on the planet, saying, here's a toy, do what you want. How do we counter that? I have to say, I don't really have a good solution. I think one of the things we have to do is to make designs for safe AI very clear and simple, and sort of make it unthinkable to do anything other than that, right? Just like it's unthinkable to have a program with an infinite loop that produces a spinning pizza of death on your, oh sorry. Or it's unthinkable to have a buffer overflow that allows your software to be hacked into. The other thing that keeps me awake is actually the possibility that success would lead to AI as a helicopter parent for the human race that would sort of ossify and gradually enfeeble us, so then there would be no point at which it was obvious to us that this was happening. And I think the mitigation, which you asked for, to look on the bright side, is that in some sense the meta-value of human evolvability, the freedom to change the future, is something that the AI needs to adopt, and in some sense that would result eventually with the AI receding into the background, and saying, OK, now I've got you through your adolescence, now it's time for the human race to grow up, now that we have the capabilities to eliminate scarcity, to eliminate needless conflict and coordination failures and all of those things that we suffer from right now. So I can imagine a distant future where, in fact, AI is perhaps even less visible than it is today.

Max Tegmark: Great, finally you, Elon, have as far as I know never ever expressed any concerns about AI, right I'm just wondering if there are any concerns, in particular any concerns where you see there's a very clear thing we should be doing now that are going to help.

Elon Musk: I'm trying to think of what is an actual good future, what does that actually look like, or least bad, or however you want to characterize it. Because to a point that was made earlier by Sam and maybe made by others, we're headed towards either superintelligence or civilization ending. Those are the two things that will happen - intelligence will keep advancing, the only thing that would stop it from advancing is something that puts civilization into stasis or destroys civilization. So, we have to figure out, what is a world that we would like to be in where there is this digital superintelligence? I think, another point that is really important to appreciate is that we are, all of us, already are cyborgs. So you have a machine extension of yourself in the form of your phone and your computer and all your applications. You are already superhuman. By far you have more power, more capability, than the President of the United States had 30 years ago. If you have an Internet link you have an article of wisdom, you can communicate to millions of people, you can communicate to the rest of Earth instantly. I mean,

these are magical powers that didn't exist, not that long ago. So everyone is already superhuman, and a cyborg. The limitation is one of bandwidth. So we're bandwidth-constrained, particularly on output. Our input is much better but our output is extremely slow. If you want to be generous you could say maybe it's a few hundred bits per second, or a kilobit or something like that output. The way we output is like we have our little meat sticks that we move very slowly and push buttons, or tap a little screen. And that's extremely slow. Compare that to a computer which can communicate at the terabyte level. These are very big orders of magnitude differences. Our input is much better because of vision, but even that could be enhanced significantly. So I think the two things that are needed for a future that we would look at and conclude is good, most likely, is, we have to solve that bandwidth constraint with a direct neural interface. I think a high bandwidth interface to the cortex, so that we can have a digital tertiary layer that's more fully symbiotic with the rest of us. We've got the cortex and the limbic system, which seem to work together pretty well – they've got good bandwidth, whereas the bandwidth to additional tertiary layer is weak. So I think if we can solve that bandwidth issue and then AI can be widely available. The analogy to a nuclear bomb is not exactly correct – it's not as though it's going to explode and create a mushroom cloud, it's more like if there were just a few people that had it they would be able to be essentially dictators of Earth, or whoever acquired it and if it was limited to a small number of people and it was ultra-smart, they would have dominion over Earth. So I think it's extremely important that it be widespread and that we solve the bandwidth issue. And if we do those things, then it will be tied to our consciousness, tied to our will, tied to the sum of individual human will, and everyone would have it so it would be sort of still a relatively even playing field, in fact, it would be probably more egalitarian than today.

Max Tegmark: Great, thank you so much, that's in fact the perfect segue into the last question I want to ask you before we open it up to everybody. Something I have really missed in the discussion about really advanced intelligence, beyond human, is more thought about the upside. We have so much talk about existential risk, and not just in the academic context, but just flip on your TV, check out Netflix, what do you see there in these scientific visions of the future? It's almost always dystopias, right? For some reason fear gives more clicks than the positive visions, but if I have a student coming into my office at MIT asking for career advice, the first thing I'm going to ask her is, where will you want to be in 20 years? And if she just says, well maybe I'll get cancer, maybe I'll get run over by a bus, that's a terrible way to think about career planning, right? I want her to be on fire and say my vision is I want to do this – and here are the things that could go wrong, and then you can plan out how to avoid those problems and get it out – I would love to see more discussion about the upsides, futures we're really excited about, so we cannot just try to avoid problems for the sake of avoiding problems, but to get to something that we're all really on fire about. So to start off I'll just tell you something that makes me really excited about advanced artificial intelligence. Everything I love about civilization is a product of intelligence. If we for some reason were to say, well, you know, I'm scared about this technology thing, let's just press pause on it forever, there's no interesting question about if we're going to have human extinction, the question is just 'when' Is it going to be a supervolcano, is it going to be the next dinosaur killing class asteroid the last one happened 60 million years ago, so how long is it going to be? Pretty horrible future to just sit

and wonder when we're going to get taken out here without the technology when we know that we totally have the brainpower to solve all of these problems if we proceed forward and develop technology. So that was just one thing that makes me very excited about moving forward rather than pressing 'Pause.' I want to just ask the same question to all of you guys in turn. So tell us, just pretty briefly, about something that you are really excited about. Some future vision you imagine with very advanced artificial intelligence that you're really excited about, that you would like to see. Jaan

Jaan Tallinn: So I want to be careful when I imagine concrete fruits of AGI. On a meta-level I think as a first approximation, I think we should just maximize the amount of fun and minimize the amount of suffering. I think Eliezer has written a sequence called 'Fun Theory', where he points out that people have been horrible imagining, are very unimaginative imagining paradises of various sorts, just like really boring places, actually, when you think about them. I think Eliezer has this sketch where he says, 'It was hard to spend like one weekend with my relatives. Imagine spending eternity with your dead relatives.' So I think we should be concerned about side effects and try to capture dynamics of improvement, and basically go from there – make sure that we're going to adjust the trajectory as we get smarter and more grown together.

Max Tegmark: Great, thank you, Jaan. Sam, what do you get excited about?

Sam Harris Well, strangely, what excites me really just abuts the parts that scare me the most. I think what is nice about this conversation, in particular about the alignment problem, is that it's forcing us to realize that there are better and worse answers to questions of human value. And as someone said, perhaps at this last meeting in Puerto Rico, we really have to do philosophy on a deadline, and we have to admit to ourselves that there are better and worse answers and we have to converge on the better ones. And what would excite me about actually the birth of superintelligent AI – one of the things, apart from solving obvious problems like curing disease and energy issues and all the rest, perhaps differs a little bit with what Stuart said. I'm not so worried about idiocracy or all of us just losing our way as apes and living unproductive lives in dialogue with these oracles. I think actually, I would want a truly value-aligned superintelligence to incrementally show us, not merely conserve what we want, but show us what we should want to keep improving our values so that we can navigate in the space of all possible experiences and converge on better and better ones.

Max Tegmark: Thank you, Sam, and what about you, Demis?

Demis Hassabis: So obviously this is why I spend my whole career working on this, is that, I think if we do this right, it's going to be the greatest thing ever to happen to humanity, and in some ways I think unlock our full potential. I mean, I've talked to a lot about, in all my talks about using it as a tool to help us make science and medical breakthroughs faster. And so I think that's an obvious one. But taking that longer-term, one reason I got so into AI is that, like probably many of you in this room, I'm interested in the biggest questions of why we're here, understanding our minds, what is consciousness, what's the nature of the universe, what's our purpose and if we're going to try and really grapple with any of those questions I think we're going to need something like AI, perhaps with ourselves enhanced as well. And I think in that future world we'll have a chance to actually find out about some of these really deep questions

in the same way we're finding out with AlphaGo just about Go, but what if we could do that with all of science and physics and the biggest questions in the universe. And I think that's going to be the most exhilarating journey of all, to find that out. To just carry out on a few other things that people commented on is in terms of us as the most intelligent beings on the planet right now, and treating animals badly and these sorts of things, I think if you think about it though let's take tigers or something in India. They have huge ranges and those people are very poor and they're resource poor, but if they had abundant resources I don't think they're intentionally trying to kill off these tigers in some cases they are but often it's just because they need the land for their cattle, and the tiger needs whatever number of kilometers squared to live, one tiger. And it's just difficult with the number of people that are there. So I think if we solve the kind of abundance and scarcity problem, then I think that opens up a lot of conflicts both between humans as well as to do with resource scarcity, at the heart of it. So I see, if we can solve a lot of these problems I can see a much better future.

Max Tegmark: So Nick, you pointed out, the upside part of your book was a little shorter, so now you have a chance to add something positive. What are you excited about?
Nick Bostrom: There are really two sides to that. So one is getting rid of a lot of the negatives, like the compassionate use to cure diseases and all other kinds of horrible miseries that exist on the planet today. So that is a large chunk of the potential. But then beyond that, if one really wants to see realistically what the positive things are that could be developed, I think one has to think outside the constraints of our current human biological nature. That it's unrealistic to imagine a trajectory stretching hundreds of thousands of years into the future, we have superintelligence, we have material abundance, and yet we are still these bipedal organisms with three pounds of gray tissue matter, with a fixed set of emotional sensitivities and the hedonic set point that is kind of OKish for most people but if you get – if something really good happens it lasts for a time and then you're back to the baseline. I think all of these basic parameters that sort of define the human game today, I think become up for grabs in this future. And it opens up this much vaster space of post-human modes of beings, some of which I think could be wonderful, literally beyond our ability to imagine, in terms of the mental states, the types of activities, the understanding, the ways of relating. So I don't think we need a detailed blueprint for utopia now, what we need is to get ourselves in a position later on where we can have the ability to use this to realize the values that come into view once we've taken steps forward.

Max Tegmark: Thank you, Nick. What about you, David?
David Chalmers: I'm excited about the possibilities for AI making us humans smarter. I mean some of it is selfish I turned 50 last year, my brain is gradually becoming slower and older and dumber, but I'm not sure that I am, and that's partly because of all of the augmented intelligence technology we're using. Smartphones, and the Internet, and so on, they're giving me all kinds of capacities, extended capacities that I didn't have before. And I'm really looking forward to AI helping with that. In ten years or so once everyone is wearing augmented reality glasses with deep learning built into it, then I'm really going to need that around 60. And if you guys really get on the case and by the time I'm 70 or so we've got real genuine AI or AI modules out there which can somehow come to be integrated with my brain processes or maybe

eventually we get to upload our entire brains onto AI, then there's a way potentially to get smarter, more intelligent forever. And this is not just selfish, although I can't say that doesn't motivate me, but Demis talked about the AI scientists; I also like to think about the AI philosopher. The problems of philosophy are really hard and many people have speculated that we humans are just too dumb to solve some of them. But once we've actually got AIs on the scene, maybe AI-enhanced humans, then maybe we're going to be able to cross those thresholds where the AI-enhanced humans or maybe just the AGIs end up solving some of those hard problems of philosophy for once and for all.

Max Tegmark: Great, Ray, you have been a true pioneer in articulating positive visions of the future in your writing. So if you picked the one that you're most excited about now, what would that be?
Ray Kurzweil: So imagine going back 10,000 years and asking the quintessential caveman and woman, Gee, what is a beneficial future? What would you like to see? And they would say, well I would like this fire to stop going out and I would like a bigger boulder to prevent the animals from getting in the cave. Anything else? Well no I think that would be pretty perfect. Well don't you want a better website and apps and search engines? Imagine going back 2 million years ago and talking to primates imagine if you could do that, and saying, isn't it great that frontal cortex is coming and we're going to have additional neocortex and a hierarchy and they say, well what's the point of that? And you say, well you'll have music and humor, and their answer would be, what's music? What's humor? So they couldn't imagine concepts that they couldn't imagine, and by analogy I think we will have new phenomena that are as profound as music and humor, you could call it more profound music and we'll be funnier, but I think it'll be as profound as these great leaps that evolution has brought us, because we will become profoundly smarter and if music and humor are up here and we go to even higher levels of the neocortex, we're going to have more profound ways of expressing ourselves and once we have that we would not want to go back.

Max Tegmark: What about you, Bart?
Bart Selman: Well, I pretty much agree that we can't really predict much in advance, what we would like to have. For myself personally I see the developments in mathematics and science and discovery, and computers are just the hybrids of human computers there is quite incredible and makes the field makes what we do much more exciting. So I think that will be in the near future the first thing.

Max Tegmark: Great, and what about you, Stuart?
Stuart Russell: Well, so like Jeffrey Sachs I think that for many of us, and probably like the cavemen - that for many of us life is pretty amazing, and for many more of us it isn't. And I think the best thing that AI can do, the big upside, is actually to fix the latter problem. I mean I love Nick's feeling that there are higher states of being that are so far above our current 'pretty good', that that balances out all the 'pretty bad' that a lot of people are suffering. But I really think the emphasis should be on the 'pretty bad' and fixing it, and eliminating so Demis was reading my notes apparently, from across the room but eliminating the scarcity basically eliminates the need for people to act in a zero-sum fashion where they can only get by, by

making it less possible for someone else to get by, and I think that's the source of a lot of the nastiness that Jeffrey mentioned earlier. So I think that would be my main upside, and not having to read so much email, that would be the second one.

Max Tegmark: And for you, Elon, you've never articulated any visionary ideas about the future as far as I know, either. What about now?
Elon Musk: I think I just I have thought about this a lot, and I think it just really comes down to two things, and it's solving the machine-brain bandwidth constraint and democratization of AI. I think if we have those two things, the future will be good. There was a great quote by Lord Acton which is that 'freedom consists of the distribution of power and despotism in its concentration.' And I think as long as we have - as long as AI powers, like anyone can get it if they want it, and we've got something faster than meat sticks to communicate with, then I think the future will be good.
Max Tegmark: Fantastic, so let's get I know your caffeine levels are dropping dangerously low, and we also have another panel after this, which is going to be really exciting to listen to, so let's do a just a few quick questions. Make sure that they are actually questions, and say your name and also say, pick one person on the panel and address it just to them, OK? Yoshua?

Questioner 1: Yoshua Bengio, Montreal. And it's for Jaan I found your presentation very inspiring, and one question I have is related to the question of eliciting preferences and values from people. Do you think this line of investigation could lead to better democracy, better society, more direct democracy, and you know, what do you think about this direction to deal with the issue of misuse and things like that?
Jaan Tallinn: Yes, absolutely. There could be one code name for this, even, could be like 'Democracy 2.0' or 'U.N. 2.0' or something like that. So, and as I mentioned in my presentation, just a lot of people today basically want to make the world better, but it's kind of hard to distinguish them from people who say they want to make the world better. So if there was actually kind of like a very easy measuring, like a metric that basically would work as a Schelling point, focal point, then I think that would be super helpful. And yeah, like democracy was invented like hundreds of years ago so, and clearly we have advanced as a civilization and we have better knowledge about how to aggregate preferences.
Max Tegmark: And Nicholas Berggruen, over there.

Questioner 2: Thank you, Max. Nicholas Berggruen, so I have a very almost naive question. This is a very well-meaning group in terms of, let's say, intentions, but who sort of, looking at who else is doing, potentially, AGI, it could be well beyond this group, it could be in China, it could be any place. And what happens because we've talked about how powerful AGI is, and if Elon is correct, if it is distributed fairly, fine. But if it isn't, is there a way to monitor today or in a year or in 10 years, because once it's out it'll be fast. Who is monitoring it, who has a tab on it? Because this is self-selected, but beyond...

Max Tegmark: Elon or Demis does either one of you want to take a swing at this?
Demis Hassabis: Well I think this sort of relates to my point I said earlier about trying to build AI at the hard part of the S curve, so, which I think is where we sort of are at the moment, as far as

we can tell, because, you know, it's not easy to make this kind of progress, so you need quite a lot of people who are quite smart and that community is pretty small, still, even though it's getting rapidly bigger at places like NIPS. And so most people know each other, so this is pretty representative of everyone in the West, at least, obviously it's harder to know what's happening in China or in Russia, maybe. But, you know, I think that you need quite a large footprint of resources, people and very smart people and lots of computers and so on. So I think that narrows down the scope of the number of groups who can do that, and it also means that they're more visible. So, you know, I think certainly in the West I think most people around here, someone in this room will have contact with somebody who's in those groups who are capable of making meaningful progress towards AGI. It's harder to know in the East and further apart, but we should try and make links to those Chinese National Academy of Sciences, and so on, to find out more. But you know that may change in the future, I think that's the current state of it.

Max Tegmark: Great, it's the bad news is it's getting later in the day and we only have time for one more question. The good news is there's a coffee break right after this so you can ask all of your questions if you swarm the panel. And the last question goes to you, Erik. Do you want to stand up?

Questioner 3: Erik Brynjolfsson, MIT. I'm going to pick up on the thing that Elon said at the end about democratizing the outcome and tie it back to the panel yesterday where Reid Hoffman talked about people caring a lot about not just absolute income but relative income, and I wanted to get the panelists' reactions to the thoughts about whether or not AI had tendencies towards winner-take-all effects, that there's a tendency for concentration, that whoever's ahead can pull further ahead, or whether there's potential for more widespread democratic access to it, and what kinds of mechanisms we can put in place if we want to have the widely shared prosperity that Elon suggested?

Max Tegmark: Elon, do you want to take that?

Elon Musk: Yeah, well, I mean I have to say that when something is a danger to the public, then there needs to be some I hate to say government agency, like regulators – I'm not the biggest fan of regulators, 'cause they're a bit of a buzzkill. But the fact is we've got regulators in the aircraft industry, car industry, I deal with them all the time, with drugs, food – and anything that's sort of a public risk. And I think this has to fall into the category of a public risk. So I think that the right thing to do, and I think it will happen, the question is whether the government reaction speed matches the advancement speed of AI. Governments react slowly or governments move slowly and they tend to be reactive, as opposed to proactive. But you can look at these other industries and say, does anybody really want the FAA to go away? and it's like people could just be a free for all for aircraft – like, probably not. You know, there's a reason it's there or just people could just do any kind of drugs and maybe they work, maybe they don't. You know, we have that in supplements, kind of ridiculous. But I think on balance FDA is good, so I think we probably need some kind of regulatory authority and I think it's, like a rebuttal to that is, well people will just move to Costa Rica or something. That's not true. OK, we don't see Boeing moving to Costa Rica or to Venezuela or wherever it's like free and loose. To Demis'

point, the AI is overwhelmingly likely to be developed where there is a concentration of AI research talent. And that happens to be in a few places in the world. It's Silicon Valley, London, at Boston, if you sort of figure out a few other places, but it's really just a few places that really regulators could reasonably access. And I want to be clear, it's not because I love regulators, OK? They're a pain in the neck but they're necessary to preserve the public at times.

Max Tegmark: Alright, on that note, let's thank the panel for a fascinating discussion.