Transcript

WHY. The Problem with AI. SafeAI Forever.

WANTED: Provably Safe AGI for The Benefit of People, Forever.

"Our thinking machines must be *contained*, forever. *Never* allow them to take control of our humanity"

Learn more: The AI Safety Problem. Attribution of Quotes and Sources…

"Cogito, ergo sum." (I think, therefore I am.)
René **Descartes** (1637)

"There is thus this completely decisive property of complexity, that there exists a critical size below which the process of synthesis is degenerative, but above which **the phenomenon of [AI] synthesis, if properly arranged, can become explosive**, in other words, where syntheses of automata can proceed in such a manner that each automaton will produce other automata which are more complex and of higher potentialities than itself… **Technological power.** as such is always an ambivalent achievement, and science is neutral all through, providing only means of control applicable to any purpose, and indifferent to all. It is not the particularly perverse destructiveness of one specific invention that creates danger. **The danger is intrinsic. For progress there is no cure."**
John von Neumann (1946)

"**It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers**… They would be able to converse with each other to sharpen their wits. At some stage therefore, we should have to **expect the machines to take control**."
Alan Turing (1951)

"As machines learn they may develop unforeseen strategies at rates that baffle their programmers."
Norbert Weiner (1960)

"Let an ultra-intelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then **unquestionably** be **an "intelligence explosion"** and the intelligence of man would be left far behind. Thus the first ultra-intelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. It is curious that this point is made so seldom outside of science fiction. It is sometimes worthwhile to take science fiction seriously."
I.J. Good (1966)

"The development of **full artificial intelligence could spell the end of the human race**. It would take off on its own and re-design itself at an ever increasing rate. Humans, who are limited by slow biological evolution, couldn't compete, and would be superseded."
Stephen Hawking (2014)

These things understand… My intuition is: we're toast. This is the actual end of history.
Geoffrey Hinton (2023)

"If someone does crack the code and build a superintelligence … I'd like to make sure that we **treat this at least as seriously as** we treat, say, **nuclear material**.... We think this technology, the benefits, the access to it, the governance of it, belongs to humanity as a whole." --- Sam **Altman** (2023)

"A.I. might be even **more urgent than climate change** if you can imagine that." --- Christiane **Amanpour** (2023)

"**These systems are inherently unpredictable.** Controlling them is different than controlling expert systems. Mechanistic interpretability... is the science of figuring out what's going on inside the models. " --- Dario **Amodei** (2023)

"**Because there is such uncertainty**, I think we collectively in our governments, in particular, have a responsibility to **prepare for the plausible worst case which might be like five years.**" --- Yoshua **Bengio** (2023)

"**Tech companies have a responsibility, in my view, to make sure their products are safe before making them public.**" --- President Joe **Biden** (2023)

"I think there is a significant chance that we'll have **an intelligence explosion**. So that within a short period of time, we go from something that was only moderately affecting the world, to something that **completely transforms the world.**" --- Nick **Bostrom** (2023)

"**Once super intelligence is achieved, there's a takeoff**, it becomes exponentially smarter and in a matter of time, they're just, **we're ants and they're gods.**" --- Lex **Fridman** (2023)

"**This is a Promethean moment** we've entered." --- Thomas **Friedman** (2023)

"**It really is about a point of no return**. Where if we cross that point of no return we have very little chance to bring the genie back into the bottle." - Mo **Gowdat** (2023)

"Storytelling computers will change the course of human history... There are two things to know about AI, it's the first technology in history that can make decisions by itself and it's the first technology in history that can create ideas by itself... this is **nothing like anything we have seen in history**... **it's really an existential risk to humanity** and what we need above all is time. Human societies are extremely adaptable, we are good at it, but it takes time... ultimately the problem is us, not the AI... One of the dangers we are facing now with this new technology is that if we don't make sure that everybody benefits then we might see the greatest inequality ever emerging because of these new technologies- this is a certainly a very very big danger." --- Yuval **Hariri** (2023)

"**Just think about how we relate to ants.** We don't hate them. We don't go out of our way to harm them. In fact, sometimes we take pains not to harm them. We step over them on the sidewalk. But whenever their presence seriously conflicts with one of our goals, let's say when constructing a building like this one, we annihilate them without a qualm. The concern is that we will one day build machines that, whether they're conscious or not, **[AI] could treat us with similar disregard.**" --- Sam **Harris** (2023)

"I've always believed that it's going to be **[AI is] the most important invention that humanity will ever make**." --- Demis **Hassabis** (2023)

"**50% of AI researchers believe there's a 10% or greater chance that humans go extinct from our inability to control AI.**" --- Tristan **Harris** (2023)

"I should certainly hope that the risk of extinction is not unsolvable or else we're in big trouble... AI scientists from all the top universities and many of the people who created it are concerned that AI could even lead to extinction... we need to cooperate... **we don't want another arms race where we build extremely powerful technologies that could potentially destroy us all**." --- Dan **Hendrycks** (2023)

"**My intuition is: we're toast. This is the actual end of history**... It is like aliens have landed on our planet and we haven't quite realised it yet because they speak very good English... It's conceivable that the genie is already out of the bottle... If we allow it to take over, it will be bad for all of us. We're all in the same boat with respect to the existential threat. So we all ought to be able to cooperate on trying to stop it... put comparable amount of effort into making them better and understanding how to keep them under control... it's a whole different world when you're dealing with things more intelligent than you... you don't really understand something until you've built one." — Geoffrey **Hinton** (2023)

"**Every expert I talk to says basically the same thing:** We have made no progress on interpretability, and while there is certainly a chance we will, it is only a chance. For now, **we have no idea what is happening inside these prediction systems**... If you told me you were building a next generation nuclear power plant, but there was no way to get accurate readings on whether the reactor core was going to blow up, I'd say you shouldn't build it." - Ezra **Klein** (2023)

"The pace of change and its potential presents **an almighty challenge to governments around the world**." --- Laura **Kuenssberg** (2023)

"When thinking about black swan events such as the creation of **AGI or human extinction**, we need to **be open to extreme possibilities** that have never happened before and don't follow previous events." --- Stephen **McAleese** (2023)

"**With artificial intelligence we are summoning the demon**… If it's engaged in recursive self-improvement... and it's optimisation or utility function is something that's detrimental to humanity, then it will have a very bad effect." --- Elon **Musk** (2023)

"It appears that AGI and ASI are imminent… weak AGI which is sort of an AGI which can do anything a remote human worker can do is due in about 2026 and a stronger one based on robotics is due in 2031, and once AGI shows up, artificial super intelligence (ASI) is estimated six months after that- so we're talking very near term probably the next decade or so... Oh yes. AI should not be able to independently launch nukes on their own. Whoa great. But **we need absolute technical guarantees of [AI safety]** that." --- Steve **Omohundro** (2023)

"**We may look on our time as the moment civilization was transformed** as it was by fire, agriculture and electricity. In 2023 we learned that a machine taught itself how to speak to humans like a peer which is to say with creativity, truth, error and lies. The technology known as a chatbot is only one of the recent breakthroughs in artificial intelligence machines that can teach themselves superhuman skills." --- Scott **Pelly** (2023)

"**It can be very harmful if deployed wrongly** and we don't have all the answers there yet – and the technology is moving fast. So does that keep me up at night? Absolutely... When it comes to AI, avoid what I would call race conditions where people working on it across companies, get so caught up in who's first that we lose, you know, the potential pitfalls and downsides to it." --- Sundar **Pichai** (2023)

"**A very strange conversation** with the chatbot built into Microsoft's search engine **left me deeply unsettled. Even frightened.** Then, out of nowhere, Sydney declared that it loved me — and wouldn't stop, even after I tried to change the subject." --- Kevin **Roose** (2023)

"But the real thing to know is that we honestly don't know what it's capable of. The **researchers don't know what it's capable of**. There's going to be a lot more research that's required to understand its capacities. And even though that's true, it's already been deployed to the public." --- Aza **Raskin** (2023)

"**Our planet**, this 'pale blue dot' in the cosmos, **is a special place**. It may be a unique place. And **we are it's stewards in an especially crucial era**... It's an important maxim that the unfamiliar is not the same as the improbable."--- Martin **Rees** (2023)

"**We're at a Frankenstein moment.**" --- Robert **Reich** (2023)

"**How we choose to control AI is possibly the most important question facing humanity.**.. Imitation learning is not alignment. Machines are beneficial to the extent that their actions can be expected to achieve our objectives." **---** Stuart **Russell** (2023)

"**The existential risk of AI is defined as many, many, many, many people harmed or killed**... I've been through time-sharing and the PC industry, the web revolution, the Unix revolution, and Linux, and Facebook, and Google. And this is growing faster than the sum of all of them." --- Eric **Schmidt** (2023)

"**We need rules. We need laws. We need responsibility. And we need it quickly**... Let's start to get some legislation moving. Let's figure out how we can implement voluntary safety standards." --- Brad **Smith** (2023)

"**This situation needs worldwide popular attention**. It needs answers, answers that no one yet has. Containment is not, on the face of it, possible. And yet for all our sakes, containment must be possible." --- Mustafa **Suleyman** (2023)

"Artificial intelligence or **AI is evolving at light speed** with profound consequences for our culture, our politics and our national security." --- George **Stephanopoulos** (2023)

"AI has an incredible potential to transform our lives for the better. But **we need to make sure it is developed and used in a way that is safe and secure.** Time and time again throughout history we have invented paradigm-shifting new technologies and we have harnessed them for the good of humanity. That is what we must do again. No one country can do this alone. **This is going to take a global effort**. But with our vast expertise and commitment to an open, democratic international system, the UK will stand together with our allies to lead the way." --- The Prime Minister, The Rt Hon **Rishi Sunak MP** (2023)

"The problem is that with **these experiments** is they **are producing uncontrollable minds**... I've not met anyone in AI labs who says the risk is less than 1% of blowing up the planet. It's important that people know **lives are being risked**.... So the question is: what kind of firewalls are in place to make sure that it can't self-deploy." --- Jaan **Tallin** (2023)

"There might simply not be any humans on the planet at all. **This is not an arms race it's a suicide race.** We should build AI for Humanity, by Humanity." --- Max **Tegmark** (2023)

"The trouble is it does good things for us, but **it can make horrible mistakes by not knowing what humanness is**." --- Steve **Wozniak** (2023)

"OK, **it's time to freak out about AI**... So different is AI from past technological challenges that the jury's out on whether truly effective international regulation is possible." --- Robert **Wright** (2023)

"**This is a very lethal problem**, it has to be solved one way or another... **and failing on the first really dangerous try is fatal**." --- Eliezer **Yudkowsky** (2023)

Official Music Video and lyrics (Rhino Records)

Teach Your Children
by Crosby, Stills, Nash & Young

You, who are on the road
Must have a code you try to live by
And so become yourself
Because the past is just a goodbye

Teach your children well
Their father's hell did slowly go by
Feed them on your dreams
The one they pick's the one you'll know by

Don't you ever ask them why
If they told you, you would cry
So just look at them and sigh
And know they love you

And you, of tender years
Can't know the fears your elders grew by
Help them with your youth
They seek the truth before they can die

Teach your parents well
Their children's hell will slowly go by
And feed them on your dreams
The one they pick's the one you'll know by

Don't you ever ask them why
If they told you, you would cry
So just look at them and sigh
And know they love you

Don't you ever ask them why
If they told you, you would cry
So just look at them and sigh
And know they love you

Ooh, and know they love you
And know they love you, yeah
And know they love you
And know they love you

——---

Not-for-profit, fair use of Teach Your Children, for social benefit.
Songwriter: Graham Nash
Teach Your Children Lyrics: © Nash Notes.
Music: © Rhino
Quote sources here: The AI Safety Problem.
Thank you: Crosby, Stills, Nash & Young.
RIP David Crosby (1941 - 2023)