Max Tegmark starts at 36:14 into the above film

The Helen Caldicott Foundation Presents

Max Tegmark

Artificial Intelligence and the Risk of Accidental Nuclear War: A Cosmic Perspective

Symposium: The Dynamics of Possible Nuclear Extinction

The New York Academy of Medicine, 28 February - 1 March 2015


Introduction by Dr. Helen Caldicott

The next speaker is Max Tegmark who, as I mentioned earlier, was mentioned in the *Atlantic Monthly* piece along with Steven Hawking. Max Tegmark has been concerned about nuclear war risk since his teens and started publishing articles about it at the age of 20. He is President of the Future Of Life Institute which aims to prevent human extinction as discussed in his popular book, *Our Mathmatical Universe*. His scientific interests also include precision cosmology and the ultimate nature of reality. He is an MIT physics professor with more than 200 technical papers and is featured in dozens of science documentaries. His work with the Sloan Digital Sky Survey on galaxy clustering, shared the first prize in *Science Magazine*'s breakthrough of the year 2003. His title is "Artificial Intelligence and the Risk of Accidental Nuclear War." Max Tegmark:

Thank you so much for inviting me. It's a great honor to be here. Can I borrow the clicker? It seems like you invited a lot of people from MIT this morning.

So, as you heard, I am physicist, a cosmologist; I spend much of my time studying our universe, trying to figure out what's out there, how old it is, how big it is, how it got here. And I want to share with you my cosmic perspective.

You heard from Theodore Postol here that what we've done with nuclear weapons is probably the dumbest thing we've ever done here on earth. I'm going to argue that it might also be the dumbest thing ever done in our universe.
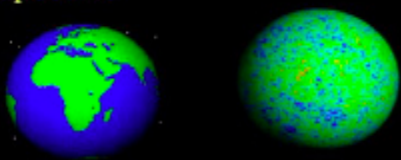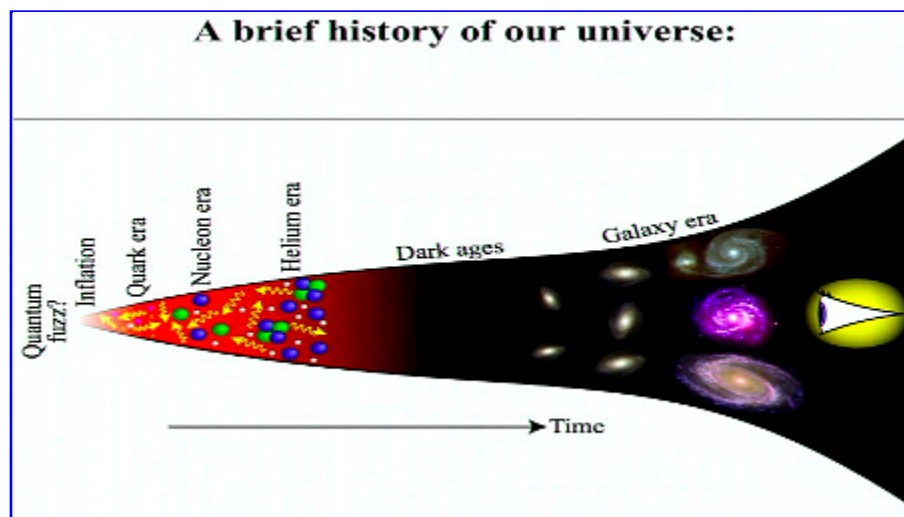

When we look at it from a cosmic perspective: here we are 13.8 billion years after a big bang, something quite remarkable has happened. Life has evolved. Our universe has become aware of itself. This life has done a lot of really fantastic things that are truly inspiring. We have created great music, theater, literature, and by using our curious minds we've been able to figure out more and more about our cosmos. How enormously vast it is, how grand it is, how beautiful it is. And through this understanding we've also come to discover technologies that enable us to take more control and actually start to shape our cosmos giving us the opportunity to make life flourish far beyond what our ancestors had dreamt of.
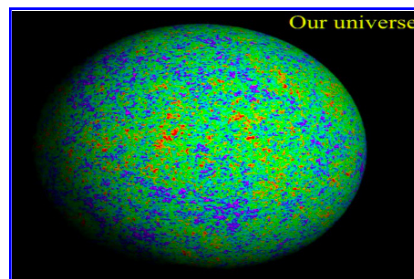
So we've done a lot of various inspiring things. But we've also done some dumb things, and even some extremely dumb things here in our universe. One of the bad habits I have as a professor is I like to give grades, sometimes unsolicited. So I thought what grade should I give for humanity for Risk Management 101 here, 13.8 billion years in?

And I figured, well, you know, I asked some friends. They said, maybe a B-plus. You know, we've done a lot of dumb stuff, a lot of close calls like the Cuban Missile Crisis, but we're still here, so maybe a B-plus. But from a cosmic perspective actually, I really have to give a D-minus. Even though, as Theodore can certify, that's not an allowed grade at MIT. D is the lowest—above F—there is.

Why a D-minus? Because from a standard perspective a lot of people feel that humans are the pinnacle of evolution, this is where we got this planet, we're limited to it. Some people are very obsessed about the next 50 years, maybe the next election cycle even. Right? So, if we wipe ourselves out in 50 years maybe it's not such a big deal. From a cosmic perspective that is completely retarded. We ain't seen nothing yet. It would be completely naïve, in a cosmic perspective, to think this is best, this is as good as it can possibly get.

We have 10 to the power 57 times more volume at our disposal for life out there. We don't have 50 years. We have billions and billions of years available. We have an incredible future opportunity that we stand to squander if we go extinct or in other ways screw up.

People argue passionately about what the probability is that we wipe out in any given year. Some might say it's one percent. Some might say it's much lower—a percent of a percent. Some might

say it's higher—10 percent. Any of these numbers are just completely pathetic. If it's one percent you expect maybe we'll last for 100 years. That's pretty far from the billions of years of potential we have there, right? So, come on, let's be a little more ambitious here.

Let me just summarize in one single slide why I think it's so pathetic, how reckeless we are being stewards of life. Namely this slide.

Which one of these two people is more famous? Let me ask you one more question. Which one of these two people should we thank for us all being alive here today because he single-handedly perhaps stopped, or prevented a Soviet nuclear attack during the Cuban Missile Crisis? One clue: wasn't Canadian.

So these are some pretty screwed up priorities we have as a species.

When I first became aware of this nuclear situation, and I was about 14, I was really quite shocked by how so many grownups could be so dumb. When I was 17 I felt I wanted to do whatever little things I can do for this. I was in Stockholm, Sweden, I went and volunteered to write some articles for a local magazine. I wrote a bunch of articles about nuclear weapons and nuclear war and so on.

The oldest article, ever, to my knowledge about the U.S. hydrogen bomb project—which from my physicist point-of-view was when it started getting incredibly scary—is this one ["Experiment in Annihilation," by Jules Laurents], from 1954, which, to my knowledge, for the first time really lays out what had largely been unknown to the broader public; the fact that America had just done its fourth hydrogen bomb test and that there had actually been three [earlier] ones before that.

EXPERIMENT IN ANNIHILATION*

MARCH 1, 1954, the same day that shots were ringing on the floor of the House of Representatives, another "shot", unheralded but of every ing significance, was fired in the Marshall Islands. On that day an American AEC task group detonated a hydrogen bomb of monstrous size. In its widest implications that bomb has not yet ceased to reverberate. A long chain of incidents, ranging from the curious to the tragic, has made it clear that "peacetime" nuclear explosions present a substantial threat to our well-being. Storm signals from earlier atomic tests such as fogged photographic film and radioactive rain have given way to the storm which has already resulted in the radioactive poisoning of several hundred people. The March 1 explosion also blasted the lid of secrecy from the AEC's thermonuclear adventures, giving the public its first real look behind the "uranium curtain"; thus it is now known that the AEC touched of three prior hydrogen explosions, the third of which (November, 1952) gave more than five times as great an energy release as predicted by its creators.

1. Chronicle of Events

The March 1 bomb was expected to explode with a force of four to six megatons (a megaton denotes the energy released by exploding one million tons of TNT) but developed instead about fourteen, according to Joseph and Stewart Alsop, *New York Herald Tribune*, April 7, 1954. It left scientific measuring instruments unable to record its full effects. Sound waves from the blast were detected in London, and an American astronomer said the flash could have been seen from Mars. Rep. Holifield of the JCAE (Joint Committee on Atomic Energy) described it as "so far beyond what was expected you might say it was out of control". Defense Secretary Wilson called it "unbelievable", and President Eisenhower admitted it "surprised and astonished" the scientists. Rep. Van Zandt of the JCAE stated that the "explosion had left an area of total destruction about twelve miles in diameter, with light damage extending in a circle with a diameter of forty miles". The AEC called it a "routine atomic test". As with the November, 1952 H-bomb, the first inkling the public had that something extraordinary had occurred was through "leaks". Intent on maintaining secrecy, the AEC ordered all task force personnel to refrain from divulging any information about the tests. Such an order was not given on Kwajalein, however, 176 miles from Bikini, it being assumed apparently that at this distance details of the explosion could not be perceived. Yet a marine corporal stationed there wrote to his mother:

"I was walking back to the barracks . . . just as it was getting daylight, when all of a sudden the sky lighted up a bright orange. . . . About ten

*This is the first of two articles dealing with the recent hydrogen explosions and studies the damage they have wrought upon man and his environment. Historical and scientific background material is presented in the Supplement at the end. The second article will go more deeply into the social implications.

214

or fifteen minutes later . . . we heard very loud rumbling that sounded like thunder. Then the whole barracks began shaking as if there had been an earthquake. This was followed by a very high wind."

In a second letter he reported: "There were two destroyers here to-day bearing natives of one of the Marshall Islands that was within seventy-five miles of the blast. They were suffering from various burns and radioactivity".

Directly thereafter the AEC issued the following statement:

"During the course of a routine atomic test in the Marshall Islands, twenty-eight U.S. personnel and 236 residents were transported from neighboring atolls to Kwajalein Island according to plans as a precautionary measure. The individuals were unexpectedly exposed to some radiation. There were no burns. All were reported well. After the completion of the atomic tests, they will be returned to their homes."

The AEC never acknowledged the statement of the corporal, nor his assertion that some victims were suffering from burns. (We shall see that the AEC statement is false.) When the announcement was made some observers were puzzled over how, after the victims were "unexpectedly" exposed to radiation, they were evacuated "according to plans as a precautionary measure". *Time* magazine introduced additional cause for apprehension by reporting that American casualties from the March 1 explosion were exposed to radiation "ten times greater than scientists deem safe".

On March 13* a grave new consequence of the "routine atomic test" was reported. The Japanese fishing trawler *Fukuryu Maru* docked in Yaezu, Japan, with its twenty-three crew members showing symptoms of acute radiation exposure. They told how on March 1 they were some eighty to ninety miles from Bikini, when at 4 a.m. they fancied they saw the sun rising prematurely "in a strange manner". Six or seven minutes later they heard a roar, and two hours later they were showered with a white ash, which continued to fall for several hours. The ash was, of course, fall-out from the explosion, consisting mainly of irradiated coral dust. Only after they had become quite ill did they suspect that they had been rained with *shi no hai* (ashes of death) and head for port. They had on board 40 tons of freshly caught tuna and shark, which, according to a *New York Times* dispatch, exhibited radioactivity "sufficient to be fatal to any person who remained for eight hours within thirty yards of the fish". Two of them were in worse condition than the rest, having eaten some of the fish. The crewmen were hospitalized, the sampan was ordered burned at sea and sunk, the fish buried; but not before several thousand pounds of the contaminated fish had been unloaded and shipped to market. A "hot fish" panic ensued in Japan, and police, in a frantic effort to track it all down, ordered a thousand tons of other fish destroyed, with which it had got mixed. Fish prices dropped to half overnight, and Tokyo's numerous *sushi* houses (*sushi* — a popular fish dish) reported business at a standstill. It should not be necessary to give all details — it is sufficient to recall the Japanese experience with atomic bombs, coupled with the fact that fish is the mainstay of the Japanese diet (a million pounds a day of tuna alone are consumed in Japan) to appreciate the extent of the panic. The people's

*All dates without years refer to 1954.

As you see here this explosion here had lifted the "uranium curtain"—a lot of things which had been done very much behind the back of the American people and even a lot of politicians actually seemed kind of reckless—that the fourth blast here was five times more powerful than had been anticipated. A lot of Japanese fisherman got radiation poisoning for being in the area, et cetera, et cetera.

Now this article was translated into French by Jean-Paul Sartre. It was actually read into the Congressional Record by an American politican who gave no attribution whatsoever to where he had gotten this article. And nobody knew actually. Still nobody really knows publically, who wrote this article because Jules Laurents doesn't exist.

This was written by someone who was so worried about getting in trouble with the McCarthy folks during the time that he wrote this under a false name. So I figured in honor of this meeting I would tell you who wrote this. You will be the first to know. It was my father. Harold Shapiro wrote this article. And if anyone wants it I can e-mail you a copy.
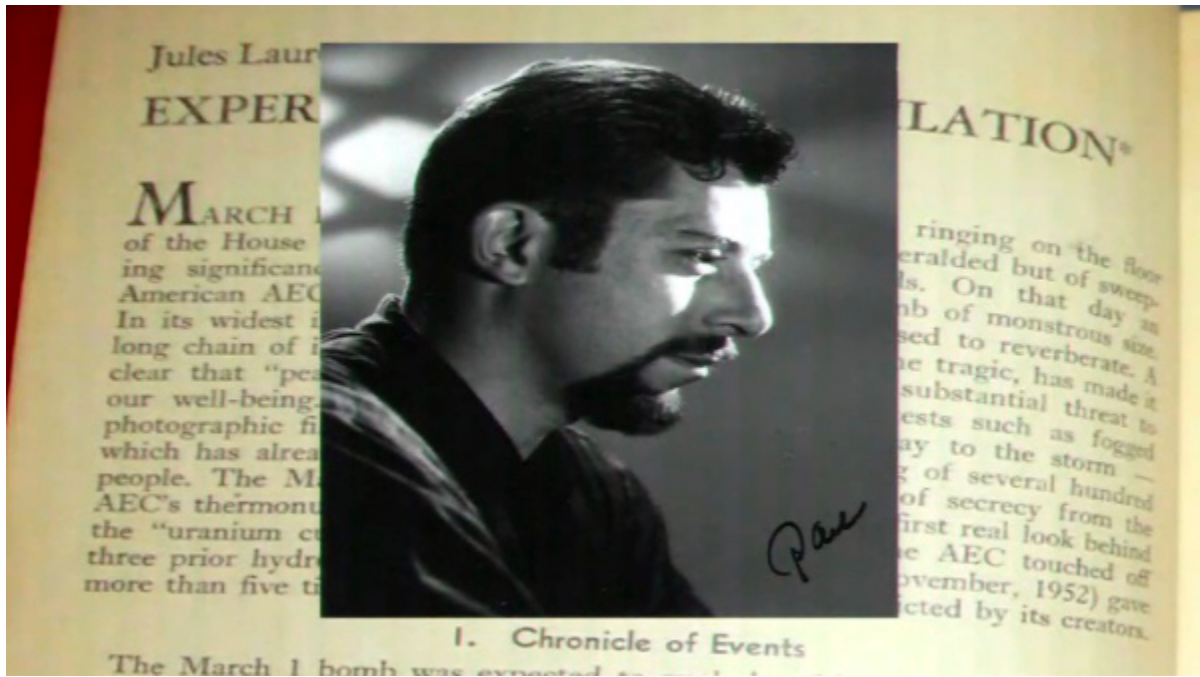
Coming back to the cosmic perspective, to emphasize how stupid I feel we're being as a life form, let me just tell you the way I see this in simple cartoon form.

Here we are on this planet, and we humans have decided to build this device. Let's cartoon-fashion draw it like this okay?
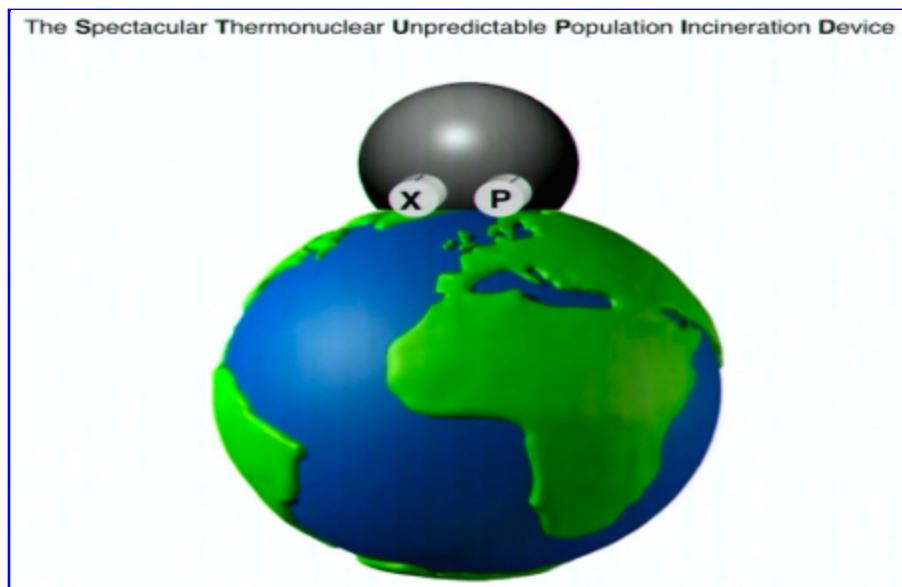
It's called the Spectacular Thermonuclear Unpredictable Population Incineration Device. Okay, I'm a little bit inspired by Dr. Seuss here, I have to confess. This is a long mouthful so let's just abbreviate it: S-T-U-P-I-D. Okay?

This device—it's a very complicated device—it's a bit like a Rube Goldberg machine inside. A *very* elaborate system. Nobody—there's not a single person on the planet who actually understands how 100 percent of it works. Okay?

But we do know some things about it. It has two knobs on the front, X and P, which I'll explain to you shortly. And it was so complicated to build that it really took the talent and resources from more than one country, they worked really hard on it, for many, many years. And not just on the technical side—to invent the technology to be able to create what this device does. Namely, massive explosions around the planet.
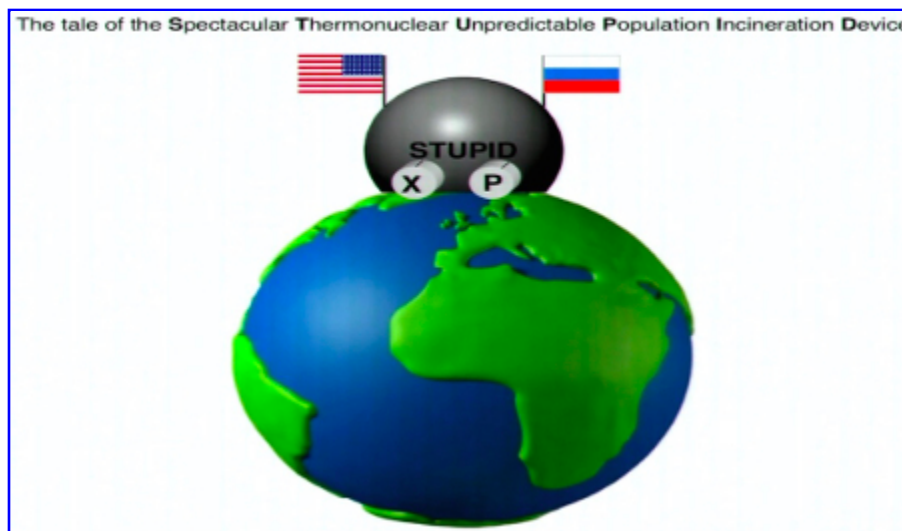
But also to overcome a lot of human inhibitions towards doing just this. So this system actually involves also a lot of very clever social engineering where you put people in special uniforms and have a lot of peer pressure and you use all the latest social coercion technology to make people do things they otherwise normally wouldn't do. Okay?

And you do fake tests and in the event of people who fail to launch the missiles, you fire them, replace them.

And so a *lot* of clever thought has gone into building STUPID.

That's what this device does. It's kind of remarkable that we went ahead and put so much effort into building it since actually, really, there's almost nobody on this spinning ball in space who really wants it to ever get used, who ever wants this stuff to blow up.

But we'll continue talking throughout the conference about why we humans decided—made it anyway.

Let's focus now instead a bit on how it works. What are these two knobs? The X knob determines the total explosive power that this thing brings to bear and the P knob, it determines the probability that this thing will just, BOOM!, go off in any random year for whatever reason.

As we'll see, one of the cool features of it is that it can spontaneously go off even if nobody actually wants it to. Alright?

So you can tune these two knobs X and P. Let's look a little bit at how this has evolved over time, the settings of these dials.

Of course in 1945, I feel personally guilty about this being a physics professor because the knob was set to zero until we physicists came on the scene and figured out how to ramp up X here.
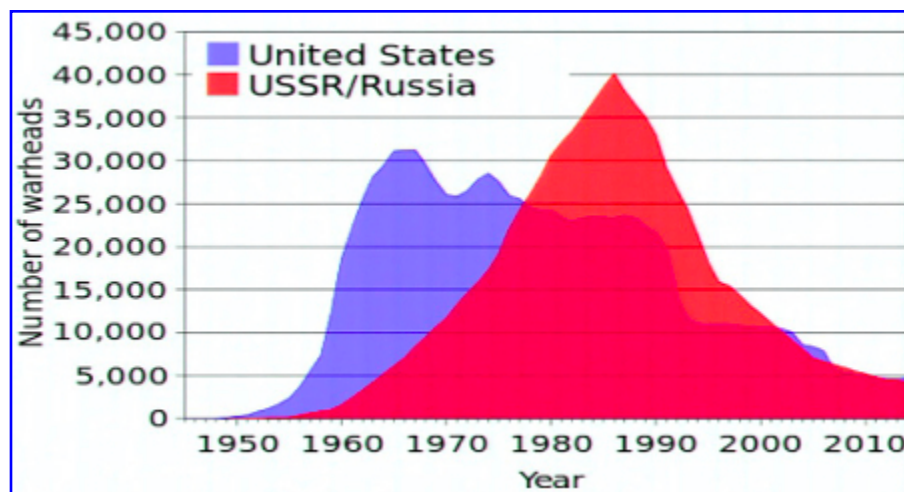
This is a plot of how the number of warheads has evolved over time. You guys are all quite familiar with this. Of course, it's not just that the number of warheads has changed. We started

out below 20 kilotons with Hiroshima and Nagasaki. By the time we're up to Tsar Bomba we're up to 50 megatons; 3,000 tims more powerful.

We peaked in the total, in the setting of the X knob around the mid-eighties with about 63,000 warheads. Since then, the total number of warheads, as you know, has gone down quite a bit. But sadly the drop has stopped and things haven't gone down much at all in the last decade.

Status of World Nuclear Forces 2014*

Country	Operational Strategic	Operational Nonstrategic	Reserve/ Nondeployed	Military Stockpile	Total Inventory
Russia	1,800	0	2,500	4,300	8,000
United States	1,920	184	2,661	4,765	7,315
France	290	n.a.	?	300	300
China	0	?	250	250	250k
United Kingdom	160	n.a.	65	225	225l
Israel	0	n.a.	80	80	80
Pakistan	0	n.a.	100-120	100-120	100-120
India	0	n.a.	90-110	90-110	90-110
North Korea	0	n.a.	<10	<10	<10
Total:	~4,100	~180	~5,800	~10,100	~16,300

* All numbers are approximate estimates and further described in the Nuclear Notebook in the *Bulletin of the Atomic Scientists*, and the nuclear appendix in the SIPRI Yearbook. Seealso status and 10-year projection of U.S. and Russian forces. Additional reports are published on the FAS Strategic Security Blog. Unlike those publications, this table is updated continuously

This is roughly where we stand today. About 16,000 hydrogen bombs, about 4,100 of them on hair-trigger alert—meaning they can be launched on 5 to 15 minutes notice. [Hans M. Kristensen and Robert S. Norris, "Worldwide deployments of nuclear weapons, 2014", *Bulletin of the Atomic Scientists*, Nuclear Notebook, Aug 26, 2014]

A lot of my friends, unfortunately, take the mere fact that this has gone done—this curve—is their reason to stop worrying about this. Which I think is a very bad idea.
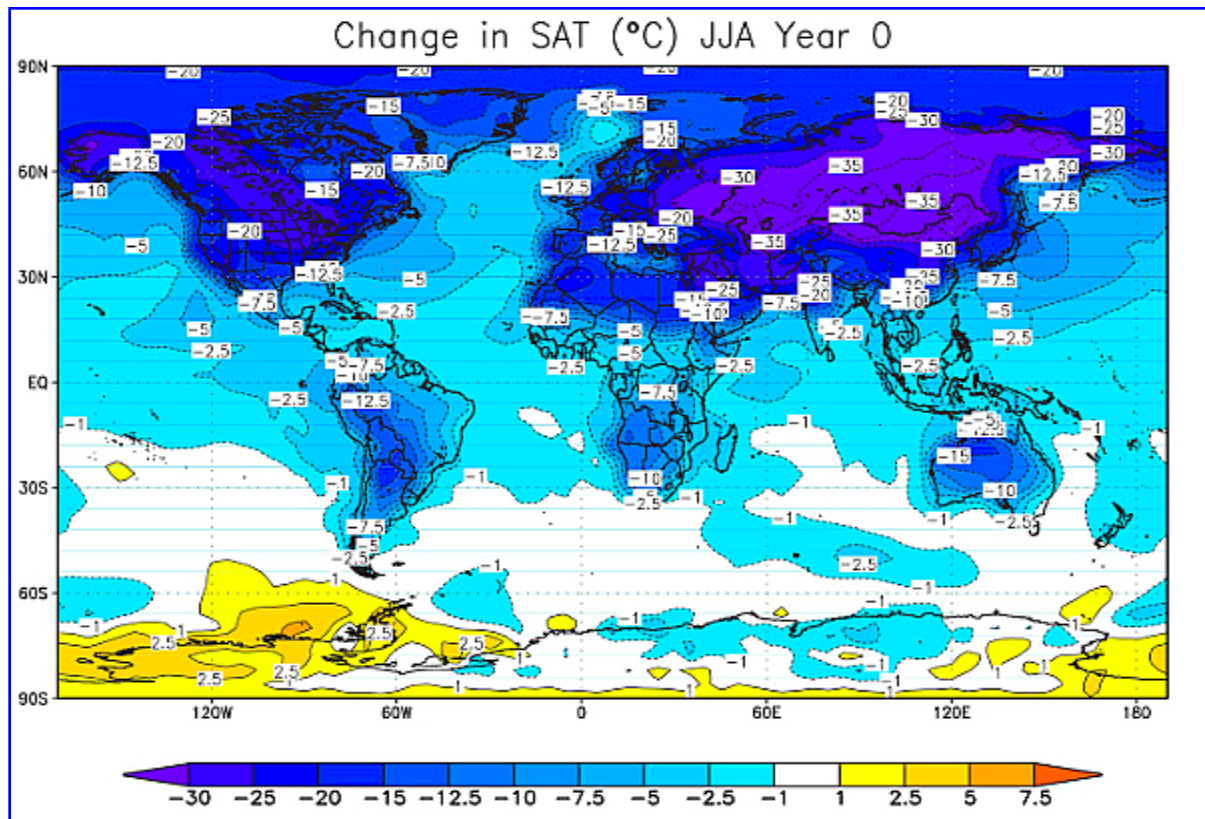
Nuclear winter: average change in surface air temperature (°C) during 1st 2 years

Now that we have much better climate modelling—this is a paper I really liked by Robock et al [Robock, Alan, Luke Oman, and Georgiy L. Stenchikov, 2007b: Nuclear winter revisited with a modern climate model and current nuclear arsenals: Still catastrophic consequences. *J. Geophys. Res.*, 112, D13107, doi:2006JD008235.] and when I made this talk and I put in this graph I had absolutely no idea that the speaker after me is, in fact, Robock! So I am very honored as I did not put this in to make you feel good. I put it in because it's a fantastic piece of work.

What you're seeing here—and we'll hear much more about it, of course, in the next talk—is simply the average surface temperature change during the two years after a global nuclear war with roughly today's arsenals. And it's in Celsius—so you can see, typically drop the temperature by about 20 Celsius throughout most of the American breadbasket here. And some parts in the Soviet farming areas it drops by 35 Celsius—70 degress Farenheit.

What does that mean in plain English? You don't have to think very hard, you don't have to have a great imagination to imagine that if you turn this corn field into this, you might have some impact on the world food supply.

You don't have to be very creative either to imagine that if you have total infrastructure collapse and mass starvation there are going to be a lot of other things which are really hard for us to predict. But we certainly can't rule out pandemics on a scale we haven't seen since the Great Plague. And moreover, having massive amounts of handguns and things around, that whoever survives, we'll have, obviously armed gangs going from house to house doing enormous damage to whatever survivors there are. It's clearly not a situation we would like to put ourselves in.

So taking the setting of this X knob is so low now that we should stop worrying would be the ultimate naïvete in my opinion.
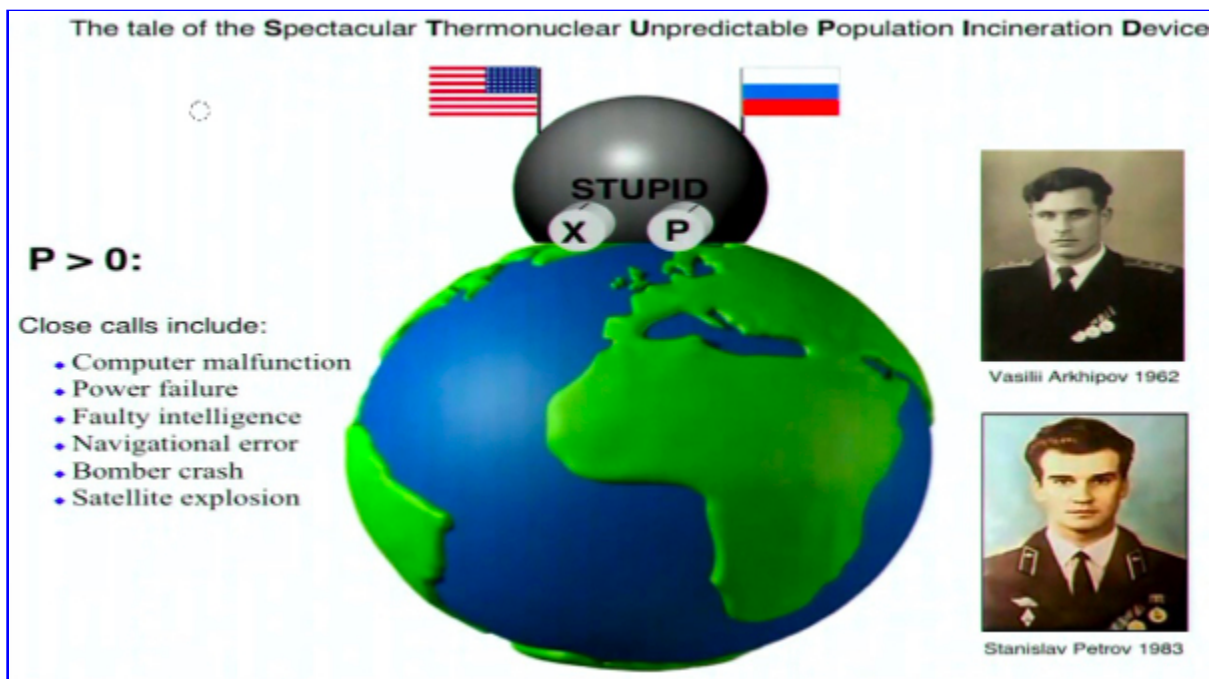

Let's talk about the other knob, P: the probability that this thing just goes Ka-boom!, for whatever reason. My own view is that the most likely way we'll get a nuclear war going is by accident—which can also include people through various sorts of misunderstandings.

We know for sure—we don't know what P is, obviously; there's good debate about it, we should

discusss it here at the meeting—but we know very rigorously it's not zero. Because as so many of your are very well aware there have been enormous numbers of close calls caused by all sorts of things: computer malfunctions, power failure, faulty intel, navigational error, a crashing bomber, an exploding satellite, et cetera. [Eric Schlosser, *Command and Control: Nuclear Weapons, the Damascus Accident, and the Illusion of Safety* (New York: *Penguin*), 2013]

And, in fact, if it weren't for very heroic acts of certain people—we'll talk more about, Vasili Arkhipov for example, and Stanislav Petrov—we might already have had a global nuclear war.

So P is not zero. What about the change of P over time? We talked about how X has changed. How has P changed?
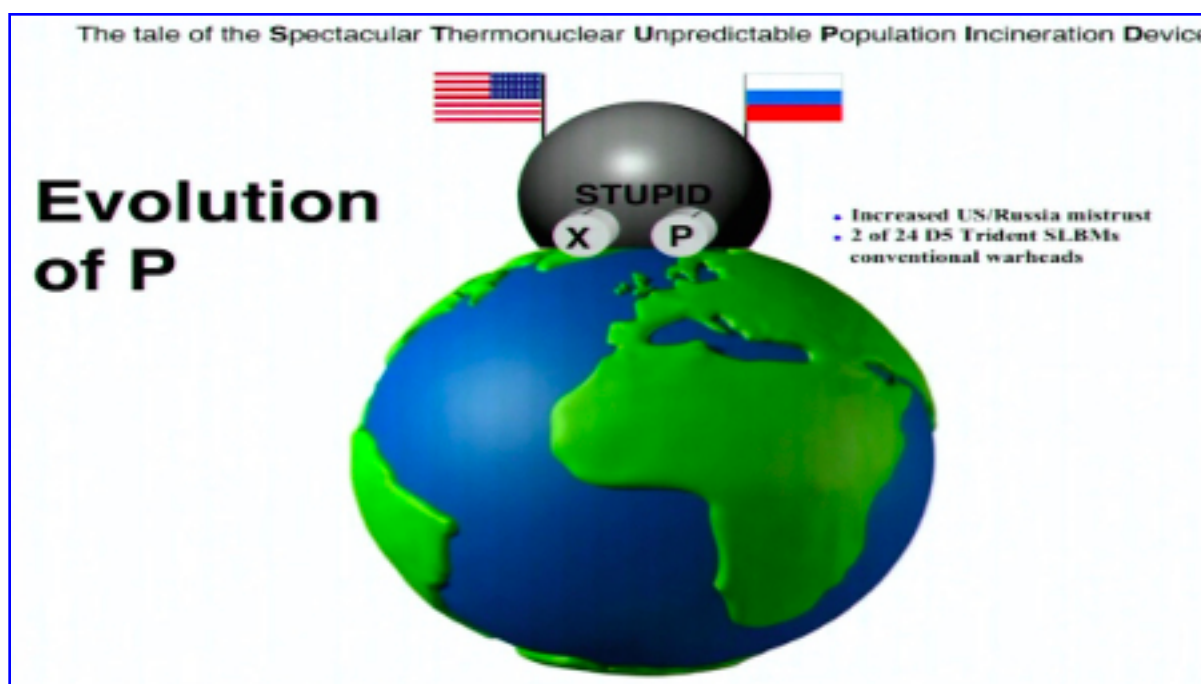

We heard a very powerful argument here from Theodore Postol that even though P certainly dropped after 1990, when the U.S. and Russia decided to chill out quite a bit, it might very well have gone up quite a bit again.

There are various reasons for this. Obviously increasing U.S.-Russian mistrust is a very bad thing and that's certainly happening now.

Then there are a lot of just random, dumb things that we do which increase P. Just one little example amoung many that's been discussed is this plan to replace 2 out of the 24 submarine launched ballistic missiles on the Tridents by conventional warheads that you can fire at North Korea. A *great* setup for misunderstanding, since if you're the Russians and you see this missile

coming you have absolutely no way of knowing what kind of warheads it has.
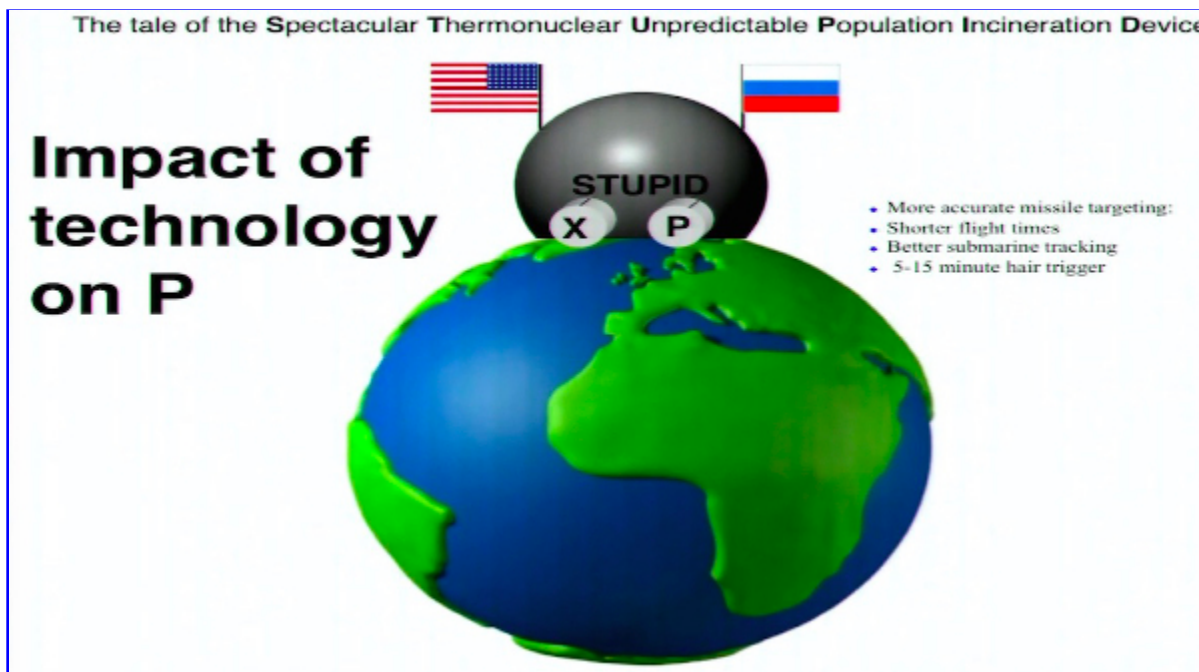
Let me spend my last five minutes talking about the impact of technology on P; the impact of technology on the risk of accidental nuclear war. A lot of the dumb things have been caused by just people and social things. But technology, obviously, has a powerful effect on these things.

We heard from Theodore Postol already various examples of how technology is perhaps increasing the risk of accidental nuclear war. Mutual Assured Destruction worked great when missiles were accurate enough to destroy a city but not accurate enough to destroy a silo. That made it very disadvantageous to do any kind of first strike.

Now, we're seeing—thanks to early forms of artificial intelligence that are really enabling very precise targeting of missiles—you can hit very very accurately, that's better for a first strike. Having submarine-launched balistic missiles very close to their targets also is good for a first strike—you get less time for the enemy to react—and these very short flight times and also the better ability to track where the enemy submarines are and take them out means a lot of people are a lot jumpier, there's very short times to decide, so both the U.S. and Russia, of course, are on hair-trigger alert—launch-on-warning where decisions, you have only 5 to 15 minutes to decide; obviously to that some things like this can increase P.

What about artificial intelligence? We heard from Helen how, there's a broad consensus that artificial intelligence is progressing very rapidly. In fact I've spent a lot of time—I just came back last month from a conference in Puerto Rico, that my wife and I and many of my colleagues

organized where we brought together many of the top AI builders in the world to discuss the future of AI.

And a lot of people felt that things that they thought were going to take 20 years to happen, 5 years ago, have already happened. There's huge progress. Obviously, it's very hard to forecast what will happen looking decades ahead if we get human-level AI or beyond.

But we can say some things about what's going to happen much sooner. And what's already happening as computers get more powerful and have more and more impact on the world.

For example, if you can make computer systems that are more reliable than people in properly following proper protocol, it's an almost irresistible temptation for the military to implement them.

We've already seen a lot of the communications and command and even analysis being computerized in the military. Now properly following proper protocol might sound like a pretty good thing until you read of the Stanislav Petrov incident. Why was it, in 1983, when he got this alarm that the U.S. was attacking the Soviet Union, that he decided *not* to pass this along to his superiors? He decided to *not* follow proper protocol. He was a person. If he had been a computer he *would have* followed proper protocol. And maybe something much worse would have happened.

Another thing which is disturbing about computerizing things more and more is that we know that more the you de-personalize decisions, the more you take "system 1" (as Kahneman would say) out of the loop, the more likely we are to do dumb things. [Daniel Kahneman, *Thinking Fast and Slow* (New York: Farra, Straus and Giroux), 2013]

If President Obama had a person with him who he was friends with who carried the nuclear launch codes surgically implanted next to her heart, and the only way for him to get them was to stab her to death first, that would actually make him think twice before starting a nuclear war. And it might be a good thing, right?

If you take that away, if all you need to do is press a button, less inhibitions. If you have a super-advanced artificial intelligence system that you just delegate the decision to it's even easier because you're not actually authorizing launch. Right? You're just delegating the authority to this system that IF something happens in the future, then please go ahead and properly follow proper protocol. Right? That worries me.

Then there are bugs, right? Raise your hand if you've even been given the blue screen of death

from your computer. Let's hope the blue screen of death never turns into the red sky of death. This is funny if it's just 2 hours of your presentation that got destroyed but it's not so funny if it's your planet.

Finally, another thing which is happening as artificial intelligence systems get more and more advanced is they become more and more inscrutable black boxes where we just don't understand what reasoning they use though we still trust them.

I'm driving on my GPS, I was just last week, we were up in New Hampshire with the kids and my GPS said "Turn Left on Rufus Colby Road". We drive down there and suddenly there's this enormous snow bank blocking the road. I have no idea how it came to that conclusion but I trusted it.

If we have a super-advanced computer system which is telling all the Russian military leadership and Putin that, 'Yes, there is an American missile attack happening right now—here is the cool map, high-res graphic'—they might just trust it without knowing how it came to the conclusion. If it's a human, you can ask the human, 'How did you come to this conclusion?' You can challenge them. You can speak the same language. It's much harder, these days, to query a computer and clear up misunderstandings.

So I'm not standing here saying We know for sure that AI is going to increase the risk of accidental nuclear war. But we certainly can't say it won't. And it's very likely that it will have strong effects. This is something we need to think about. It would be naïve to think that the rise of artificial intelligence is going to have no impact on the nuclear situation.
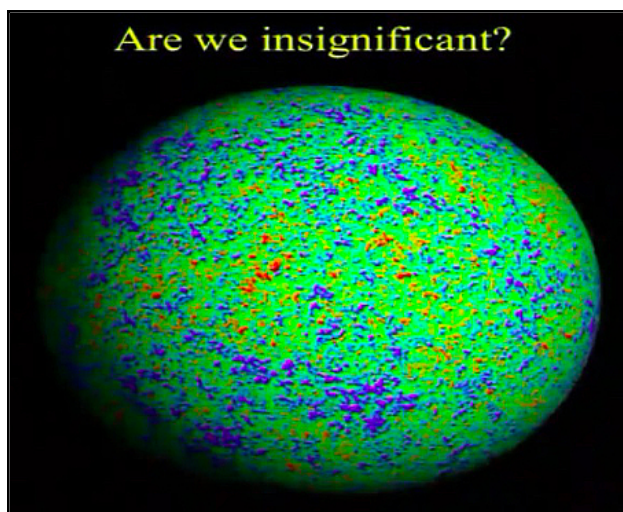
Let me conclude by coming back to the cosmic perspective again. It's easy, when you look up into our cosmos and see how big it is, to feel small and insignificant. In fact, I started feeling

more and more insignificant the more I learned about the size of the cosmos in my scientific career.

Until I had a total u-turn. Because, we have discovered that, Yes, first of all, there are way more planets than we thought there were. But we've also discovered that it seems like life, advanced enough to build telescopes and technology like we have, is much more rare than you might have thought.

In fact we haven't found any evidence at all so far that there is any [life] anywhere in our observable universe, besides us. We don't know which way it is. I argue in my book that we are probably the only life within this region of space that we have access to that's come this far.

Which, if it's true, makes a huge responsibility. Why are all these galaxies beautiful? It's because you see them. That's why they're beautiful.

If we annihilate life and there's no consciousness with telescopes they are not beautiful anymore. They're just a giant waste of space.

So what I'm saying here is that rather than look to our universe to give meaning to us. It's we who are giving meaning to our universe. And we should really be good stewards of this.

Because of this, as Helen Caldicott mentioned, I'm the President of the Future of Life Institute we founded to really try to focus humanity on to being better stewards of this incredible

opportunity we have.

Of course, we all love technology. Every way in which 2015 is better than the stone age is because of technology. But it's absolutely crucial that before we just go ahead and develop technologies to be powerful also develop the wisdom to handle that technology well.

Nuclear technology is the first technology powerful enough that it's really, really driven this whole—artificial intelligence is another example of this.

We, with our organization have so far spent most of our effort on things to do with AI. But we care deeply about nuclear issues as well and we have a lot of awesome people in our organization.

We are very eager to hear from you ideas for how we can help make the future of life actually exist. How we can actually help all of your efforts to keep the world, to keep ourselves safe from nuclear weapons.

Thank you.