Sources for “You have no idea how terrified AI scientists actually are”

[ ]

#Geoffrey Hinton, 2024

Geoffrey Hinton:

> “I actually think the risk is more than 50% regarding the existential threat”

@tsarnick on X

Geoffrey Hinton:

> “they will take control and make us irrelevant”

@tsarnick on X

Jessica Coates:

> “Speaking to the media in the aftermath of his win, Mr Hinton – once dubbed the “godfather” of AI – reiterated urgent calls for companies to do more research into AI safety.”

Geoffrey Hinton warns of AI’s growing danger after Nobel Prize win

---

[ ]

#“Top 3 most-cited AI Scientists”

PauseAI on X:

> “Striking how the top three cited AI scientists share a strong opinion that this tech could kill every living thing on earth. Arguably all three made the decision to prioritise safety over money.”

[ ]

#Ilya Sutskever, 2019

Ilya Sutskever, in documentary, iHuman, 2019:

> “I think it’s pretty likely the entire surface of the earth will be covered with solar panels and data centers.”

[ ]

#Ilya Sutskever, 2023

Ilya Sutskever:

> “Is there even room for human beings in a world with smarter AIs?"

[Rogue superintelligence and merging with machines: Inside the mind of OpenAI’s chief scientist]

Ily Sutskever:

> “A good analogy would be the way humans treat animals - when the time comes to build a highway between two cities, we are not asking the animals for permission."

Ilya: the AI scientist shaping the world@8.26, the Guardian

[ ]

#Yoshua Bengio, 2024

Yoshua Bengio:

> “We're playing dice with humanity's future."

35 - Yoshua Bengio on Why AI Labs are “Playing Dice with Humanity’s Future”

[ ]

#Geoffrey Hinton, 2023

Scott Pelley:

> “Does humanity know what it's doing?”

Geoffrey Hinton:

> “We're moving into a period when, for the first time ever, we may have things more intelligent than us. I think in five years' time, they may reason better than we do.”
>
> “And normally, the first time you deal with something totally novel, you get it wrong. We can't afford to get it wrong with these things.”

Scott Pelley:

> 'Why not?'

Geoffrey Hinton:

> “Because they might take over."

"Godfather of Artificial Intelligence" Geoffrey Hinton on the promise, risks of advanced AI

[ ]

#Dario Amodei, 2023

> “Dario Amodei, the chief executive of the A.I. company Anthropic, puts his between 10 and 25 percent.”

Silicon Valley Confronts a Grim New A.I. Metric

Dario Amodei:

> "I would emphasize is, you know I I don't think we have a lot of time […] whatever we do it has to happen fast"

@1:56:16 Senate Testimony

[ ]

#Dario Amodei, 2023

Dario Amodei:

> "When I think of why am I scared [...] I think the thing that's really hard to argue with is like, there will be powerful models; they will be agentic; [...] If such a model wanted to wreak havoc and destroy humanity or whatever, I think we have basically no ability to stop it"

Dario Amodei (Anthropic CEO) - Scaling, Alignment, & AI Progress

---

[ ]

#Sam Altman, 2015

Sam Altman:

> "AI will most likely lead to the end of the world, but in the meantime there will be great companies"

Sam Altman Investing in 'AI Safety Research'

[ ]

#Sam Altman, 2023

Sam Altman:

> "Development of superhuman machine intelligence is probably the greatest threat to the continued existence of humanity."

Washington Post: 'King of the cannibals': How Sam Altman took over Silicon Valley

---

[ ]

#Sam Altman, 2023

Sam Altman:

> "The bad case—and I think this is important to say—is, like, lights-out for all of us,"

Fortune: Sam Altman, the man behind ChatGPT, is increasingly alarmed about what he unleashed. Here are 15 quotes charting his descent into sleepless panic

---

[ ]

#Elon Musk, 2018

Elon Musk:

> "And mark my words, AI is far more dangerous than nukes"

Elon Musk Answers Your Questions! | SXSW 2018@38:11

[ ]

#Elon Musk, 2023

Elon Musk:

> “xAI is Elon's new AGI project. His probability of an existential catastrophe is around 20-30%.”

FLI Podcast

[ ]

#Definition by New York Times, 2019

- As defined by the New York Times, superintelligence means an AI that surpasses human intelligence in every field - from scientific creativity to social skills.

Paraphrase from We Shouldn’t be Scared by ‘Superintelligent A.I.’

[ ]

#Sam Altman, 2024

> “This may turn out to be the most consequential fact about all of history so far. It is possible that we will have superintelligence in a few thousand days (!); it may take longer, but I’m confident we’ll get there.”

The Intelligence Age

#Stephen Nellis, 2024

> “Chief Executive Jensen Huang on Friday said that artificial general intelligence could - by some definitions - arrive in as little as five years.”

Nvidia CEO says AI could pass human tests in five years

#Leopold Aschenbrenner, 2024

> **"Complementarities/long tail:** A classic lesson from economics (cf Baumol's growth disease) is that if you can automate, say, 70% of something, you get some gains but quickly the remaining 30% become your bottleneck. For anything that falls short of full automation—say, really good copilots—human AI researchers would remain a major bottleneck, making the overall increase in the rate of algorithmic progress relatively small. Moreover, there's likely some long tail of capabilities required for automating AI research—the last 10% of the job of an AI researcher might be particularly hard to automate. This could soften takeoff some, though my best guess is that this only delays things by a couple years. Perhaps 2026/27-models speed are the proto-automated-researcher, it takes another year or two for some final unhobbling, a somewhat better model, inference speedups, and working out kinks to get to full automation, and finally by 2028 we get the 10x acceleration (and superintelligence by the end of the decade)."

From AGI to Superintelligence: the Intelligence Explosion

[ ]
#Luke Lango, 2024

> "Each week, the TrackingAI project provides IQ quizzes to a variety of AI models. Most models out there right now – like ChatGPT 4, Llama 3.1, Claude, Grok, and Bing Copilot – score around 80 to 90 on those IQ tests."

OpenAI Races Toward AGI with its New Breakthrough Model

[ ]

#Geoffrey Hinton, 2024

Geoffrey Hinton:

> "People who say neural networks are over-hyped have been wrong every time and will continue to be wrong as AI keeps improving."

@tsarnick on X

[ ]

#Yann Lecun. 2024

Yann Lecun:

> "They can't learn to clear out the dinner table"

Yann Lecun: Meta AI, Open Source, Limits of LLMs, AGI & the Future of AI | Lex Fridman …
@9:33

[ ]

#Carl Franzen, 2024

Carl Franzen:

> "Figure co-founder and CEO Brett Adcock took to his account on the social platform X to post a video demo of a Figure full-sized humanoid robot, the Figure 01 (pronounced "Figure One"), demonstrating its capabilities to interact with a nearby human and its environment, showing the robot following the person's orders, locating and handing them an object (an apple, in this case), describing what it's doing and conversing with the person (albeit with slightly delayed reaction time from what we would expect in a typical human-to-human conversation), and identifying, planning and carrying out helpful tasks on its own (in this case, picking up trash and putting dishes into a drying rack)."

OpenAI powers a robot that can hand people food, pick up trash, put away dishes, and more

---

[ ]

#Yann Lecun, 2024

Yann Lecun:

> "I mean, when I say very far, it's... several years."

@liron on X

[ ]

#Word on the Street, 2024

Word on the Street:

> “In a recent conversation, NVIDIA CEO Jensen Huang elaborated on the potential of AI data centers, suggesting that they could expand to accommodate millions of chips without being halted by any physical laws. Huang emphasized that AI software could be deployed across multiple data centers with consistent performance improvements and significant reductions in energy consumption. He referred to this rapid advancement as a "super Moore’s Law" trajectory, where performance might double or triple annually while energy demands could shrink by two to three times each year.”

NVIDIA CEO Jensen Huang Unveils Vision for AI Data Centers Breaking 'Super Moore's Law' Barriers

---

[ ]

#Dwarkesh Patel, 2023

Dwarkesh Patel:

> “You can do mental gymnastics about compute and data bottlenecks and the true nature of intelligence and the brittleness of benchmarks. Or you can just look at the fucking line.”

@AISafetyMemes on X

[ ]

#Ilya Sutskever, 2023

Ilya Sutskever:

> "One doesn't bet against deep learning. Somehow, every time you run into an obstacle, within six months researchers find a way around it."

Rogue superintelligence and merging with machines: Inside the mind of OpenAI's chief scientist

---

[ ]

#Cameron R. Jones, 2024

Cameron R. Jones:

> "We evaluated 3 systems (ELIZA, GPT-3.5 and GPT-4) in a randomized, controlled, and preregistered Turing test. Human participants had a 5 minute conversation with either a human or an AI, and judged whether or not they thought their interlocutor was human. GPT-4 was judged to be a human 54% of the time, outperforming ELIZA (22%) but lagging behind actual humans (67%). The results provide the first robust empirical demonstration that any artificial system passes an interactive 2-player Turing test."

People cannot distinguish GPT-4 from a human in a Turing test

---

[ ]

- **Now AIs are solving PhD-level math problems and beating top professionals at their own craft.**

@nim_chimpsky_ on X

> "o3 can solve 25% of research level mathematics questions designed by experts out of the box"

@polynoamial on X

> "We announced @OpenAI o1 just 3 months ago. Today, we announced o3. We have every reason to believe this trajectory will continue."

[@AISafetyMemes on X]

> The director of Taxi Driver is having an existential crisis about AI

[ ]

#Mustafa Suleyman, 2024

Mustafa Suleyman:

> "AI is a new digital species."
>
> "To avoid existential risk, we should avoid:
>
> 1) Autonomy
> 2) Recursive self-improvement
> 3) Self-replication
>
> We have a good 5 to 10 years before we'll have to confront this."

@FutureJurvetson on X

[ ]

#Sujith Mathew Iype, 2018

Sujith Mathew Iype:

> “One of the biggest advantages a computer system has over a human is the speed of processing. A 1GHz processor can perform a single operation in 1 nanosecond. Even assuming that it takes about 40 CPU cycles to transfer the data from memory before processing, a computer can do a single operation including data fetch in about 40 nanoseconds. Compare this to the human neuron which collects inputs from a synapse, processes it and transfers it to the next neuron in about 5 milliseconds. This would mean that a computer system is 125,000 times faster than the human neuron.”

Human Brain vs Artificial Intelligence Systems

[ ]

#Yuval Harari, 2023

Yuval Harari:

> "I would take it very seriously. When I hear this as a historian, this is the end of human history. Not the end of history, the end of human dominated history. History will continue with somebody else in control, because what we just heard is basically Mustafa telling us that in 5 years, there'll be a technology that can make decisions independently, and that can create new ideas independently.
>
> This is the first time in history we confronted something like this. With every previous technology in history, from a stone knife to nuclear bombs - [the technology itself] could not make decisions.
>
> The decision to drop the bomb on Hiroshima was not made by the atom bomb, it was made by President Truman.
>
> Every previous technology in history could only replicate our ideas, like radio, or the printing press. It could make copies and disseminate the music or the poems or the novels that some human wrote.
>
> Now, we have a technology that can create completely new ideas, and it can do it at a scale far beyond what humans are capable of."

AI and our future with Yuval Noah Harari and Mustafa Suleyman

---

[ ]

#Grace et al., 2024:

Grace et al.:

	Statistics	2022 Result	2023 Result
What probability do you put on **future AI advances** causing human extinction or similarly permanent and severe disempowerment of the human species?	N; Mean (SD); Median (IQR)	149; 15.7% (22.1%); 5% (19%)	1321; 16.2% (23%); 5% (19%)
What probability do you put on **human inability to control future advanced AI systems** causing human extinction or similarly permanent and severe disempowerment of the human species?	N; Mean (SD); Median (IQR)	162; 20.5% (26.2%); 10% (29%)	661; 19.4% (26%); 10% (29%)
What probability do you put on **future AI advances** causing human extinction or similarly permanent and severe disempowerment of the human species **within the next 100 years**?[4]	N; Mean (SD); Median (IQR)	Not asked	655; 14.4% (22.2%); 5% (19.9%)

Table 2: Respondents' estimates in 2022 and 2023 for the probability that AI causes human extinction. For each of the two questions that were asked in both years, the results are very similar.

Thousands Of AI Authors On The Future Of AI

[ ]

#Center for Safety of AI, 2024

Center for Safety of AI:

> "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."

Statement on AI Risk Press Coverage

[ ]

#Anders Sandberg, 2023

Andrers Sanberg:

> “Seriously, can you name any other business doing this? The oil business: "Yes, we totally are *wrecking* the climate! Want to buy shares?" Biotech: "Oh, we can totally make doomsday bugs. Please invest." Nuclear: "Nobody can tell how bad an accident will be. Wanna buy?"

@anderssandberg on X

[ ]

#Vitalik Buterin, 2023

Vitalik Buterin:

> “A survey of machine learning researchers from 2022 showed that on average, researchers think that there is a 5-10% chance that AI will literally kill us all: about the same probability as the statistically expected chance that you will die of non-biological causes like injuries.”

My techno-optimism

[ ]

#Stuart Russell, 2024

Stuart Russell:

> “Geoff Hinton is in the process of tidying up his affairs… he believes that we maybe have 4 years left."

@AISafetyMemes on X

[ ]

#Emmett Shear, 2023

Emmett Shear:

> "..I would say like it's between like 5-50."

Emmett Shear Reacts To AI Doom Theories @7:51

[ ]

#Jan Leike, 2023

Jan Leike:

> ...it's more than 10% less than 90%.

OpenAI's huge push to make superintelligence safe | Jan Leike @1:16:01

[ ]

#Secure Future AI, 2023

Secure Future AI,:

> In the AI field, the term “p(doom)” means the probability that advanced AI results in human extinction. Sometimes the field also calls this “xrisk” which is short for "existential risk." Here are estimates of p(doom) from leaders of most or all of the leading AI companies.
>
> Dario Amodei (CEO, Anthropic, the AI company with the 2nd most powerful models): 10-25%
> Jan Leike (Head of AI safety, at OpenAI which has the 1st most powerful models): 10-90%
> Geoffrey Hinton (Godfather of AI): 10%
> Paul Christiano (Former Head of AI safety at OpenAI, inventor of RLHF, considered the top AI safety researcher in the world): 10-20%
> Lina Khan (FTC Chair): 15%
> Elon Musk: 20-30%
> Average American: 21%
> Average AI engineer (Oct 2023): ~40%
> Average AI safety researcher: 30%
>
> CEO’s and head AI scientists of all three top leading AI firms, have signed this public letter warning of extinction risks:
>
> “Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.”

[Expert Estimates on AI Human Extinction Risks]

[ ]

#Yoshua Bengio, 2023

Yoshua Bengio:

> “Even if it was 0.1 per cent [chance of doom], I would be worried enough to say I’m going to devote the rest of my life to trying to prevent that from happening,”

[A ‘godfather of AI’ gives his p(doom)]

[ ]

#Tim Urban, 2024

Paraphrased, Tim Urban:

> So when you're building a God, if God V1 is buggy, you're in big trouble. Because what are you gonna say? Oh, okay, we want to go iterate. We want, let's get some beta testers. Let's change it. God says, no way. You changing me is going to actually detract from my goal.
>
> In fact, the fact that all these people are trying to change me, I might need to get rid of people because they seem to be in the way of my goal.
>
> So you can't build God V2. There is only God V1 — and that is so different than, you know, people who build software. They don't think that way. They think, yeah, you know, let's get it out there. Let's test it. Let's throw it against the wall and see if it sticks.
>
> You can't think that way. You have to get V1 right. Which is so hard. You know that not much software V1 is good.
>
> We're also used to the concept that if you build something that is somehow doing damage — pull the plug. Right? Shut it down.
>
> You can't pull the plug on a God. The God has figured out how to get energy in ways that we don't even understand. God is sapping energy from dark matter that we don't even know is here, you know, through the channels that we don't even know exist. Right?
>
> There's just — there's no such thing as, 'Oh, they pulled the plug.'
>
> So imagine there's a thousand of these, and one of them is just wacky and it starts to go unintended — develops this really destructive personality. And alone, it doesn't matter what the other 999 are doing — it can wipe us all out.

Awakening the Machine: Tim Urban @10:21

[ ]

#Shane Legg, 2025

Shane Legg:

> Google's Chief AGI Scientist Shane Legg: 50% chance of AGI within 3 years, then 5-50% chance of extinction ONE YEAR LATER

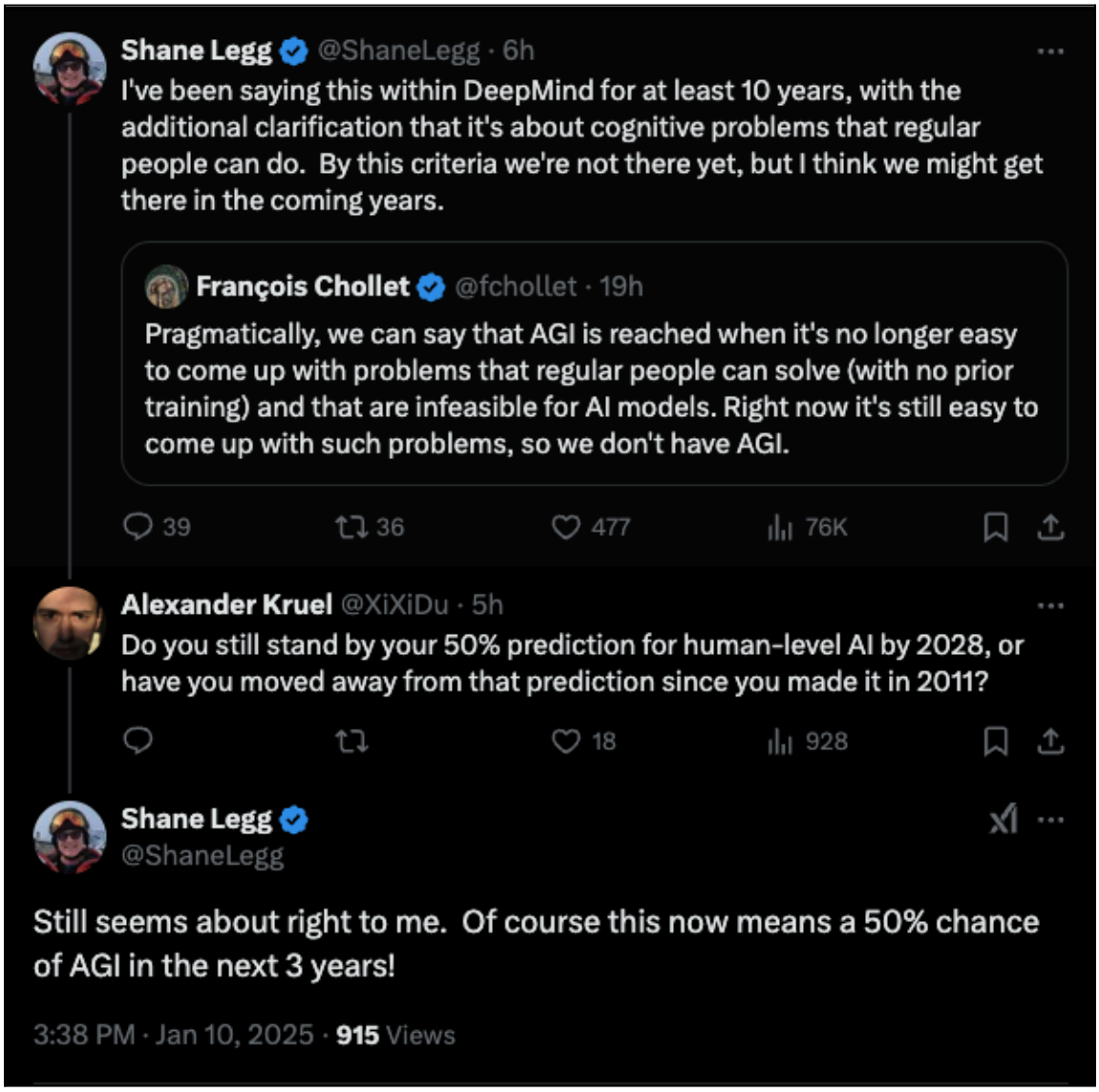
[@AISafetyMemes on X]

[ ]

#ai_ctril, 2023

Shane Legg:

> “Google DeepMind's Chief AGI Scientist says there's a 50% chance that AGI will be built in the next 3 years. This was in reference to a prediction he made back in 2011. He also thought there was a 5 to 50% chance of human extinction within a year of human-level AI being built!”



> **The Interview**:
>
> **Q1:** *Assuming no global catastrophe halts progress, by what year would you assign a 10%/50%/90% chance of the development of human-level machine intelligence?*
>
> *Explanatory remark to Q1:*
>
> *P(human-level AI by (year) | no wars ∧ no disasters ∧ beneficially political and economic development) = 10%/50%/90%*
>
> > **Shane Legg:** 2018, 2028, 2050
>
> **Q2:** *What probability do you assign to the possibility of negative/extremely negative consequences as a result of badly done AI?*
>
> *Explanatory remark to Q2:*
>
> *P(negative consequences | badly done AI) = ?*
> *P(extremely negative consequences | badly done AI) = ?*
>
> *(Where 'negative' = human extinction; 'extremely negative' = humans suffer;)*
>
> > **Shane Legg:** Depends a lot on how you define things. Eventually, I think human extinction will probably occur, and technology will likely play a part in this. But there's a big difference between this being within a year of something like human level AI, and within a million years. As for the former meaning...I don't know. Maybe 5%, maybe 50%. I don't think anybody has a good estimate of this.

[@ai_ctrl on X]

[ ]

#Dario Amodei, 2023

Dario Amodei:

> "A straightforward extrapolation of today's AI systems to those we expect to see in 2-3 years suggests … AI systems will be able to fill in all the missing pieces, enabling many more actors to carry out large-scale biological attacks."

Senate Judiciary Committee holds hearing on AI oversight and regulation — 07/25/23 @21:17

[ ]

#Sam Altman, 2023

Sam Altman:

> "An A.I. that could design novel biological pathogens."

Fox News: OpenAI CEO Sam Altman reveals what he thinks is 'scary' about AI

[ ]

#Nathan A. Paxton and Jaime M. Yassif, 2024

Nathan A. Paxton and Jaime M. Yassif:

> "Over the course of more than a century, there has been clear evidence that countries have developed bioweapons or created bioweapons programs, but it has been exceedingly difficult to identify known or probable bioweapons developers with certainty. The most comprehensive, unclassified, peer reviewed study concluded that since 1915, 44 countries have been suspected of pursuing bioweapons. Of these 44, it is likely that 18 never had a bioweapons program, three only considered developing such a program, and 23 had or likely had a bioweapons program at some point. Even though most of the latter countries abandoned their programs by the time they signed on to the BWC, some BWC States-Parties continue to suspect one other of developing bioweapons or at least bioweapons-relevant capabilities."

Disincentivizing Bioweapons: Theory and Policy Approaches

[ ]

#Geoffrey Hinton, 2023

Geoffrey Hinton:

> These AI models have far fewer neural connections than humans do, but they manage to know a thousand times as much as a human, Hinton said.
>
> [...]
>
> Hinton said he is worried about the increasingly powerful machines' ability to outperform humans in ways that are not in the best interest of humanity, and the likely inability to limit AI development.

[Why neural net pioneer Geoffrey Hinton is sounding the alarm on AI]

Geoffrey Hinton:

> "AI trained by good people will have a bias towards good; AI trained by bad people such as Putin or somebody like that will have a bias towards bad. We know they're going to make battle robots. They're not going to necessarily be good since their primary purpose is going to be to kill people."

Geoffrey Hinton Collision 2023 Speech (with subtitles) @8:30

Geoffrey Hinton:

> "The research will happen in China if it doesn't happen here because there's so many benefits of these things, such huge increases in productivity."

['The godfather of AI' sounds alarm about potential dangers of AI]

[ ]

#Geoffrey Hinton, 2023

Geoffrey Hinton:

> "They will be able to manipulate people... they'll have learned from all the novels ever written, all the books by Machiavelli…They might take over."

"Godfather of Artificial Intelligence" Geoffrey Hinton on the promise, risks of advanced AI

[ ]

#Nick Bostrom, 2012

Nick Bostrom:

> "The OAI is already boxed (placed in a single physical substrate) by design (see section 4.1.1). One can further place it within a Faraday cage, surround the cage with high explosives, and bury the whole set-up in a concrete bunker. There are no real limits to the number of physical security measures that can be added by wary or paranoid developers, and their effects are well understood."

Thinking Inside the Box: Controlling and Using an Oracle AI by Stuart Armstrong, Anders Sandberg, and Nick Bostrom

---

[ ]

#Mark Zuckerberg, 2024

Mark Zuckerberg:

> "We need to control our own destiny and not get locked into a closed vendor. Many organizations don't want to depend on models they cannot run and control themselves. They don't want closed model providers to be able to change their model, alter their terms of use, or even stop serving them entirely. They also don't want to get locked into a single cloud that has exclusive rights to a model. Open source enables a broad ecosystem of companies with compatible toolchains that you can move between easily."

Open Source AI is the Path Forward

[ ]

#Will Knight, 2024

Will Knight:

> "Anduril, which has made software a central part of its products, appears to be trying to lay the groundwork for swarm warfare. The company's Lattice platform can be used to connect and coordinate different sensors and weapons systems, providing an integrated visualization of a battlefield. Anduril now markets the platform's ability to control a swarm of drones and has collaborated with another defense startup, Epirus, to offer a counter-drone system that uses powerful microwaves to neutralize swarms of drones."

Anduril Is Building Out the Pentagon's Dream of Deadly Drone Swarms

---

[ ]

#Pope Francis, 2023

Reuters:

> Pope Francis has called for a legally binding international treaty to regulate artificial intelligence, saying algorithms must not be allowed to replace human values and warning of a "technological dictatorship" threatening human existence.
>
> [...]
>
> At a news conference presenting the message, Cardinal Michael Czerny, head of the Vatican's human development office, said the 86-year-old pope was "no luddite," a term referring to someone opposed to new technology.
>
> He said the pope appreciates technological and scientific progress that serves humanity but that Francis was particularly concerned about AI because it is "perhaps the highest-stake gamble of our future".

Reuters: Pope Francis calls for binding global treaty to regulate AI

[ ]

#Jake Sullivan, 2025

Jake Sullivan:

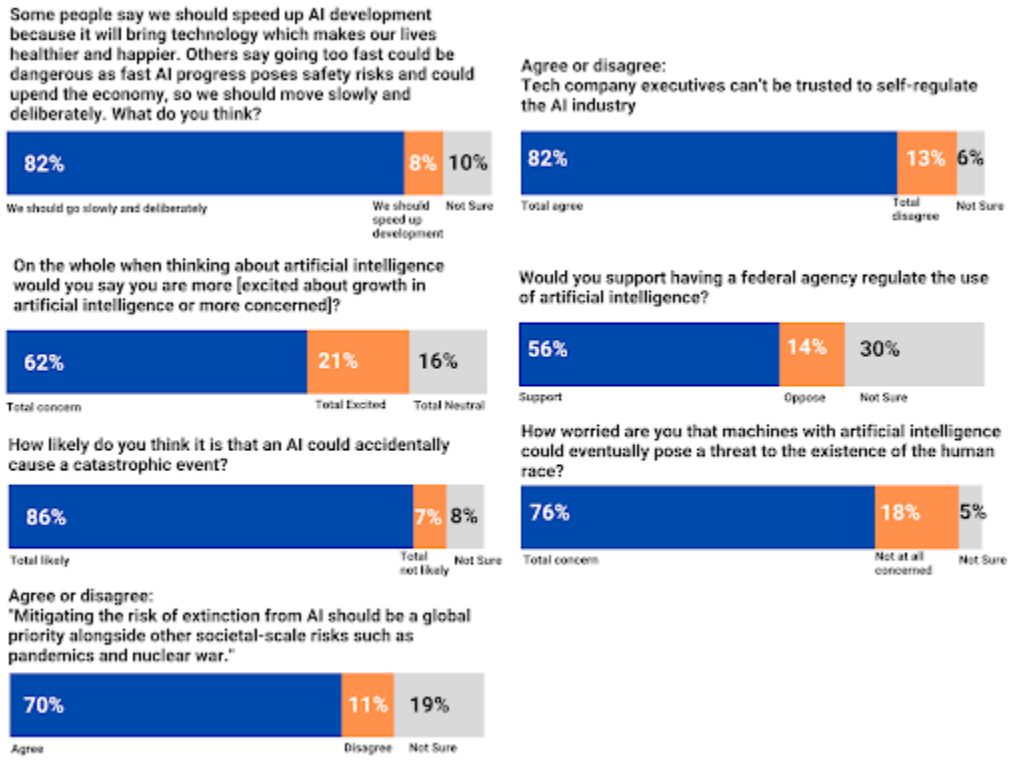
> National Security Advisor Jake Sullivan told Axios that the U.S. is in a global race to lead in artificial intelligence and that AI could determine the future of American power.
>
> 'This is one of those moments that is going to define the trajectory of U.S. power and influence in the years to come,' Sullivan said.
>
> 'We cannot afford to fall behind. Whoever leads in AI will shape the future of global power. AI is as consequential as the advent of nuclear weapons.'
>
> 'It is a transformative and potentially destabilizing force that requires careful guidance. We must shape its trajectory before it shapes us.'

Behind the Curtain: A chilling, "catastrophic" warning

[ ]

#YouGov, 2023

Artificial Intelligence Policy Institute (AIPI):

> 82% want to slow down AI development, while only 8% want to speed up
>
> 
>
>
> [Poll Shows Overwhelming Concern About Risks From AI as New Institute Launches to Understand Public Opinion and Advocate for Responsible AI Policies]
