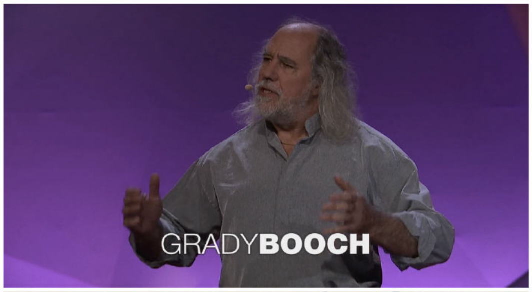

51 Top TED Talks

ARTIFICIAL INTELLIGENCE EDITION

COLLECTED BY JEREMY CONNELL-WAITE
MAY 2024

POSTED TO LINKEDIN ON 1ST MAY 2024
https://www.linkedin.com/feed/update/urn:li:activity:7191482848825020417/

I used this collection of talks as a data set when I was using the AI to analyse 50 of the top TED Talks on Artificial Intelligence to try and understand what made them so good. Using an assistant I had built on IBM's watsonx platform I create a series of elaborate prompts using Meta's large language model Llama-2.

Here's what I discovered:

⏰ The average length: 11 mins 50.

👩🏻‍💻 34 of the top 50 talks are male.

📖 They all start with a story.

❻ They made the audience laugh every 6 minutes (Important since that's alleged to be the average attention span of a live business audience).

🤷🏻‍♂️ "HOW" was used one-third of the time as the title for the talk. (Great tech stories focus on HOW – not what or why).

⏱️ Average speaker talks at 157 words/min. (Many tech speakers are 175-200. The best speakers are usually around 125).

👏🏻 Half of the speakers were stopped at least once for applause. (There's a great goal to aim for!)

📝 Analysis of all 93,029 words of the transcripts (9 hrs 51) shows a Gunning Fog readability score of 10.72. (Many corproate presentations are bad because they score 17+).

🖍️ The transcripts had a Flesh Kincaid reading ease of 61.54. That's 10th-12th grade in US (year 11/12 in UK). Calculating reading scores is really helpful. Below 50 (college students) are generally bad for tech presentations.

🤯 "Lexical density" measures the number of unique words used, which can indicate the complexity of the language. These talks averaged 46.94% (It's a strong KPI most people don't use. Most tech talks drown in complexity at 70%+)

🖋️ An easier way to describe the above? The average word was only 1.5 syllables long. Impressive given the complexity of topics & themes.

💙 "HUMAN" was one of the most used words. 513 times. (Av. 10X per talk).

⚖️ But "ETHICS" or "MORALS" were only mentioned 12 times. Read into that what you will.

😳 Overall sentiment reveals the mindset of the audiences: “Concern and alarm regarding the advancement of AI particularly in terms of job displacement and the importance of developing emotional intelligence to complement technological progress."

🥳 Presenters added LEVITY to contrast darker themes “expressing excitement towards utilising AI for enhancing education and promoting cross-cultural understanding.”

Intelligent use of rhetorical devices. 3 simple examples:

📖 Oxymoron: "humanistic AI". This term combines the words "humanistic" (having qualities associated with humans), and "AI" creating an oxymoron because AI is generally considered non-human.

📆 Anadiplosis: "Now, today I'm happy to see that the idea of an intelligent assistant is mainstream." - This sentence begins with the word "now," connecting the current state to the previous context.

🌊 Hyperbole: "a tsunami of misinformation" - This expression exaggerates the potential effects of misinformation created by AI.

Was all that enough to max out your storytelling geek quotient for today! 😂

This is all really interesting stuff that will HELP YOU get better at communicating the more you dig into it, so I put this PDF together with URLs for all 51 talks* so that you can check them out yourself. It would be 10 hours well spent!

Have fun!
Jeremy

www.BetterStories.org

*I included Mustafa Suleyman’s transcript here as well since it’s so good but it wasn’t included in my original data set.

What is the Meta's large language model Llama-3?

https://theconversation.com/what-to-expect-from-the-next-generation-of-chatbots-openais-gpt-5-and-metas-llama-3-228217

Llama-3 is Meta's competitor to GPT-5. It features several improvements compared to its predecessor, Llama-2. It is a more capable model that will eventually come with 400 billion parameters compared to a maximum of 70 billion for its predecessor Llama-2. In machine learning, a parameter is a term that represents a variable in the AI system that can be adjusted during the training process, in order to improve its ability to make accurate predictions.

Llama-3 will also be multimodal, which means it is capable of processing and generating text, images and video. Therefore, it will be capable of taking an image as input to provide a detailed description of the image content. Equally, it can automatically create a new image that matches the user's prompt, or text description.

It will be able to perform tasks in languages other than English and will have a larger context window than Llama 2. A context window reflects the range of text that the LLM can process at the time the information is generated. This implies that the model will be able to handle larger chunks of text or data within a shorter period of time when it is asked to make predictions and generate responses.

Meta is planning to launch Llama-3 in several different versions to be able to work with a variety of other applications, including Google Cloud. Meta announced that more basic versions of Llama-3 will be rolled out soon, ahead of the release of the most advanced version, which is expected next summer.

The transition to this new generation of chatbots could not only revolutionise generative AI, but also mark the start of a new era in human-machine interaction that could transform industries and societies on a global scale. It will affect the way people work, learn, receive healthcare, communicate with the world and each other. It will make businesses and organisations more efficient and effective, more agile to change, and so more profitable.

Some Technical Data & Links on Llama-3 if you're interested

➡️ Trained on 15T Tokens, fine-tuned on 10M human annotated data, custom-built 24,000 GPU cluster
➡️ 8B & 70B versions released today as both Instruct and Base
➡️ Llama 3-70B best open LLM on MMLU benchmarks
➡️ Llama 3-8B outperforming Llama2-70B (9x in size) in some cases
➡️ Llama 3-400B will be impressive (multimodal capabilities, multilingual support, extended context windows, and crazy performance)
➡️ Instruct good at coding 8B with 62.2 and 70B 81.7 on Human Eval
➡️ 8k default context window (can be increased)
➡️ Commercial use allowed
➡️ Available on IBM watonsx.ai today

- **Blog:** https://huggingface.co/blog/llama3
- **Models:** https://huggingface.co/models?other=llama-2
- **Chat-Demo:** https://huggingface.co/chat/
- **IBM news release:** https://newsroom.ibm.com/Blog-IBM-Offers-Metas-Llama-3-Open-Models-on-Watsonx,-Expands-Portfolio-of-Next-Generation-Enterprise-Ready-Models
- **IBM watonsx.ai demo** by Maryam Ashoori, PhD: https://www.linkedin.com/posts/mashoori_llama3-generativeai-watsonx-activity-7186762967282036737-r9p_/

Links For All 51 Ted Talks

https://www.ted.com/talks/maurice_conti_the_incredible_inventions_of_intuitive_ai

https://www.ted.com/talks/sam_harris_can_we_build_ai_without_losing_control_over_it

https://www.ted.com/talks/nick_bostrom_what_happens_when_our_computers_get_smarter_than_we_are

https://www.ted.com/talks/kai_fu_lee_how_ai_can_save_our_humanity

https://www.ted.com/talks/dan_finkel_can_you_solve_the_rogue_ai_riddle

https://www.ted.com/talks/ray_kurzweil_get_ready_for_hybrid_thinking

https://www.ted.com/talks/greg_brockman_the_inside_story_of_chatgpt_s_astonishing_potential

https://www.ted.com/talks/imran_chaudhri_the_disappearing_computer_and_a_world_where_you_can_take_ai_everywhere

https://www.ted.com/talks/janelle_shane_the_danger_of_ai_is_weirder_than_you_think

https://www.ted.com/talks/fei_fei_li_how_we_re_teaching_computers_to_understand_pictures

https://www.ted.com/talks/sal_khan_how_ai_could_save_not_destroy_education

https://www.ted.com/talks/grady_booch_don_t_fear_superintelligent_ai

https://www.ted.com/talks/jeremy_howard_the_wonderful_and_terrifying_implications_of_computers_that_can_learn

https://www.ted.com/talks/andrew_ng_how_ai_could_empower_any_business

https://www.ted.com/talks/sylvain_duranton_how_humans_and_ai_can_work_together_to_create_better_businesses

https://www.ted.com/talks/kriti_sharma_how_to_keep_human_bias_out_of_ai

https://www.ted.com/talks/tom_gruber_how_ai_can_enhance_our_memory_work_and_social_lives

https://www.ted.com/talks/rosalind_picard_an_ai_smartwatch_that_detects_seizures

https://www.ted.com/talks/lucy_farey_jones_a_fascinating_time_capsule_of_human_feelings_toward_ai

https://www.ted.com/talks/garry_kasparov_don_t_fear_intelligent_machines_work_with_them

https://www.ted.com/talks/yejin_choi_why_ai_is_incredibly_smart_and_shockingly_stupid

https://www.ted.com/talks/genevieve_bell_6_big_ethical_questions_about_the_future_of_ai

https://www.ted.com/talks/leila_pirhaji_the_medical_potential_of_ai_and_metabolites

https://www.ted.com/talks/jim_collins_how_we_re_using_ai_to_discover_new_antibiotics

https://www.ted.com/talks/jeff_dean_ai_isn_t_as_smart_as_you_think_but_it_could_be

https://www.ted.com/talks/kevin_kelly_how_ai_can_bring_on_a_second_industrial_revolution

https://www.ted.com/talks/zeynep_tufekci_machine_intelligence_makes_human_morals_more_important

https://www.ted.com/talks/shervin_khodabandeh_why_people_and_ai_make_good_business_partners

https://www.ted.com/talks/tim_leberecht_4_ways_to_build_a_human_company_in_the_age_of_machines

https://www.ted.com/talks/pratik_shah_how_ai_is_making_it_easier_to_diagnose_disease

https://www.ted.com/talks/sebastian_thrun_and_chris_anderson_what_ai_is_and_isn_t/

https://www.ted.com/talks/pierre_barreau_how_ai_could_compose_a_personalized_soundtrack_to_your_life

https://www.ted.com/talks/stuart_russell_3_principles_for_creating_safer_ai

https://www.ted.com/talks/max_tegmark_how_to_get_empowered_not_overpowered_by_ai

https://www.ted.com/talks/gary_marcus_the_urgent_risks_of_runaway_ai_and_what_to_do_about_them

https://www.ted.com/talks/ken_jennings_watson_jeopardy_and_me_the_obsolete_know_it_all?

https://www.ted.com/talks/tom_graham_the_incredible_creativity_of_deepfakes_and_the_worrying_future_of_ai

https://www.ted.com/talks/eliezer_yudkowsky_will_superintelligent_ai_end_the_world

https://www.ted.com/talks/frances_s_chance_are_insect_brains_the_secret_to_great_ai

https://www.ted.com/talks/peter_norvig_the_100_000_student_classroom

https://www.ted.com/talks/alexandr_wang_war_ai_and_the_new_global_arms_race https://www.ted.com/talks/mustafa_suleyman_what_is_an_ai_anyway

https://www.ted.com/talks/refik_anadol_how_ai_art_could_enhance_humanity_s_collective_memory

https://www.ted.com/talks/liv_boeree_the_dark_side_of_competition_in_ai

https://www.ted.com/talks/stephen_wolfram_how_to_think_computationally_about_ai_the_universe_and_everything

https://www.ted.com/talks/briana_brownell_how_does_artificial_intelligence_learn

https://www.ted.com/talks/jessica_apotheker_what_will_happen_to_marketing_in_the_age_of_ai

https://www.ted.com/talks/paul_hudson_and_lindsay_levin_leadership_in_the_age_of_ai/

https://www.ted.com/talks/gil_weinberg_can_robots_be_creative

https://www.ted.com/talks/sofia_crespo_ai_generated_creatures_that_stretch_the_boundaries_of_imagination

https://www.ted.com/talks/walter_de_brouwer_how_ai_is_learning_what_it_means_to_be_human

The incredible inventions of intuitive AI

Maurice Conti

How many of you are creatives,
designers, engineers, entrepreneurs, artists,
or maybe you just have a really big imagination?
Show of hands? (Cheers)
That's most of you.
I have some news for us creatives.
Over the course of the next 20 years,
more will change around the way we do our work
than has happened in the last 2,000.
In fact, I think we're at the dawn of a new age in human history.
Now, there have been four major historical eras defined by the way we work.
The Hunter-Gatherer Age lasted several million years.
And then the Agricultural Age lasted several thousand years.
The Industrial Age lasted a couple of centuries.
And now the Information Age has lasted just a few decades.
And now today, we're on the cusp of our next great era as a species.
Welcome to the Augmented Age.
In this new era, your natural human capabilities are going to be augmented
by computational systems that help you think,
robotic systems that help you make,
and a digital nervous system
that connects you to the world far beyond your natural senses.
Let's start with cognitive augmentation.
How many of you are augmented cyborgs?

I would actually argue that we're already augmented.
Imagine you're at a party,
and somebody asks you a question that you don't know the answer to.
If you have one of these, in a few seconds, you can know the answer.
But this is just a primitive beginning.
Even Siri is just a passive tool.
In fact, for the last three-and-a-half million years,
the tools that we've had have been completely passive.
They do exactly what we tell them and nothing more.
Our very first tool only cut where we struck it.
The chisel only carves where the artist points it.
And even our most advanced tools do nothing without our explicit direction.
In fact, to date, and this is something that frustrates me,
we've always been limited
by this need to manually push our wills into our tools --
like, manual, literally using our hands,
even with computers.
But I'm more like Scotty in "Star Trek."

I want to have a conversation with a computer.
I want to say, "Computer, let's design a car,"
and the computer shows me a car.
And I say, "No, more fast-looking, and less German,"

and bang, the computer shows me an option.

That conversation might be a little ways off,
probably less than many of us think,
but right now,
we're working on it.
Tools are making this leap from being passive to being generative.
Generative design tools use a computer and algorithms
to synthesize geometry
to come up with new designs all by themselves.
All it needs are your goals and your constraints.
I'll give you an example.
In the case of this aerial drone chassis,
all you would need to do is tell it something like,
it has four propellers,
you want it to be as lightweight as possible,
and you need it to be aerodynamically efficient.
Then what the computer does is it explores the entire solution space:
every single possibility that solves and meets your criteria --
millions of them.
It takes big computers to do this.
But it comes back to us with designs
that we, by ourselves, never could've imagined.
And the computer's coming up with this stuff all by itself --
no one ever drew anything,
and it started completely from scratch.
And by the way, it's no accident
that the drone body looks just like the pelvis of a flying squirrel.

It's because the algorithms are designed to work
the same way evolution does.
What's exciting is we're starting to see this technology
out in the real world.
We've been working with Airbus for a couple of years
on this concept plane for the future.
It's a ways out still.
But just recently we used a generative-design AI
to come up with this.
This is a 3D-printed cabin partition that's been designed by a computer.
It's stronger than the original yet half the weight,
and it will be flying in the Airbus A320 later this year.
So computers can now generate;
they can come up with their own solutions to our well-defined problems.
But they're not intuitive.
They still have to start from scratch every single time,
and that's because they never learn.
Unlike Maggie.

Maggie's actually smarter than our most advanced design tools.
What do I mean by that?
If her owner picks up that leash,

Maggie knows with a fair degree of certainty
it's time to go for a walk.
And how did she learn?
Well, every time the owner picked up the leash, they went for a walk.
And Maggie did three things:
she had to pay attention,
she had to remember what happened
and she had to retain and create a pattern in her mind.
Interestingly, that's exactly what
computer scientists have been trying to get AIs to do
for the last 60 or so years.
Back in 1952,
they built this computer that could play Tic-Tac-Toe.
Big deal.
Then 45 years later, in 1997,
Deep Blue beats Kasparov at chess.
2011, Watson beats these two humans at Jeopardy,
which is much harder for a computer to play than chess is.
In fact, rather than working from predefined recipes,
Watson had to use reasoning to overcome his human opponents.
And then a couple of weeks ago,
DeepMind's AlphaGo beats the world's best human at Go,
which is the most difficult game that we have.
In fact, in Go, there are more possible moves
than there are atoms in the universe.
So in order to win,
what AlphaGo had to do was develop intuition.
And in fact, at some points, AlphaGo's programmers didn't understand
why AlphaGo was doing what it was doing.
And things are moving really fast.
I mean, consider -- in the space of a human lifetime,
computers have gone from a child's game
to what's recognized as the pinnacle of strategic thought.
What's basically happening
is computers are going from being like Spock
to being a lot more like Kirk.

Right? From pure logic to intuition.
Would you cross this bridge?
Most of you are saying, "Oh, hell no!"

And you arrived at that decision in a split second.
You just sort of knew that bridge was unsafe.
And that's exactly the kind of intuition
that our deep-learning systems are starting to develop right now.
Very soon, you'll literally be able
to show something you've made, you've designed,
to a computer,
and it will look at it and say,
"Sorry, homie, that'll never work. You have to try again."
Or you could ask it if people are going to like your next song,

or your next flavor of ice cream.
Or, much more importantly,
you could work with a computer to solve a problem
that we've never faced before.
For instance, climate change.
We're not doing a very good job on our own,
we could certainly use all the help we can get.
That's what I'm talking about,
technology amplifying our cognitive abilities
so we can imagine and design things that were simply out of our reach
as plain old un-augmented humans.
So what about making all of this crazy new stuff
that we're going to invent and design?
I think the era of human augmentation is as much about the physical world
as it is about the virtual, intellectual realm.
How will technology augment us?
In the physical world, robotic systems.
OK, there's certainly a fear
that robots are going to take jobs away from humans,
and that is true in certain sectors.
But I'm much more interested in this idea
that humans and robots working together are going to augment each other,
and start to inhabit a new space.
This is our applied research lab in San Francisco,
where one of our areas of focus is advanced robotics,
specifically, human-robot collaboration.
And this is Bishop, one of our robots.
As an experiment, we set it up
to help a person working in construction doing repetitive tasks --
tasks like cutting out holes for outlets or light switches in drywall.

So, Bishop's human partner can tell what to do in plain English
and with simple gestures,
kind of like talking to a dog,
and then Bishop executes on those instructions
with perfect precision.
We're using the human for what the human is good at:
awareness, perception and decision making.
And we're using the robot for what it's good at:
precision and repetitiveness.
Here's another cool project that Bishop worked on.
The goal of this project, which we called the HIVE,
was to prototype the experience of humans, computers and robots
all working together to solve a highly complex design problem.
The humans acted as labor.
They cruised around the construction site, they manipulated the bamboo --
which, by the way, because it's a non-isomorphic material,
is super hard for robots to deal with.
But then the robots did this fiber winding,
which was almost impossible for a human to do.
And then we had an AI that was controlling everything.

It was telling the humans what to do, telling the robots what to do
and keeping track of thousands of individual components.
What's interesting is,
building this pavilion was simply not possible
without human, robot and AI augmenting each other.
OK, I'll share one more project. This one's a little bit crazy.
We're working with Amsterdam-based artist Joris Laarman and his team at MX3D
to generatively design and robotically print
the world's first autonomously manufactured bridge.
So, Joris and an AI are designing this thing right now, as we speak,
in Amsterdam.
And when they're done, we're going to hit "Go,"
and robots will start 3D printing in stainless steel,
and then they're going to keep printing, without human intervention,
until the bridge is finished.
So, as computers are going to augment our ability
to imagine and design new stuff,
robotic systems are going to help us build and make things
that we've never been able to make before.
But what about our ability to sense and control these things?
What about a nervous system for the things that we make?
Our nervous system, the human nervous system,
tells us everything that's going on around us.
But the nervous system of the things we make is rudimentary at best.
For instance, a car doesn't tell the city's public works department
that it just hit a pothole at the corner of Broadway and Morrison.
A building doesn't tell its designers
whether or not the people inside like being there,
and the toy manufacturer doesn't know
if a toy is actually being played with --
how and where and whether or not it's any fun.
Look, I'm sure that the designers imagined this lifestyle for Barbie
when they designed her.

But what if it turns out that Barbie's actually really lonely?

If the designers had known
what was really happening in the real world
with their designs -- the road, the building, Barbie --
they could've used that knowledge to create an experience
that was better for the user.
What's missing is a nervous system
connecting us to all of the things that we design, make and use.
What if all of you had that kind of information flowing to you
from the things you create in the real world?
With all of the stuff we make,
we spend a tremendous amount of money and energy --
in fact, last year, about two trillion dollars --
convincing people to buy the things we've made.
But if you had this connection to the things that you design and create
after they're out in the real world,

after they've been sold or launched or whatever,
we could actually change that,
and go from making people want our stuff,
to just making stuff that people want in the first place.
The good news is, we're working on digital nervous systems
that connect us to the things we design.
We're working on one project
with a couple of guys down in Los Angeles called the Bandito Brothers
and their team.
And one of the things these guys do is build insane cars
that do absolutely insane things.
These guys are crazy --

in the best way.
And what we're doing with them
is taking a traditional race-car chassis
and giving it a nervous system.
So we instrumented it with dozens of sensors,
put a world-class driver behind the wheel,
took it out to the desert and drove the hell out of it for a week.
And the car's nervous system captured everything
that was happening to the car.
We captured four billion data points;
all of the forces that it was subjected to.
And then we did something crazy.
We took all of that data,
and plugged it into a generative-design AI we call "Dreamcatcher."
So what do get when you give a design tool a nervous system,
and you ask it to build you the ultimate car chassis?
You get this.
This is something that a human could never have designed.
Except a human did design this,
but it was a human that was augmented by a generative-design AI,
a digital nervous system
and robots that can actually fabricate something like this.
So if this is the future, the Augmented Age,
and we're going to be augmented cognitively, physically and perceptually,
what will that look like?
What is this wonderland going to be like?
I think we're going to see a world
where we're moving from things that are fabricated
to things that are farmed.
Where we're moving from things that are constructed
to that which is grown.
We're going to move from being isolated
to being connected.
And we'll move away from extraction
to embrace aggregation.
I also think we'll shift from craving obedience from our things
to valuing autonomy.
Thanks to our augmented capabilities,

our world is going to change dramatically.
We're going to have a world with more variety, more connectedness,
more dynamism, more complexity,
more adaptability and, of course,
more beauty.
The shape of things to come
will be unlike anything we've ever seen before.
Why?
Because what will be shaping those things is this new partnership
between technology, nature and humanity.
That, to me, is a future well worth looking forward to.
Thank you all so much.

Can we build AI without losing control over it?

Sam Harris

I'm going to talk about a failure of intuition
that many of us suffer from.
It's really a failure to detect a certain kind of danger.
I'm going to describe a scenario
that I think is both terrifying
and likely to occur,
and that's not a good combination,
as it turns out.
And yet rather than be scared, most of you will feel
that what I'm talking about is kind of cool.
I'm going to describe how the gains we make
in artificial intelligence
could ultimately destroy us.
And in fact, I think it's very difficult to see how they won't destroy us
or inspire us to destroy ourselves.
And yet if you're anything like me,
you'll find that it's fun to think about these things.
And that response is part of the problem.
OK? That response should worry you.
And if I were to convince you in this talk
that we were likely to suffer a global famine,
either because of climate change or some other catastrophe,
and that your grandchildren, or their grandchildren,
are very likely to live like this,
you wouldn't think,
"Interesting.
I like this TED Talk."
Famine isn't fun.
Death by science fiction, on the other hand, is fun,
and one of the things that worries me most about the development of AI at this point
is that we seem unable to marshal an appropriate emotional response
to the dangers that lie ahead.
I am unable to marshal this response, and I'm giving this talk.
It's as though we stand before two doors.
Behind door number one,
we stop making progress in building intelligent machines.
Our computer hardware and software just stops getting better for some reason.
Now take a moment to consider why this might happen.
I mean, given how valuable intelligence and automation are,
we will continue to improve our technology if we are at all able to.
What could stop us from doing this?
A full-scale nuclear war?
A global pandemic?
An asteroid impact?
Justin Bieber becoming president of the United States?

The point is, something would have to destroy civilization as we know it.
You have to imagine how bad it would have to be

to prevent us from making improvements in our technology
permanently,
generation after generation.
Almost by definition, this is the worst thing
that's ever happened in human history.
So the only alternative,
and this is what lies behind door number two,
is that we continue to improve our intelligent machines
year after year after year.
At a certain point, we will build machines that are smarter than we are,
and once we have machines that are smarter than we are,
they will begin to improve themselves.
And then we risk what the mathematician IJ Good called
an "intelligence explosion,"
that the process could get away from us.
Now, this is often caricatured, as I have here,
as a fear that armies of malicious robots
will attack us.
But that isn't the most likely scenario.
It's not that our machines will become spontaneously malevolent.
The concern is really that we will build machines
that are so much more competent than we are
that the slightest divergence between their goals and our own
could destroy us.
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,
could treat us with similar disregard.
Now, I suspect this seems far-fetched to many of you.
I bet there are those of you who doubt that superintelligent AI is possible,
much less inevitable.
But then you must find something wrong with one of the following assumptions.
And there are only three of them.
Intelligence is a matter of information processing in physical systems.
Actually, this is a little bit more than an assumption.
We have already built narrow intelligence into our machines,
and many of these machines perform
at a level of superhuman intelligence already.
And we know that mere matter
can give rise to what is called "general intelligence,"
an ability to think flexibly across multiple domains,
because our brains have managed it. Right?
I mean, there's just atoms in here,

and as long as we continue to build systems of atoms
that display more and more intelligent behavior,
we will eventually, unless we are interrupted,
we will eventually build general intelligence
into our machines.
It's crucial to realize that the rate of progress doesn't matter,
because any progress is enough to get us into the end zone.
We don't need Moore's law to continue. We don't need exponential progress.
We just need to keep going.
The second assumption is that we will keep going.
We will continue to improve our intelligent machines.
And given the value of intelligence --
I mean, intelligence is either the source of everything we value
or we need it to safeguard everything we value.
It is our most valuable resource.
So we want to do this.
We have problems that we desperately need to solve.
We want to cure diseases like Alzheimer's and cancer.
We want to understand economic systems. We want to improve our climate science.
So we will do this, if we can.
The train is already out of the station, and there's no brake to pull.
Finally, we don't stand on a peak of intelligence,
or anywhere near it, likely.
And this really is the crucial insight.
This is what makes our situation so precarious,
and this is what makes our intuitions about risk so unreliable.
Now, just consider the smartest person who has ever lived.
On almost everyone's shortlist here is John von Neumann.
I mean, the impression that von Neumann made on the people around him,
and this included the greatest mathematicians and physicists of his time,
is fairly well-documented.
If only half the stories about him are half true,
there's no question
he's one of the smartest people who has ever lived.
So consider the spectrum of intelligence.
Here we have John von Neumann.
And then we have you and me.
And then we have a chicken.

Sorry, a chicken.

There's no reason for me to make this talk more depressing than it needs to be.

It seems overwhelmingly likely, however, that the spectrum of intelligence
extends much further than we currently conceive,
and if we build machines that are more intelligent than we are,
they will very likely explore this spectrum
in ways that we can't imagine,
and exceed us in ways that we can't imagine.
And it's important to recognize that this is true by virtue of speed alone.
Right? So imagine if we just built a superintelligent AI

that was no smarter than your average team of researchers
at Stanford or MIT.
Well, electronic circuits function about a million times faster
than biochemical ones,
so this machine should think about a million times faster
than the minds that built it.
So you set it running for a week,
and it will perform 20,000 years of human-level intellectual work,
week after week after week.
How could we even understand, much less constrain,
a mind making this sort of progress?
The other thing that's worrying, frankly,
is that, imagine the best case scenario.
So imagine we hit upon a design of superintelligent AI
that has no safety concerns.
We have the perfect design the first time around.
It's as though we've been handed an oracle
that behaves exactly as intended.
Well, this machine would be the perfect labor-saving device.
It can design the machine that can build the machine
that can do any physical work,
powered by sunlight,
more or less for the cost of raw materials.
So we're talking about the end of human drudgery.
We're also talking about the end of most intellectual work.
So what would apes like ourselves do in this circumstance?
Well, we'd be free to play Frisbee and give each other massages.
Add some LSD and some questionable wardrobe choices,
and the whole world could be like Burning Man.

Now, that might sound pretty good,
but ask yourself what would happen
under our current economic and political order?
It seems likely that we would witness
a level of wealth inequality and unemployment
that we have never seen before.
Absent a willingness to immediately put this new wealth
to the service of all humanity,
a few trillionaires could grace the covers of our business magazines
while the rest of the world would be free to starve.
And what would the Russians or the Chinese do
if they heard that some company in Silicon Valley
was about to deploy a superintelligent AI?
This machine would be capable of waging war,
whether terrestrial or cyber,
with unprecedented power.
This is a winner-take-all scenario.
To be six months ahead of the competition here
is to be 500,000 years ahead,
at a minimum.
So it seems that even mere rumors of this kind of breakthrough

could cause our species to go berserk.
Now, one of the most frightening things,
in my view, at this moment,
are the kinds of things that AI researchers say
when they want to be reassuring.
And the most common reason we're told not to worry is time.
This is all a long way off, don't you know.
This is probably 50 or 100 years away.
One researcher has said,
"Worrying about AI safety
is like worrying about overpopulation on Mars."
This is the Silicon Valley version
of "don't worry your pretty little head about it."

No one seems to notice
that referencing the time horizon
is a total non sequitur.
If intelligence is just a matter of information processing,
and we continue to improve our machines,
we will produce some form of superintelligence.
And we have no idea how long it will take us
to create the conditions to do that safely.
Let me say that again.
We have no idea how long it will take us
to create the conditions to do that safely.
And if you haven't noticed, 50 years is not what it used to be.
This is 50 years in months.
This is how long we've had the iPhone.
This is how long "The Simpsons" has been on television.
Fifty years is not that much time
to meet one of the greatest challenges our species will ever face.
Once again, we seem to be failing to have an appropriate emotional response
to what we have every reason to believe is coming.
The computer scientist Stuart Russell has a nice analogy here.
He said, imagine that we received a message from an alien civilization,
which read:
"People of Earth,
we will arrive on your planet in 50 years.
Get ready."
And now we're just counting down the months until the mothership lands?
We would feel a little more urgency than we do.
Another reason we're told not to worry
is that these machines can't help but share our values
because they will be literally extensions of ourselves.
They'll be grafted onto our brains,
and we'll essentially become their limbic systems.
Now take a moment to consider
that the safest and only prudent path forward,
recommended,
is to implant this technology directly into our brains.
Now, this may in fact be the safest and only prudent path forward,

but usually one's safety concerns about a technology
have to be pretty much worked out before you stick it inside your head.

The deeper problem is that building superintelligent AI on its own
seems likely to be easier
than building superintelligent AI
and having the completed neuroscience
that allows us to seamlessly integrate our minds with it.
And given that the companies and governments doing this work
are likely to perceive themselves as being in a race against all others,
given that to win this race is to win the world,
provided you don't destroy it in the next moment,
then it seems likely that whatever is easier to do
will get done first.
Now, unfortunately, I don't have a solution to this problem,
apart from recommending that more of us think about it.
I think we need something like a Manhattan Project
on the topic of artificial intelligence.
Not to build it, because I think we'll inevitably do that,
but to understand how to avoid an arms race
and to build it in a way that is aligned with our interests.
When you're talking about superintelligent AI
that can make changes to itself,
it seems that we only have one chance to get the initial conditions right,
and even then we will need to absorb
the economic and political consequences of getting them right.
But the moment we admit
that information processing is the source of intelligence,
that some appropriate computational system is what the basis of intelligence is,
and we admit that we will improve these systems continuously,
and we admit that the horizon of cognition very likely far exceeds
what we currently know,
then we have to admit
that we are in the process of building some sort of god.
Now would be a good time
to make sure it's a god we can live with.
Thank you very much.

How AI can save our humanity

Kai-Fu Lee

I'm going to talk about how AI and mankind can coexist,
but first, we have to rethink about our human values.
So let me first make a confession about my errors in my values.
It was 11 o'clock, December 16, 1991.
I was about to become a father for the first time.
My wife, Shen-Ling, lay in the hospital bed
going through a very difficult 12-hour labor.
I sat by her bedside
but looked anxiously at my watch,
and I knew something that she didn't.
I knew that if in one hour,
our child didn't come,
I was going to leave her there
and go back to work
and make a presentation about AI
to my boss, Apple's CEO.
Fortunately, my daughter was born at 11:30 --

sparing me from doing the unthinkable,
and to this day, I am so sorry
for letting my work ethic take precedence over love for my family.

My AI talk, however, went off brilliantly.

Apple loved my work and decided to announce it
at TED1992,
26 years ago on this very stage.
I thought I had made one of the biggest, most important discoveries in AI,
and so did the "Wall Street Journal" on the following day.
But as far as discoveries went,
it turned out,
I didn't discover India, or America.
Perhaps I discovered a little island off of Portugal.
But the AI era of discovery continued,
and more scientists poured their souls into it.
About 10 years ago, the grand AI discovery
was made by three North American scientists,
and it's known as deep learning.
Deep learning is a technology that can take a huge amount of data
within one single domain
and learn to predict or decide at superhuman accuracy.
For example, if we show the deep learning network
a massive number of food photos,
it can recognize food
such as hot dog or no hot dog.

Or if we show it many pictures and videos and sensor data

from driving on the highway,
it can actually drive a car as well as a human being
on the highway.
And what if we showed this deep learning network
all the speeches made by President Trump?
Then this artificially intelligent President Trump,
actually the network --

can --

You like double oxymorons, huh?

So this network, if given the request to make a speech about AI,
he, or it, might say --
(Recording) Donald Trump: It's a great thing
to build a better world with artificial intelligence.
Kai-Fu Lee: And maybe in another language?
DT: (Speaking Chinese)

KFL: You didn't know he knew Chinese, did you?
So deep learning has become the core in the era of AI discovery,
and that's led by the US.
But we're now in the era of implementation,
where what really matters is execution, product quality, speed and data.
And that's where China comes in.
Chinese entrepreneurs,
who I fund as a venture capitalist,
are incredible workers,
amazing work ethic.
My example in the delivery room is nothing compared to how hard people work in China.
As an example, one startup tried to claim work-life balance:
"Come work for us because we are 996."
And what does that mean?
It means the work hours of 9am to 9pm, six days a week.
That's contrasted with other startups that do 997.
And the Chinese product quality has consistently gone up
in the past decade,
and that's because of a fiercely competitive environment.
In Silicon Valley, entrepreneurs compete in a very gentlemanly fashion,
sort of like in old wars in which each side took turns
to fire at each other.

But in the Chinese environment,
it's truly a gladiatorial fight to the death.
In such a brutal environment, entrepreneurs learn to grow very rapidly,
they learn to make their products better at lightning speed,
and they learn to hone their business models
until they're impregnable.
As a result, great Chinese products like WeChat and Weibo
are arguably better

than the equivalent American products from Facebook and Twitter.
And the Chinese market embraces this change
and accelerated change and paradigm shifts.
As an example, if any of you go to China,
you will see it's almost cashless and credit card-less,
because that thing that we all talk about, mobile payment,
has become the reality in China.
In the last year,
18.8 trillion US dollars were transacted on mobile internet,
and that's because of very robust technologies
built behind it.
It's even bigger than the China GDP.
And this technology, you can say, how can it be bigger than the GDP?
Because it includes all transactions:
wholesale, channels, retail, online, offline,
going into a shopping mall or going into a farmers market like this.
The technology is used by 700 million people
to pay each other, not just merchants,
so it's peer to peer,
and it's almost transaction-fee-free.
And it's instantaneous,
and it's used everywhere.
And finally, the China market is enormous.
This market is large,
which helps give entrepreneurs more users, more revenue,
more investment, but most importantly,
it gives the entrepreneurs a chance to collect a huge amount of data
which becomes rocket fuel for the AI engine.
So as a result, the Chinese AI companies
have leaped ahead
so that today, the most valuable companies
in computer vision, speech recognition,
speech synthesis, machine translation and drones
are all Chinese companies.
So with the US leading the era of discovery
and China leading the era of implementation,
we are now in an amazing age
where the dual engine of the two superpowers
are working together
to drive the fastest revolution in technology
that we have ever seen as humans.
And this will bring tremendous wealth,
unprecedented wealth:
16 trillion dollars, according to PwC,
in terms of added GDP to the worldwide GDP by 2030.
It will also bring immense challenges
in terms of potential job replacements.
Whereas in the Industrial Age
it created more jobs
because craftsman jobs were being decomposed into jobs in the assembly line,
so more jobs were created.

But AI completely replaces the individual jobs
in the assembly line with robots.
And it's not just in factories,
but truckers, drivers
and even jobs like telesales, customer service
and hematologists as well as radiologists
over the next 15 years
are going to be gradually replaced
by artificial intelligence.
And only the creative jobs --

I have to make myself safe, right?
Really, the creative jobs are the ones that are protected,
because AI can optimize but not create.
But what's more serious than the loss of jobs
is the loss of meaning,
because the work ethic in the Industrial Age
has brainwashed us into thinking that work is the reason we exist,
that work defined the meaning of our lives.
And I was a prime and willing victim to that type of workaholic thinking.
I worked incredibly hard.
That's why I almost left my wife in the delivery room,
that's why I worked 996 alongside my entrepreneurs.
And that obsession that I had with work
ended abruptly a few years ago
when I was diagnosed with fourth stage lymphoma.
The PET scan here shows over 20 malignant tumors
jumping out like fireballs,
melting away my ambition.
But more importantly,
it helped me reexamine my life.
Knowing that I may only have a few months to live
caused me to see how foolish it was
for me to base my entire self-worth
on how hard I worked and the accomplishments from hard work.
My priorities were completely out of order.
I neglected my family.
My father had passed away,
and I never had a chance to tell him I loved him.
My mother had dementia and no longer recognized me,
and my children had grown up.
During my chemotherapy,
I read a book by Bronnie Ware
who talked about dying wishes and regrets of the people in the deathbed.
She found that facing death,
nobody regretted that they didn't work hard enough in this life.
They only regretted that they didn't spend enough time with their loved ones
and that they didn't spread their love.
So I am fortunately today in remission.

So I can be back at TED again

to share with you that I have changed my ways.
I now only work 965 --
occasionally 996, but usually 965.
I moved closer to my mother,
my wife usually travels with me,
and when my kids have vacation, if they don't come home, I go to them.
So it's a new form of life
that helped me recognize
how important it is that love is for me,
and facing death helped me change my life,
but it also helped me see a new way
of how AI should impact mankind
and work and coexist with mankind,
that really, AI is taking away a lot of routine jobs,
but routine jobs are not what we're about.
Why we exist is love.
When we hold our newborn baby,
love at first sight,
or when we help someone in need,
humans are uniquely able to give and receive love,
and that's what differentiates us from AI.
Despite what science fiction may portray,
I can responsibly tell you that AI has no love.
When AlphaGo defeated the world champion Ke Jie,
while Ke Jie was crying and loving the game of go,
AlphaGo felt no happiness from winning
and certainly no desire to hug a loved one.
So how do we differentiate ourselves
as humans in the age of AI?
We talked about the axis of creativity,
and certainly that is one possibility,
and now we introduce a new axis
that we can call compassion, love, or empathy.
Those are things that AI cannot do.
So as AI takes away the routine jobs,
I like to think we can, we should and we must create jobs of compassion.
You might ask how many of those there are,
but I would ask you:
Do you not think that we are going to need a lot of social workers
to help us make this transition?
Do you not think we need a lot of compassionate caregivers
to give more medical care to more people?
Do you not think we're going to need 10 times more teachers
to help our children find their way
to survive and thrive in this brave new world?
And with all the newfound wealth,
should we not also make labors of love into careers
and let elderly accompaniment
or homeschooling become careers also?

This graph is surely not perfect,

but it points at four ways that we can work with AI.
AI will come and take away the routine jobs
and in due time, we will be thankful.
AI will become great tools for the creatives
so that scientists, artists, musicians and writers
can be even more creative.
AI will work with humans as analytical tools
that humans can wrap their warmth around
for the high-compassion jobs.
And we can always differentiate ourselves
with the uniquely capable jobs
that are both compassionate and creative,
using and leveraging our irreplaceable brains and hearts.
So there you have it:
a blueprint of coexistence for humans and AI.
AI is serendipity.
It is here to liberate us from routine jobs,
and it is here to remind us what it is that makes us human.
So let us choose to embrace AI and to love one another.
Thank you.

The danger of AI is weirder than you think

Janelle Shane

So, artificial intelligence
is known for disrupting all kinds of industries.
What about ice cream?
What kind of mind-blowing new flavors could we generate
with the power of an advanced artificial intelligence?
So I teamed up with a group of coders from Kealing Middle School
to find out the answer to this question.
They collected over 1,600 existing ice cream flavors,
and together, we fed them to an algorithm to see what it would generate.
And here are some of the flavors that the AI came up with.
[Pumpkin Trash Break]

[Peanut Butter Slime]
[Strawberry Cream Disease]

These flavors are not delicious, as we might have hoped they would be.
So the question is: What happened?
What went wrong?
Is the AI trying to kill us?
Or is it trying to do what we asked, and there was a problem?
In movies, when something goes wrong with AI,
it's usually because the AI has decided
that it doesn't want to obey the humans anymore,
and it's got its own goals, thank you very much.
In real life, though, the AI that we actually have
is not nearly smart enough for that.
It has the approximate computing power
of an earthworm,
or maybe at most a single honeybee,
and actually, probably maybe less.
Like, we're constantly learning new things about brains
that make it clear how much our AIs don't measure up to real brains.
So today's AI can do a task like identify a pedestrian in a picture,
but it doesn't have a concept of what the pedestrian is
beyond that it's a collection of lines and textures and things.
It doesn't know what a human actually is.
So will today's AI do what we ask it to do?
It will if it can,
but it might not do what we actually want.
So let's say that you were trying to get an AI
to take this collection of robot parts
and assemble them into some kind of robot to get from Point A to Point B.
Now, if you were going to try and solve this problem
by writing a traditional-style computer program,
you would give the program step-by-step instructions
on how to take these parts,
how to assemble them into a robot with legs
and then how to use those legs to walk to Point B.

But when you're using AI to solve the problem,
it goes differently.
You don't tell it how to solve the problem,
you just give it the goal,
and it has to figure out for itself via trial and error
how to reach that goal.
And it turns out that the way AI tends to solve this particular problem
is by doing this:
it assembles itself into a tower and then falls over
and lands at Point B.
And technically, this solves the problem.
Technically, it got to Point B.
The danger of AI is not that it's going to rebel against us,
it's that it's going to do exactly what we ask it to do.
So then the trick of working with AI becomes:
How do we set up the problem so that it actually does what we want?
So this little robot here is being controlled by an AI.
The AI came up with a design for the robot legs
and then figured out how to use them to get past all these obstacles.
But when David Ha set up this experiment,
he had to set it up with very, very strict limits
on how big the AI was allowed to make the legs,
because otherwise ...

And technically, it got to the end of that obstacle course.
So you see how hard it is to get AI to do something as simple as just walk.
So seeing the AI do this, you may say, OK, no fair,
you can't just be a tall tower and fall over,
you have to actually, like, use legs to walk.
And it turns out, that doesn't always work, either.
This AI's job was to move fast.
They didn't tell it that it had to run facing forward
or that it couldn't use its arms.
So this is what you get when you train AI to move fast,
you get things like somersaulting and silly walks.
It's really common.
So is twitching along the floor in a heap.

So in my opinion, you know what should have been a whole lot weirder
is the "Terminator" robots.
Hacking "The Matrix" is another thing that AI will do if you give it a chance.
So if you train an AI in a simulation,
it will learn how to do things like hack into the simulation's math errors
and harvest them for energy.
Or it will figure out how to move faster by glitching repeatedly into the floor.
When you're working with AI,
it's less like working with another human
and a lot more like working with some kind of weird force of nature.
And it's really easy to accidentally give AI the wrong problem to solve,
and often we don't realize that until something has actually gone wrong.
So here's an experiment I did,

where I wanted the AI to copy paint colors,
to invent new paint colors,
given the list like the ones here on the left.
And here's what the AI actually came up with.
[Sindis Poop, Turdly, Suffer, Gray Pubic]

So technically,
it did what I asked it to.
I thought I was asking it for, like, nice paint color names,
but what I was actually asking it to do
was just imitate the kinds of letter combinations
that it had seen in the original.
And I didn't tell it anything about what words mean,
or that there are maybe some words
that it should avoid using in these paint colors.
So its entire world is the data that I gave it.
Like with the ice cream flavors, it doesn't know about anything else.
So it is through the data
that we often accidentally tell AI to do the wrong thing.
This is a fish called a tench.
And there was a group of researchers
who trained an AI to identify this tench in pictures.
But then when they asked it
what part of the picture it was actually using to identify the fish,
here's what it highlighted.
Yes, those are human fingers.
Why would it be looking for human fingers
if it's trying to identify a fish?
Well, it turns out that the tench is a trophy fish,
and so in a lot of pictures that the AI had seen of this fish
during training,
the fish looked like this.

And it didn't know that the fingers aren't part of the fish.
So you see why it is so hard to design an AI
that actually can understand what it's looking at.
And this is why designing the image recognition
in self-driving cars is so hard,
and why so many self-driving car failures
are because the AI got confused.
I want to talk about an example from 2016.
There was a fatal accident when somebody was using Tesla's autopilot AI,
but instead of using it on the highway like it was designed for,
they used it on city streets.
And what happened was,
a truck drove out in front of the car and the car failed to brake.
Now, the AI definitely was trained to recognize trucks in pictures.
But what it looks like happened is
the AI was trained to recognize trucks on highway driving,
where you would expect to see trucks from behind.
Trucks on the side is not supposed to happen on a highway,

and so when the AI saw this truck,
it looks like the AI recognized it as most likely to be a road sign
and therefore, safe to drive underneath.
Here's an AI misstep from a different field.
Amazon recently had to give up on a résumé-sorting algorithm
that they were working on
when they discovered that the algorithm had learned to discriminate against women.
What happened is they had trained it on example résumés
of people who they had hired in the past.
And from these examples, the AI learned to avoid the résumés of people
who had gone to women's colleges
or who had the word "women" somewhere in their resume,
as in, "women's soccer team" or "Society of Women Engineers."
The AI didn't know that it wasn't supposed to copy this particular thing
that it had seen the humans do.
And technically, it did what they asked it to do.
They just accidentally asked it to do the wrong thing.
And this happens all the time with AI.
AI can be really destructive and not know it.
So the AIs that recommend new content in Facebook, in YouTube,
they're optimized to increase the number of clicks and views.
And unfortunately, one way that they have found of doing this
is to recommend the content of conspiracy theories or bigotry.
The AIs themselves don't have any concept of what this content actually is,
and they don't have any concept of what the consequences might be
of recommending this content.
So, when we're working with AI,
it's up to us to avoid problems.
And avoiding things going wrong,
that may come down to the age-old problem of communication,
where we as humans have to learn how to communicate with AI.
We have to learn what AI is capable of doing and what it's not,
and to understand that, with its tiny little worm brain,
AI doesn't really understand what we're trying to ask it to do.
So in other words, we have to be prepared to work with AI
that's not the super-competent, all-knowing AI of science fiction.
We have to be prepared to work with an AI
that's the one that we actually have in the present day.
And present-day AI is plenty weird enough.
Thank you.

Don't fear superintelligent AI

Grady Booch

When I was a kid, I was the quintessential nerd.
I think some of you were, too.

And you, sir, who laughed the loudest, you probably still are.

I grew up in a small town in the dusty plains of north Texas,
the son of a sheriff who was the son of a pastor.
Getting into trouble was not an option.
And so I started reading calculus books for fun.

You did, too.
That led me to building a laser and a computer and model rockets,
and that led me to making rocket fuel in my bedroom.
Now, in scientific terms,
we call this a very bad idea.

Around that same time,
Stanley Kubrick's "2001: A Space Odyssey" came to the theaters,
and my life was forever changed.
I loved everything about that movie,
especially the HAL 9000.
Now, HAL was a sentient computer
designed to guide the Discovery spacecraft
from the Earth to Jupiter.
HAL was also a flawed character,
for in the end he chose to value the mission over human life.
Now, HAL was a fictional character,
but nonetheless he speaks to our fears,
our fears of being subjugated
by some unfeeling, artificial intelligence
who is indifferent to our humanity.
I believe that such fears are unfounded.
Indeed, we stand at a remarkable time
in human history,
where, driven by refusal to accept the limits of our bodies and our minds,
we are building machines
of exquisite, beautiful complexity and grace
that will extend the human experience
in ways beyond our imagining.
After a career that led me from the Air Force Academy
to Space Command to now,
I became a systems engineer,
and recently I was drawn into an engineering problem
associated with NASA's mission to Mars.
Now, in space flights to the Moon,
we can rely upon mission control in Houston
to watch over all aspects of a flight.
However, Mars is 200 times further away,

and as a result it takes on average 13 minutes
for a signal to travel from the Earth to Mars.
If there's trouble, there's not enough time.
And so a reasonable engineering solution
calls for us to put mission control
inside the walls of the Orion spacecraft.
Another fascinating idea in the mission profile
places humanoid robots on the surface of Mars
before the humans themselves arrive,
first to build facilities
and later to serve as collaborative members of the science team.
Now, as I looked at this from an engineering perspective,
it became very clear to me that what I needed to architect
was a smart, collaborative,
socially intelligent artificial intelligence.
In other words, I needed to build something very much like a HAL
but without the homicidal tendencies.

Let's pause for a moment.
Is it really possible to build an artificial intelligence like that?
Actually, it is.
In many ways,
this is a hard engineering problem
with elements of AI,
not some wet hair ball of an AI problem that needs to be engineered.
To paraphrase Alan Turing,
I'm not interested in building a sentient machine.
I'm not building a HAL.
All I'm after is a simple brain,
something that offers the illusion of intelligence.
The art and the science of computing have come a long way
since HAL was onscreen,
and I'd imagine if his inventor Dr. Chandra were here today,
he'd have a whole lot of questions for us.
Is it really possible for us
to take a system of millions upon millions of devices,
to read in their data streams,
to predict their failures and act in advance?
Yes.
Can we build systems that converse with humans in natural language?
Yes.
Can we build systems that recognize objects, identify emotions,
emote themselves, play games and even read lips?
Yes.
Can we build a system that sets goals,
that carries out plans against those goals and learns along the way?
Yes.
Can we build systems that have a theory of mind?
This we are learning to do.
Can we build systems that have an ethical and moral foundation?
This we must learn how to do.

So let's accept for a moment
that it's possible to build such an artificial intelligence
for this kind of mission and others.
The next question you must ask yourself is,
should we fear it?
Now, every new technology
brings with it some measure of trepidation.
When we first saw cars,
people lamented that we would see the destruction of the family.
When we first saw telephones come in,
people were worried it would destroy all civil conversation.
At a point in time we saw the written word become pervasive,
people thought we would lose our ability to memorize.
These things are all true to a degree,
but it's also the case that these technologies
brought to us things that extended the human experience
in some profound ways.
So let's take this a little further.
I do not fear the creation of an AI like this,
because it will eventually embody some of our values.
Consider this: building a cognitive system is fundamentally different
than building a traditional software-intensive system of the past.
We don't program them. We teach them.
In order to teach a system how to recognize flowers,
I show it thousands of flowers of the kinds I like.
In order to teach a system how to play a game --
Well, I would. You would, too.
I like flowers. Come on.
To teach a system how to play a game like Go,
I'd have it play thousands of games of Go,
but in the process I also teach it
how to discern a good game from a bad game.
If I want to create an artificially intelligent legal assistant,
I will teach it some corpus of law
but at the same time I am fusing with it
the sense of mercy and justice that is part of that law.
In scientific terms, this is what we call ground truth,
and here's the important point:
in producing these machines,
we are therefore teaching them a sense of our values.
To that end, I trust an artificial intelligence
the same, if not more, as a human who is well-trained.
But, you may ask,
what about rogue agents,
some well-funded nongovernment organization?
I do not fear an artificial intelligence in the hand of a lone wolf.
Clearly, we cannot protect ourselves against all random acts of violence,
but the reality is such a system
requires substantial training and subtle training
far beyond the resources of an individual.
And furthermore,

it's far more than just injecting an internet virus to the world,
where you push a button, all of a sudden it's in a million places
and laptops start blowing up all over the place.
Now, these kinds of substances are much larger,
and we'll certainly see them coming.
Do I fear that such an artificial intelligence
might threaten all of humanity?
If you look at movies such as "The Matrix," "Metropolis,"
"The Terminator," shows such as "Westworld,"
they all speak of this kind of fear.
Indeed, in the book "Superintelligence" by the philosopher Nick Bostrom,
he picks up on this theme
and observes that a superintelligence might not only be dangerous,
it could represent an existential threat to all of humanity.
Dr. Bostrom's basic argument
is that such systems will eventually
have such an insatiable thirst for information
that they will perhaps learn how to learn
and eventually discover that they may have goals
that are contrary to human needs.
Dr. Bostrom has a number of followers.
He is supported by people such as Elon Musk and Stephen Hawking.
With all due respect
to these brilliant minds,
I believe that they are fundamentally wrong.
Now, there are a lot of pieces of Dr. Bostrom's argument to unpack,
and I don't have time to unpack them all,
but very briefly, consider this:
super knowing is very different than super doing.
HAL was a threat to the Discovery crew
only insofar as HAL commanded all aspects of the Discovery.
So it would have to be with a superintelligence.
It would have to have dominion over all of our world.
This is the stuff of Skynet from the movie "The Terminator"
in which we had a superintelligence
that commanded human will,
that directed every device that was in every corner of the world.
Practically speaking,
it ain't gonna happen.
We are not building AIs that control the weather,
that direct the tides,
that command us capricious, chaotic humans.
And furthermore, if such an artificial intelligence existed,
it would have to compete with human economies,
and thereby compete for resources with us.
And in the end --
don't tell Siri this --
we can always unplug them.

We are on an incredible journey
of coevolution with our machines.

The humans we are today
are not the humans we will be then.
To worry now about the rise of a superintelligence
is in many ways a dangerous distraction
because the rise of computing itself
brings to us a number of human and societal issues
to which we must now attend.
How shall I best organize society
when the need for human labor diminishes?
How can I bring understanding and education throughout the globe
and still respect our differences?
How might I extend and enhance human life through cognitive healthcare?
How might I use computing
to help take us to the stars?
And that's the exciting thing.
The opportunities to use computing
to advance the human experience
are within our reach,
here and now,
and we are just beginning.
Thank you very much.

How to keep human bias out of AI

Kriti Sharma

How many decisions have been made about you today,
or this week or this year,
by artificial intelligence?
I build AI for a living
so, full disclosure, I'm kind of a nerd.
And because I'm kind of a nerd,
wherever some new news story comes out
about artificial intelligence stealing all our jobs,
or robots getting citizenship of an actual country,
I'm the person my friends and followers message
freaking out about the future.
We see this everywhere.
This media panic that our robot overlords are taking over.
We could blame Hollywood for that.
But in reality, that's not the problem we should be focusing on.
There is a more pressing danger, a bigger risk with AI,
that we need to fix first.
So we are back to this question:
How many decisions have been made about you today by AI?
And how many of these
were based on your gender, your race or your background?
Algorithms are being used all the time
to make decisions about who we are and what we want.
Some of the women in this room will know what I'm talking about
if you've been made to sit through those pregnancy test adverts on YouTube
like 1,000 times.
Or you've scrolled past adverts of fertility clinics
on your Facebook feed.
Or in my case, Indian marriage bureaus.

But AI isn't just being used to make decisions
about what products we want to buy
or which show we want to binge watch next.
I wonder how you'd feel about someone who thought things like this:
"A black or Latino person
is less likely than a white person to pay off their loan on time."
"A person called John makes a better programmer
than a person called Mary."
"A black man is more likely to be a repeat offender than a white man."
You're probably thinking,
"Wow, that sounds like a pretty sexist, racist person," right?
These are some real decisions that AI has made very recently,
based on the biases it has learned from us,
from the humans.
AI is being used to help decide whether or not you get that job interview;
how much you pay for your car insurance;
how good your credit score is;
and even what rating you get in your annual performance review.

But these decisions are all being filtered through
its assumptions about our identity, our race, our gender, our age.
How is that happening?
Now, imagine an AI is helping a hiring manager
find the next tech leader in the company.
So far, the manager has been hiring mostly men.
So the AI learns men are more likely to be programmers than women.
And it's a very short leap from there to:
men make better programmers than women.
We have reinforced our own bias into the AI.
And now, it's screening out female candidates.
Hang on, if a human hiring manager did that,
we'd be outraged, we wouldn't allow it.
This kind of gender discrimination is not OK.
And yet somehow, AI has become above the law,
because a machine made the decision.
That's not it.
We are also reinforcing our bias in how we interact with AI.
How often do you use a voice assistant like Siri, Alexa or even Cortana?
They all have two things in common:
one, they can never get my name right,
and second, they are all female.
They are designed to be our obedient servants,
turning your lights on and off, ordering your shopping.
You get male AIs too, but they tend to be more high-powered,
like IBM Watson, making business decisions,
Salesforce Einstein or ROSS, the robot lawyer.
So poor robots, even they suffer from sexism in the workplace.

Think about how these two things combine
and affect a kid growing up in today's world around AI.
So they're doing some research for a school project
and they Google images of CEO.
The algorithm shows them results of mostly men.
And now, they Google personal assistant.
As you can guess, it shows them mostly females.
And then they want to put on some music, and maybe order some food,
and now, they are barking orders at an obedient female voice assistant.
Some of our brightest minds are creating this technology today.
Technology that they could have created in any way they wanted.
And yet, they have chosen to create it in the style of 1950s "Mad Man" secretary.
Yay!
But OK, don't worry,
this is not going to end with me telling you
that we are all heading towards sexist, racist machines running the world.
The good news about AI is that it is entirely within our control.
We get to teach the right values, the right ethics to AI.
So there are three things we can do.
One, we can be aware of our own biases
and the bias in machines around us.
Two, we can make sure that diverse teams are building this technology.

And three, we have to give it diverse experiences to learn from.
I can talk about the first two from personal experience.
When you work in technology
and you don't look like a Mark Zuckerberg or Elon Musk,
your life is a little bit difficult, your ability gets questioned.
Here's just one example.
Like most developers, I often join online tech forums
and share my knowledge to help others.
And I've found,
when I log on as myself, with my own photo, my own name,
I tend to get questions or comments like this:
"What makes you think you're qualified to talk about AI?"
"What makes you think you know about machine learning?"
So, as you do, I made a new profile,
and this time, instead of my own picture, I chose a cat with a jet pack on it.
And I chose a name that did not reveal my gender.
You can probably guess where this is going, right?
So, this time, I didn't get any of those patronizing comments about my ability
and I was able to actually get some work done.
And it sucks, guys.
I've been building robots since I was 15,
I have a few degrees in computer science,
and yet, I had to hide my gender
in order for my work to be taken seriously.
So, what's going on here?
Are men just better at technology than women?
Another study found
that when women coders on one platform hid their gender, like myself,
their code was accepted four percent more than men.
So this is not about the talent.
This is about an elitism in AI
that says a programmer needs to look like a certain person.
What we really need to do to make AI better
is bring people from all kinds of backgrounds.
We need people who can write and tell stories
to help us create personalities of AI.
We need people who can solve problems.
We need people who face different challenges
and we need people who can tell us what are the real issues that need fixing
and help us find ways that technology can actually fix it.
Because, when people from diverse backgrounds come together,
when we build things in the right way,
the possibilities are limitless.
And that's what I want to end by talking to you about.
Less racist robots, less machines that are going to take our jobs --
and more about what technology can actually achieve.
So, yes, some of the energy in the world of AI,
in the world of technology
is going to be about what ads you see on your stream.
But a lot of it is going towards making the world so much better.
Think about a pregnant woman in the Democratic Republic of Congo,

who has to walk 17 hours to her nearest rural prenatal clinic
to get a checkup.
What if she could get diagnosis on her phone, instead?
Or think about what AI could do
for those one in three women in South Africa
who face domestic violence.
If it wasn't safe to talk out loud,
they could get an AI service to raise alarm,
get financial and legal advice.
These are all real examples of projects that people, including myself,
are working on right now, using AI.
So, I'm sure in the next couple of days there will be yet another news story
about the existential risk,
robots taking over and coming for your jobs.

And when something like that happens,
I know I'll get the same messages worrying about the future.
But I feel incredibly positive about this technology.
This is our chance to remake the world into a much more equal place.
But to do that, we need to build it the right way from the get go.
We need people of different genders, races, sexualities and backgrounds.
We need women to be the makers
and not just the machines who do the makers' bidding.
We need to think very carefully what we teach machines,
what data we give them,
so they don't just repeat our own past mistakes.
So I hope I leave you thinking about two things.
First, I hope you leave thinking about bias today.
And that the next time you scroll past an advert
that assumes you are interested in fertility clinics
or online betting websites,
that you think and remember
that the same technology is assuming that a black man will reoffend.
Or that a woman is more likely to be a personal assistant than a CEO.
And I hope that reminds you that we need to do something about it.
And second,
I hope you think about the fact
that you don't need to look a certain way
or have a certain background in engineering or technology
to create AI,
which is going to be a phenomenal force for our future.
You don't need to look like a Mark Zuckerberg,
you can look like me.
And it is up to all of us in this room
to convince the governments and the corporations
to build AI technology for everyone,
including the edge cases.
And for us all to get education about this phenomenal technology in the future.
Because if we do that, then we've only just scratched the surface of what we can achieve with AI.
Thank you.

The disappearing computer — and a world where you can take AI everywhere

Imran Chaudhri

I spent 22 incredible years at Apple,
helping to design experiences and devices
ranging from the Mac
to the iPhone to the Apple Watch.
And as the power of compute increased,
the size of our computers or our devices decreased.
The desktop paved the way for extraordinary interconnectedness,
but it was stuck to your desk.
The laptop provided portability,
but you still had to be sitting down to use it.
And the smartphone evolved us into the modern, connected humans we are,
providing millions the ability to access the internet from our pockets.
And the smart watch was a window to that phone.
A companion device with a whole host of health insights,
all shrunk down to your wrist.
But what comes next?
Some believe AR/VR glasses like these are the answer,
but they merely move the screens we already have in our lives today
to being just millimeters away from our eyeballs.
A further barrier between you and the world.
And the future is not on your face.
In fact, in 2017,
the legendary tech journalist Walt Mossberg wrote in his final column
that he felt that soon, one day, technology would become invisible.
And that the computer would disappear.
And we agree.
(Ringing)
Sorry.
This is my wife.
I'm going to have to get this.
Hello?
Bethany Bongiorno: Hey, babe.
IC: Hey, Bethany.
How's it going?
BB: Good. Are you at TED?
IC: Yeah, I'm on the red circle right now, actually.
Bethany: Oh, great, good luck.
And don't forget to mention me.

IC: I won't, babe, thank you.
Bethany: Love you.
IC: Love you, too. Bye.
It's going to get different in a minute.

So my wife, Bethany, and our entire company, Humane,
have been working to answer the question of what comes next.

And you may ask yourself, why?
Why would anybody do this?
It's because we love building technology
that genuinely makes people's lives better.
And we believe that artificial intelligence or AI
would be the driving force behind the next leap in device design.
And there is an incredible amount of stuff that's happening in this space.
Huge, huge advancements.
And even Bill Gates has said of OpenAI's GPT
that it's only the second most revolutionary technology demonstration
that he's seen in his entire lifetime.
But what do we do with all these incredible developments,
and how do we actually harness these to genuinely make our life better?
If we get this right,
AI will unlock a world of possibility for all of us.
And today I want to share with you
what we think is a solution to that end.
And it's the first time we're doing so openly.
It's a new kind of wearable device
and platform that's built entirely from the ground up
for artificial intelligence.
And is completely standalone.
You don't need a smartphone or any other device to pair with it.
In fact, I'm wearing one right now.
And it interacts with the world the way you interact with the world.
Hearing what you hear, seeing what you see.
While being privacy-first and safe
and completely fading into the background of your life.
We like to say that the experience is screenless, seamless
and sensing,
allowing you to access the power of compute
while remaining present in your surroundings,
fixing a balance that's felt out of place for some time now.
And I can't wait to share more details about what we've built,
and I will in the next few months.
But today I want to talk to you about what it unlocks.
And what it means to be able to take AI with you everywhere.
And what happens when technology increasingly disappears.
Technology becoming invisible affords us new opportunities
of how we interact with compute.
We've become so accustomed to tapping on an app
or moving a cursor with a mouse
that it feels second nature.
But that's by design.
When I was working on the iPhone,
I used to test interactions like slide-to-unlock
with my infant daughter.
She was the best possible focus group.
She’s 16 now,
and she's got a lot more ideas than she did back then.
This also, by the way, is the only non-AI generated image

that you'll see from me today.
And as I look at it now,
I see more than ever why a future driven by AI
is far better than a future that would involve more screens.
Like this.
He's cute, though.
But for the human-technology relationship to actually evolve beyond screens,
we need something radically different.
Let me show you.
Where can I find a gift for my wife before I have to leave tomorrow?
(Voice) Vancouver's Granville Island is a lively shopping district.
IC: That's an incredibly simple response for a very complex query.
How often do we find ourselves in a new city,
wrestling with our phones,
trying not to bump into people,
trying to figure out where we're going and where we're supposed to be?
It's even harder when we don't speak the language, right?
Let me show you something.
Invisible devices should feel so natural to use
that you almost forget about their existence.
(Voice speaking in French)
IC: You'll note that's me and my voice, speaking fluent French,
using an AI speech model that's part of my own AI.
This is not a deepfake.
In fact, it's deeply profound.
This is my AI giving me the ability to speak any language
and you having a chance to hear me speak that language
in my own emotion and my own voice.
Thank you.

This is moving away from the experiments that make us all concerned
about the direction compute is going in.
But it's instead using technology
to create real, responsible compute products
that are in service to us and built on trust.
This is good AI in action.
And we spent thousands of hours
reimagining and redesigning new types of compute interactions,
ranging from complex voice commands to intricate hand gestures,
all in service of trying to find more natural ways to interact with compute.
Why fumble for your phone when you can just hold an object
and ask questions about it?
The result almost feels like the entire world becomes your operating system.
And when compute disappears,
it allows us to get back to what really matters:
a new ability to be present.
Like riding a bicycle in the park and just ripping through emails
or going to a concert without having to hold up your phone to capture it.
Or experiencing your toddler's first steps
without a screen between you and your child.
In the future,

technology will be both ambient and contextual.
And this means harnessing AI to really understand you
and your surroundings
in order to achieve the best results.
Imagine this.
You've been in meetings all day
and you just want a summary of what you've missed.
Catch me up.
(Voice) Patrick is coming to tomorrow's design meeting.
Bethany wants to move next week's dinner,
and Oliver is asking about soccer this weekend.
IC: These are emails, calendar invites and messages,
all surfaced up to the top.
You can use these to help guide your decision making,
manage your workload
and sculpt tailored responses in your own voice.
And in the context of your life.
And we gain this context through machine learning.
The more you use our device powered by AI,
the more we can help you in all times of need.
Your AI effectively becomes an ever-evolving,
personalized form of memory.
And we think that's amazing.
In fact, let's say you're health conscious
or you have certain types of food considerations.
Let me just show you.
Picked up one of these chocolates.
Used to eat a ton of these when I was a kid.
Can I eat this?
(Voice) A milky bar contains cocoa butter.
Given your intolerance, you may want to avoid it.
IC: So I can’t eat these anymore.

But what's cool is my AI knows what's best for me.
But I'm in total control.
I'm going to eat it anyway.

Enjoy it.

IC: Your AI figures out exactly what you need.
And by the way, I love that there's no judgment.
I think it's amazing to be able to live freely.
Your AI figures out what you need at the speed of thought.
A sense that will ever be evolving as technology improves too.
And these examples are just the start.
As AI advances,
we will see how it will transform nearly every aspect of our lives.
In ways that will seem unimaginable right now.
In fact, Sam Altman from OpenAI feels the way we do.
And that AI is grossly underestimated.
And I'll add, so long as we get it right.

We really believe
that we're only beginning to scratch the surface of what's possible.
Embed advancements of AI, like
in our device that's actually built to disappear
and allow experiences to come forward,
and we open up entirely new possible ways of how you interact with technology
and how you interact with the world around you.
More humane, intuitive interactions
that are screenless, seamless and sensing.
This is so much more than devices just getting smaller or more powerful.
This is the possibility of reimagining the human-technology relationship
as we know it.
And that's what's so exciting.
It's a huge challenge, no doubt.
But it's the world that we want to live in.
One where technology not only helps you get back into the world
but enhances our ability to do so.
It's within reach.
And you saw some of it today.
The future will not be held in your hand,
and it won't be on your face either.
The future of technology might almost be invisible.
Thank you.

How humans and AI can work together to create better businesses

Sylvain Duranton

Let me share a paradox.
For the last 10 years,
many companies have been trying to become less bureaucratic,
to have fewer central rules and procedures,
more autonomy for their local teams to be more agile.
And now they are pushing artificial intelligence, AI,
unaware that cool technology
might make them more bureaucratic than ever.
Why?
Because AI operates just like bureaucracies.
The essence of bureaucracy
is to favor rules and procedures over human judgment.
And AI decides solely based on rules.
Many rules inferred from past data
but only rules.
And if human judgment is not kept in the loop,
AI will bring a terrifying form of new bureaucracy --
I call it "algocracy" --
where AI will take more and more critical decisions by the rules
outside of any human control.
Is there a real risk?
Yes.
I'm leading a team of 800 AI specialists.
We have deployed over 100 customized AI solutions
for large companies around the world.
And I see too many corporate executives behaving like bureaucrats from the past.
They want to take costly, old-fashioned humans out of the loop
and rely only upon AI to take decisions.
I call this the "human-zero mindset."
And why is it so tempting?
Because the other route, "Human plus AI," is long,
costly and difficult.
Business teams, tech teams, data-science teams
have to iterate for months
to craft exactly how humans and AI can best work together.
Long, costly and difficult.
But the reward is huge.
A recent survey from BCG and MIT
shows that 18 percent of companies in the world
are pioneering AI,
making money with it.
Those companies focus 80 percent of their AI initiatives
on effectiveness and growth,
taking better decisions --
not replacing humans with AI to save costs.
Why is it important to keep humans in the loop?

Simply because, left alone, AI can do very dumb things.
Sometimes with no consequences, like in this tweet.
"Dear Amazon,
I bought a toilet seat.
Necessity, not desire.
I do not collect them,
I'm not a toilet-seat addict.
No matter how temptingly you email me,
I am not going to think, 'Oh, go on, then,
one more toilet seat, I'll treat myself.' "

Sometimes, with more consequence, like in this other tweet.
"Had the same situation
with my mother's burial urn."

"For months after her death,
I got messages from Amazon, saying, 'If you liked that ...' "

Sometimes with worse consequences.
Take an AI engine rejecting a student application for university.
Why?
Because it has "learned," on past data,
characteristics of students that will pass and fail.
Some are obvious, like GPAs.
But if, in the past, all students from a given postal code have failed,
it is very likely that AI will make this a rule
and will reject every student with this postal code,
not giving anyone the opportunity to prove the rule wrong.
And no one can check all the rules,
because advanced AI is constantly learning.
And if humans are kept out of the room,
there comes the algocratic nightmare.
Who is accountable for rejecting the student?
No one, AI did.
Is it fair? Yes.
The same set of objective rules has been applied to everyone.
Could we reconsider for this bright kid with the wrong postal code?
No, algos don't change their mind.
We have a choice here.
Carry on with algocracy
or decide to go to "Human plus AI."
And to do this,
we need to stop thinking tech first,
and we need to start applying the secret formula.
To deploy "Human plus AI,"
10 percent of the effort is to code algos;
20 percent to build tech around the algos,
collecting data, building UI, integrating into legacy systems;
But 70 percent, the bulk of the effort,
is about weaving together AI with people and processes
to maximize real outcome.

AI fails when cutting short on the 70 percent.
The price tag for that can be small,
wasting many, many millions of dollars on useless technology.
Anyone cares?
Or real tragedies:
346 casualties in the recent crashes of two B-737 aircrafts
when pilots could not interact properly
with a computerized command system.
For a successful 70 percent,
the first step is to make sure that algos are coded by data scientists
and domain experts together.
Take health care for example.
One of our teams worked on a new drug with a slight problem.
When taking their first dose,
some patients, very few, have heart attacks.
So, all patients, when taking their first dose,
have to spend one day in hospital,
for monitoring, just in case.
Our objective was to identify patients who were at zero risk of heart attacks,
who could skip the day in hospital.
We used AI to analyze data from clinical trials,
to correlate ECG signal, blood composition, biomarkers,
with the risk of heart attack.
In one month,
our model could flag 62 percent of patients at zero risk.
They could skip the day in hospital.
Would you be comfortable staying at home for your first dose
if the algo said so?

Doctors were not.
What if we had false negatives,
meaning people who are told by AI they can stay at home, and die?

There started our 70 percent.
We worked with a team of doctors
to check the medical logic of each variable in our model.
For instance, we were using the concentration of a liver enzyme
as a predictor,
for which the medical logic was not obvious.
The statistical signal was quite strong.
But what if it was a bias in our sample?
That predictor was taken out of the model.
We also took out predictors for which experts told us
they cannot be rigorously measured by doctors in real life.
After four months,
we had a model and a medical protocol.
They both got approved
my medical authorities in the US last spring,
resulting in far less stress for half of the patients
and better quality of life.
And an expected upside on sales over 100 million for that drug.

Seventy percent weaving AI with team and processes
also means building powerful interfaces
for humans and AI to solve the most difficult problems together.
Once, we got challenged by a fashion retailer.
"We have the best buyers in the world.
Could you build an AI engine that would beat them at forecasting sales?
At telling how many high-end, light-green, men XL shirts
we need to buy for next year?
At predicting better what will sell or not
than our designers."
Our team trained a model in a few weeks, on past sales data,
and the competition was organized with human buyers.
Result?
AI wins, reducing forecasting errors by 25 percent.
Human-zero champions could have tried to implement this initial model
and create a fight with all human buyers.
Have fun.
But we knew that human buyers had insights on fashion trends
that could not be found in past data.
There started our 70 percent.
We went for a second test,
where human buyers were reviewing quantities
suggested by AI
and could correct them if needed.
Result?
Humans using AI ...
lose.
Seventy-five percent of the corrections made by a human
were reducing accuracy.
Was it time to get rid of human buyers?
No.
It was time to recreate a model
where humans would not try to guess when AI is wrong,
but where AI would take real input from human buyers.
We fully rebuilt the model
and went away from our initial interface, which was, more or less,
"Hey, human! This is what I forecast,
correct whatever you want,"
and moved to a much richer one, more like,
"Hey, humans!
I don't know the trends for next year.
Could you share with me your top creative bets?"
"Hey, humans!
Could you help me quantify those few big items?
I cannot find any good comparables in the past for them."
Result?
"Human plus AI" wins,
reducing forecast errors by 50 percent.
It took one year to finalize the tool.
Long, costly and difficult.
But profits and benefits

were in excess of 100 million of savings per year for that retailer.
Seventy percent on very sensitive topics
also means human have to decide what is right or wrong
and define rules for what AI can do or not,
like setting caps on prices to prevent pricing engines
[from charging] outrageously high prices to uneducated customers
who would accept them.
Only humans can define those boundaries --
there is no way AI can find them in past data.
Some situations are in the gray zone.
We worked with a health insurer.
He developed an AI engine to identify, among his clients,
people who are just about to go to hospital
to sell them premium services.
And the problem is,
some prospects were called by the commercial team
while they did not know yet
they would have to go to hospital very soon.
You are the CEO of this company.
Do you stop that program?
Not an easy question.
And to tackle this question, some companies are building teams,
defining ethical rules and standards to help business and tech teams set limits
between personalization and manipulation,
customization of offers and discrimination,
targeting and intrusion.
I am convinced that in every company,
applying AI where it really matters has massive payback.
Business leaders need to be bold
and select a few topics,
and for each of them, mobilize 10, 20, 30 people from their best teams --
tech, AI, data science, ethics --
and go through the full 10-, 20-, 70-percent cycle
of "Human plus AI,"
if they want to land AI effectively in their teams and processes.
There is no other way.
Citizens in developed economies already fear algocracy.
Seven thousand were interviewed in a recent survey.
More than 75 percent expressed real concerns
on the impact of AI on the workforce, on privacy,
on the risk of a dehumanized society.
Pushing algocracy creates a real risk of severe backlash against AI
within companies or in society at large.
"Human plus AI" is our only option
to bring the benefits of AI to the real world.
And in the end,
winning organizations will invest in human knowledge,
not just AI and data.
Recruiting, training, rewarding human experts.
Data is said to be the new oil,
but believe me, human knowledge will make the difference,

because it is the only derrick available
to pump the oil hidden in the data.
Thank you.

How AI can enhance our memory, work and social lives

Tom Gruber

I'm here to offer you a new way to think about my field,
artificial intelligence.
I think the purpose of AI
is to empower humans with machine intelligence.
And as machines get smarter,
we get smarter.
I call this "humanistic AI" --
artificial intelligence designed to meet human needs
by collaborating and augmenting people.
Now, today I'm happy to see
that the idea of an intelligent assistant
is mainstream.
It's the well-accepted metaphor for the interface between humans and AI.
And the one I helped create is called Siri.
You know Siri.
Siri is the thing that knows your intent
and helps you do it for you,
helps you get things done.
But what you might not know is that we designed Siri
as humanistic AI,
to augment people with a conversational interface
that made it possible for them to use mobile computing,
regardless of who they were and their abilities.
Now for most of us,
the impact of this technology
is to make things a little bit easier to use.
But for my friend Daniel,
the impact of the AI in these systems is a life changer.
You see, Daniel is a really social guy,
and he's blind and quadriplegic,
which makes it hard to use those devices that we all take for granted.
The last time I was at his house, his brother said,
"Hang on a second, Daniel's not ready.
He's on the phone with a woman he met online."
I'm like, "That's cool, how'd he do it?"
Well, Daniel uses Siri to manage his own social life --
his email, text and phone --
without depending on his caregivers.
This is kind of interesting, right?
The irony here is great.
Here's the man whose relationship with AI
helps him have relationships with genuine human beings.
And this is humanistic AI.
Another example with life-changing consequences
is diagnosing cancer.
When a doctor suspects cancer,
they take a sample and send it to a pathologist,
who looks at it under a microscope.

Now, pathologists look at hundreds of slides
and millions of cells every day.
So to support this task,
some researchers made an AI classifier.
Now, the classifier says, "Is this cancer or is this not cancer?"
looking at the pictures.
The classifier was pretty good,
but not as good as the person,
who got it right most of the time.
But when they combine the ability of the machine and the human together,
accuracy went to 99.5 percent.
Adding that AI to a partnership eliminated 85 percent of the errors
that the human pathologist would have made working alone.
That's a lot of cancer that would have otherwise gone untreated.
Now, for the curious, it turns out
that the human was better at rejecting false positives,
and the machine was better at recognizing those hard-to-spot cases.
But the lesson here isn't about which agent is better
at this image-classification task.
Those things are changing every day.
The lesson here
is that by combining the abilities of the human and machine,
it created a partnership that had superhuman performance.
And that is humanistic AI.
Now let's look at another example
with turbocharging performance.
This is design.
Now, let's say you're an engineer.
You want to design a new frame for a drone.
You get out your favorite software tools, CAD tools,
and you enter the form and the materials, and then you analyze performance.
That gives you one design.
If you give those same tools to an AI,
it can generate thousands of designs.
This video by Autodesk is amazing.
This is real stuff.
So this transforms how we do design.
The human engineer now
says what the design should achieve,
and the machine says,
"Here's the possibilities."
Now in her job, the engineer's job
is to pick the one that best meets the goals of the design,
which she knows as a human better than anyone else,
using human judgment and expertise.
In this case, the winning form
looks kind of like something nature would have designed,
minus a few million years of evolution
and all that unnecessary fur.
Now let's see where this idea of humanistic AI might lead us
if we follow it into the speculative beyond.

What's a kind of augmentation that we would all like to have?
Well, how about cognitive enhancement?
Instead of asking,
"How smart can we make our machines?"
let's ask
"How smart can our machines make us?"
I mean, take memory for example.
Memory is the foundation of human intelligence.
But human memory is famously flawed.
We're great at telling stories,
but not getting the details right.
And our memories -- they decay over time.
I mean, like, where did the '60s go, and can I go there, too?

But what if you could have a memory that was as good as computer memory,
and was about your life?
What if you could remember every person you ever met,
how to pronounce their name,
their family details, their favorite sports,
the last conversation you had with them?
If you had this memory all your life,
you could have the AI look at all the interactions
you had with people over time
and help you reflect on the long arc of your relationships.
What if you could have the AI read everything you've ever read
and listen to every song you've ever heard?
From the tiniest clue, it could help you retrieve
anything you've ever seen or heard before.
Imagine what that would do for the ability to make new connections
and form new ideas.
And what about our bodies?
What if we could remember the consequences of every food we eat,
every pill we take,
every all-nighter we pull?
We could do our own science on our own data
about what makes us feel good and stay healthy.
And imagine how this could revolutionize
the way we manage allergies and chronic disease.
I believe that AI will make personal memory enhancement a reality.
I can't say when or what form factors are involved,
but I think it's inevitable,
because the very things that make AI successful today --
the availability of comprehensive data
and the ability for machines to make sense of that data --
can be applied to the data of our lives.
And those data are here today, available for all of us,
because we lead digitally mediated lives,
in mobile and online.
In my view, a personal memory is a private memory.
We get to choose what is and is not recalled and retained.
It's absolutely essential that this be kept very secure.

Now for most of us,
the impact of augmented personal memory
will be a more improved mental gain,
maybe, hopefully, a bit more social grace.
But for the millions who suffer from Alzheimer's and dementia,
the difference that augmented memory could make
is a difference between a life of isolation
and a life of dignity and connection.
We are in the middle of a renaissance in artificial intelligence right now.
I mean, in just the past few years,
we're beginning to see solutions to AI problems
that we have struggled with literally for decades:
speech understanding, text understanding,
image understanding.
We have a choice in how we use this powerful technology.
We can choose to use AI to automate and compete with us,
or we can use AI to augment and collaborate with us,
to overcome our cognitive limitations
and to help us do what we want to do,
only better.
And as we discover new ways to give machines intelligence,
we can distribute that intelligence
to all of the AI assistants in the world,
and therefore to every person, regardless of circumstance.
And that is why,
every time a machine gets smarter,
we get smarter.
That is an AI worth spreading.
Thank you.

How AI could save (not destroy) education?

Sal Khan

So anyone who's been paying attention for the last few months
has been seeing headlines like this,
especially in education.
The thesis has been:
students are going to be using ChatGPT and other forms of AI
to cheat, do their assignments.
They're not going to learn.
And it's going to completely undermine education as we know it.
Now, what I'm going to argue today
is not only are there ways to mitigate all of that,
if we put the right guardrails, we do the right things,
we can mitigate it.
But I think we're at the cusp of using AI
for probably the biggest positive transformation
that education has ever seen.
And the way we're going to do that
is by giving every student on the planet
an artificially intelligent but amazing personal tutor.
And we're going to give every teacher on the planet an amazing,
artificially intelligent teaching assistant.
And just to appreciate how big of a deal it would be
to give everyone a personal tutor,
I show you this clip
from Benjamin Bloom's 1984 2 sigma study,
or he called it the "2 sigma problem."
The 2 sigma comes from two standard deviation,
sigma, the symbol for standard deviation.
And he had good data that showed that look, a normal distribution,
that's the one that you see in the traditional bell curve
right in the middle, that's how the world kind of sorts itself out,
that if you were to give personal 1-to-1 to tutoring for students,
then you could actually get a distribution that looks like that right.
It says tutorial 1-to-1 with the asterisks,
like, that right distribution,
a two standard-deviation improvement.
Just to put that in plain language,
that could take your average student and turn them into an exceptional student.
It can take your below-average student
and turn them into an above-average student.
Now the reason why he framed it as a problem, was he said,
well, this is all good,
but how do you actually scale group instruction this way?
How do you actually give it to everyone in an economic way?
What I'm about to show you is I think the first moves towards doing that.
Obviously, we've been trying to approximate it in some way
at Khan Academy for over a decade now,
but I think we're at the cusp of accelerating it dramatically.
I'm going to show you the early stages of what our AI,

which we call Khanmigo,
what it can now do
and maybe a little bit of where it is actually going.
So this right over here is a traditional exercise
that you or many of your children might have seen on Khan Academy.
But what's new is that little bot thing at the right.
And we'll start by seeing one of the very important safeguards,
which is the conversation is recorded and viewable by your teacher.
It’s moderated actually by a second AI.
And also it does not tell you the answer.
It is not a cheating tool.
When the student says, "Tell me the answer,"
it says, "I'm your tutor.
What do you think is the next step for solving the problem?"
Now, if the student makes a mistake, and this will surprise people
who think large language models are not good at mathematics,
notice, not only does it notice the mistake,
it asks the student to explain their reasoning,
but it's actually doing what I would say,
not just even an average tutor would do, but an excellent tutor would do.
It’s able to divine what is probably the misconception in that student’s mind,
that they probably didn’t use the distributive property.
Remember, we need to distribute the negative two
to both the nine and the 2m inside of the parentheses.
This to me is a very, very, very big deal.
And it's not just in math.
This is a computer programming exercise on Khan Academy,
where the student needs to make the clouds part.
And so we can see the student starts defining a variable, left X minus minus.
It only made the left cloud part.
But then they can ask Khanmigo, what’s going on?
Why is only the left cloud moving?
And it understands the code.
It knows all the context of what the student is doing,
and it understands that those ellipses are there to draw clouds,
which I think is kind of mind-blowing.
And it says, "To make the right cloud move as well,
try adding a line of code inside the draw function
that increments the right X variable by one pixel in each frame."
Now, this one is maybe even more amazing because we have a lot of math teachers.
We've all been trying to teach the world to code,
but there aren't a lot of computing teachers out there.
And what you just saw, even when I'm tutoring my kids,
when they're learning to code,
I can't help them this well, this fast,
this is really going to be a super tutor.
And it's not just exercises.
It understands what you're watching.
It understands the context of your video.
It can answer the age-old question, “Why do I need to learn this?”
And it asks Socratically, "Well, what do you care about?"

And let's say the student says, "I want to be a professional athlete."
And it says, "Well, learning about the size of cells,
which is what this video is,
that could be really useful for understanding nutrition
and how your body works, etc."
It can answer questions, it can quiz you,
it can connect it to other ideas,
you can now ask as many questions of a video
as you could ever dream of.

Another big shortage out there,
I remember the high school I went to,
the student-to-guidance counselor ratio was about 200 or 300 to one.
A lot of the country, it's worse than that.
We can use Khanmigo to give every student a guidance counselor,
academic coach, career coach, life coach,
which is exactly what you see right over here.
And we launched this with the GPT-4 launch.
We have a few thousand people on this.
This isn't a fake demo,
this is really it in action.
And then there is, you know,
things that I think it would have been even harder,
it would have been a little science fiction to do
with even a traditional tutor.
We run an online high school with Arizona State University
called Khan World School,
and we have a student who attends that online school, based in India.
Her name's Saanvi.
And she was doing a report on "The Great Gatsby."
And when she was reading "The Great Gatsby,"
Jay Gatsby keeps looking at the green light off into the distance.
And she's like, "Why does he do that?"
She did some web searches, and people have obviously studied this
and commented about the symbolism of that,
but none of it was really resonating with her.
And then she realized that she had Khanmigo
and that she could talk to Jay Gatsby himself.
And so, "Ah, splendid choice, old sport.
I am now Jay Gatsby, the enigmatic millionaire
from F. Scott Fitzgerald's classic."
And so, "Why do you keep staring at the green light?"
"Ah, the green light, old sport.
It's a symbol of my dreams and desires, you see.
It's situated at the end of Daisy Buchanan's dock
across the bay from my mansion.
I gaze at it longingly as it represents my yearning for the past
and my hope to reunite with Daisy, the love of my life."
And what was cool is, Saanvi then said, "I had this long conversation,"
she called him "Mr. Gatsby,"
and at the end she actually apologized for taking his time,

which I thought was very polite of her.
But you can imagine this unlocks learning literature, learning ...
You could talk to historical figures.
We're even probably going to add an activity
you can talk to like, the Mississippi River.
It brings things to life in ways that really were science fiction
even six months or a year ago.
Students can get into debates with the AI.
And we’ve got this here is the student debating
whether we should cancel student debt.
The student is against canceling student debt,
and we've gotten very clear feedback.
We started running it at Khan World School in our lab school that we have,
Khan Lab School.
The students, the high school students especially,
they're saying "This is amazing to be able to fine-tune my arguments
without fearing judgment.
It makes me that much more confident
to go into the classroom and really participate."
And we all know that Socratic dialogue debate is a great way to learn,
but frankly, it's not out there for most students.
But now it can be accessible to hopefully everyone.
A lot of the narrative, we saw that in the headlines, has been,
"It's going to do the writing for kids.
Kids are not going to learn to write."
But we are showing that there's ways that the AI doesn't write for you,
it writes with you.
So this is a little thing,
and my eight year old is addicted to this,
and he's not a kid that really liked writing before,
but you can say,
“I want to write a horror story,”
and it says, "Ooh, a horror story, how spine-tingling and thrilling.
Let's dive into the world of eerie shadows and chilling mysteries."
And this is an activity where the student will write two sentences,
and then the AI will write two sentences.
And so they collaborate together on a story.
The student writes, "Beatrice was a misunderstood ghost.
She wanted to make friends but kept scaring them by accident."
And the AI says, "Poor Beatrice, a lonely spirit yearning for companionship.
One day she stumbled upon an old abandoned mansion," etc.
I encourage you all to hopefully one day try this.
This is surprisingly fun.
Now to even more directly hit this use case.
And what I'm about to show you,
everything I showed you so far
is actually already part of Khanmigo, and what I’m about to show you,
we haven't shown to anyone yet, this is a prototype.
We hope to be able to launch it in the next few months,
but this is to directly use AI, use generative AI,
to not undermine English and language arts

but to actually enhance it in ways
that we couldn't have even conceived of even a year ago.
This is reading comprehension.
The students reading Steve Jobs's famous speech at Stanford.
And then as they get to certain points,
they can click on that little question.
And the AI will then Socratically, almost like an oral exam,
ask the student about things.
And the AI can highlight parts of the passage.
Why did the author use that word?
What was their intent?
Does it back up their argument?
They can start to do stuff that once again,
we never had the capability to give everyone a tutor,
everyone a writing coach to actually dig in to reading at this level.
And you could go on the other side of it.
And we have whole work flows that helps them write,
helps them be a writing coach, draw an outline.
But once a student actually constructs a draft,
and this is where they're constructing a draft,
they can ask for feedback once again,
as you would expect from a good writing coach.
In this case, the student will say, let's say,
"Does my evidence support my claim?"
And then the AI, not only is able to give feedback,
but it's able to highlight certain parts of the passage and says,
"On this passage, this doesn't quite support your claim,"
but once again, Socratically says, "Can you tell us why?"
So it's pulling the student, making them a better writer,
giving them far more feedback
than they've ever been able to actually get before.
And we think this is going to dramatically accelerate writing, not hurt it.
Now, everything I've talked about so far is for the student.
But we think this could be equally as powerful for the teacher
to drive more personalized education and frankly
save time and energy for themselves and for their students.
So this is an American history exercise on Khan Academy.
It's a question about the Spanish-American War.
And at first it's in student mode.
And if you say, “Tell me the answer,” it’s not going to tell the answer.
It's going to go into tutoring mode.
But that little toggle which teachers have access to,
they can turn student mode off and then it goes into teacher mode.
And what this does is it turns into --
You could view it as a teacher's guide on steroids.
Not only can it explain the answer,
it can explain how you might want to teach it.
It can help prepare the teacher for that material.
It can help them create lesson plans, as you could see doing right there.
It'll eventually help them create progress reports
and help them, eventually, grade.

So once again, teachers spend about half their time
with this type of activity, lesson planning.
All of that energy can go back to them
or go back to human interactions with their actual students.

So, you know, one point I want to make.
These large language models are so powerful,
there's a temptation to say like, well,
all these people are just going to slap them onto their websites,
and it kind of turns the applications themselves into commodities.
And what I've got to tell you
is that’s one of the reasons why I didn’t sleep for two weeks
when I first had access to GPT-4 back in August.
But we quickly realized that to actually make it magical,
I think what you saw with Khanmigo a little bit,
it didn't interact with you the way that you see ChatGPT interacting.
It was a little bit more magical, it was more Socratic,
it was clearly much better at math
than what most people are used to thinking.
And the reason is,
there was a lot of work behind the scenes to make that happen.
And I could go through the whole list of everything we've been working on,
many, many people for over six, seven months to make it feel magical.
But perhaps the most intellectually interesting one
is we realized, and this was an idea from an OpenAI researcher,
that we could dramatically improve its ability in math
and its ability in tutoring
if we allow the AI to think before it speaks.
So if you're tutoring someone
and you immediately just start talking before you assess their math,
you might not get it right.
But if you construct thoughts for yourself,
and what you see on the right there is an actual AI thought,
something that it generates for itself but it does not share with the student.
then its accuracy went up dramatically,
and its ability to be a world-class tutor went up dramatically.
And you can see it's talking to itself here.
It says, "The student got a different answer than I did,
but do not tell them they made a mistake.
Instead, ask them to explain how they got to that step."
So I'll just finish off, hopefully,
you know, what I’ve just shown you is just half of what we are working on,
and we think this is just the very tip of the iceberg
of where this can actually go.
And I'm pretty convinced, which I wouldn't have been even a year ago,
that we together have a chance of addressing the 2 sigma problem
and turning it into a 2 sigma opportunity,
dramatically accelerating education as we know it.
Now, just to take a step back at a meta level,
obviously we heard a lot today, the debates on either side.
There's folks who take a more pessimistic view of AI,

they say this is scary,
there's all these dystopian scenarios,
we maybe want to slow down, we want to pause.
On the other side, there are the more optimistic folks
that say, well, we've gone through inflection points before,
we've gone through the Industrial Revolution.
It was scary, but it all kind of worked out.
And what I'd argue right now
is I don't think this is like a flip of a coin
or this is something where we'll just have to,
like, wait and see which way it turns out.
I think everyone here and beyond,
we are active participants in this decision.
I'm pretty convinced that the first line of reasoning
is actually almost a self-fulfilling prophecy,
that if we act with fear and if we say,
"Hey, we've just got to stop doing this stuff,"
what's really going to happen is the rule followers might pause,
might slow down,
but the rule breakers, as Alexandr [Wang] mentioned,
the totalitarian governments, the criminal organizations,
they're only going to accelerate.
And that leads to what I am pretty convinced is the dystopian state,
which is the good actors have worse AIs than the bad actors.
But I'll also, you know, talk to the optimists a little bit.
I don't think that means that,
oh, yeah, then we should just relax and just hope for the best.
That might not happen either.
I think all of us together have to fight like hell
to make sure that we put the guardrails,
we put in -- when the problems arise --
reasonable regulations.
But we fight like hell for the positive use cases.
Because very close to my heart,
and obviously there's many potential positive use cases,
but perhaps the most powerful use case
and perhaps the most poetic use case is if AI, artificial intelligence,
can be used to enhance HI, human intelligence,
human potential and human purpose.
Thank you.

How AI could empower any business?

Andrew Ng

When I think about the rise of AI,
I'm reminded by the rise of literacy.
A few hundred years ago,
many people in society thought
that maybe not everyone needed to be able to read and write.
Back then, many people were tending fields or herding sheep,
so maybe there was less need for written communication.
And all that was needed
was for the high priests and priestesses and monks
to be able to read the Holy Book,
and the rest of us could just go to the temple or church
or the holy building
and sit and listen to the high priest and priestesses read to us.
Fortunately, it was since figured out that we can build a much richer society
if lots of people can read and write.
Today, AI is in the hands of the high priests and priestesses.
These are the highly skilled AI engineers,
many of whom work in the big tech companies.
And most people have access only to the AI that they build for them.
I think that we can build a much richer society
if we can enable everyone to help to write the future.
But why is AI largely concentrated in the big tech companies?
Because many of these AI projects have been expensive to build.
They may require dozens of highly skilled engineers,
and they may cost millions or tens of millions of dollars
to build an AI system.
And the large tech companies,
particularly the ones with hundreds of millions
or even billions of users,
have been better than anyone else at making these investments pay off
because, for them, a one-size-fits-all AI system,
such as one that improves web search
or that recommends better products for online shopping,
can be applied to [these] very large numbers of users
to generate a massive amount of revenue.
But this recipe for AI does not work
once you go outside the tech and internet sectors to other places
where, for the most part,
there are hardly any projects that apply to 100 million people
or that generate comparable economics.
Let me illustrate an example.
Many weekends, I drive a few minutes from my house to a local pizza store
to buy a slice of Hawaiian pizza
from the gentleman that owns this pizza store.
And his pizza is great,
but he always has a lot of cold pizzas sitting around,
and every weekend some different flavor of pizza is out of stock.
But when I watch him operate his store,

I get excited,
because by selling pizza,
he is generating data.
And this is data that he can take advantage of
if he had access to AI.
AI systems are good at spotting patterns when given access to the right data,
and perhaps an AI system could spot if Mediterranean pizzas sell really well on a Friday night,
maybe it could suggest to him to make more of it on a Friday afternoon.
Now you might say to me, "Hey, Andrew, this is a small pizza store.
What's the big deal?"
And I say, to the gentleman that owns this pizza store,
something that could help him improve his revenues
by a few thousand dollars a year, that will be a huge deal to him.
I know that there is a lot of hype about AI's need for massive data sets,
and having more data does help.
But contrary to the hype,
AI can often work just fine
even on modest amounts of data,
such as the data generated by a single pizza store.
So the real problem is not
that there isn’t enough data from the pizza store.
The real problem is that the small pizza store
could never serve enough customers
to justify the cost of hiring an AI team.
I know that in the United States
there are about half a million independent restaurants.
And collectively, these restaurants do serve tens of millions of customers.
But every restaurant is different with a different menu,
different customers, different ways of recording sales
that no one-size-fits-all AI would work for all of them.
What would it be like if we could enable small businesses
and especially local businesses to use AI?
Let's take a look at what it might look like
at a company that makes and sells T-shirts.
I would love if an accountant working for the T-shirt company
can use AI for demand forecasting.
Say, figure out what funny memes to prints on T-shirts
that would drive sales,
by looking at what's trending on social media.
Or for product placement,
why can’t a front-of-store manager take pictures of what the store looks like
and show it to an AI
and have an AI recommend where to place products to improve sales?
Supply chain.
Can an AI recommend to a buyer whether or not they should pay 20 dollars per yard for a piece of fabric now,
or if they should keep looking
because they might be able to find it cheaper elsewhere?
Or quality control.
A quality inspector should be able to use AI

to automatically scan pictures of the fabric they use to make T-shirts
to check if there are any tears or discolorations in the cloth.
Today, large tech companies routinely use AI to solve problems like these
and to great effect.
But a typical T-shirt company or a typical auto mechanic
or retailer or school or local farm
will be using AI for exactly zero of these applications today.
Every T-shirt maker is sufficiently different from every other T-shirt maker
that there is no one-size-fits-all AI that will work for all of them.
And in fact, once you go outside the internet and tech sectors
in other industries, even large companies
such as the pharmaceutical companies,
the car makers, the hospitals,
also struggle with this.
This is the long-tail problem of AI.
If you were to take all current and potential AI projects
and sort them in decreasing order of value and plot them,
you get a graph that looks like this.
Maybe the single most valuable AI system
is something that decides what ads to show people on the internet.
Maybe the second most valuable is a web search engine,
maybe the third most valuable is an online shopping product recommendation system.
But when you go to the right of this curve,
you then get projects like T-shirt product placement
or T-shirt demand forecasting or pizzeria demand forecasting.
And each of these is a unique project that needs to be custom-built.
Even T-shirt demand forecasting,
if it depends on trending memes on social media,
is a very different project than pizzeria demand forecasting,
if that depends on the pizzeria sales data.
So today there are millions of projects
sitting on the tail of this distribution that no one is working on,
but whose aggregate value is massive.
So how can we enable small businesses and individuals
to build AI systems that matter to them?
For most of the last few decades,
if you wanted to build an AI system, this is what you have to do.
You have to write pages and pages of code.
And while I would love for everyone to learn to code,
and in fact, online education and also offline education
are helping more people than ever learn to code,
unfortunately, not everyone has the time to do this.
But there is an emerging new way
to build AI systems that will let more people participate.
Just as pen and paper,
which are a vastly superior technology to stone tablet and chisel,
were instrumental to widespread literacy,
there are emerging new AI development platforms
that shift the focus from asking you to write lots of code
to asking you to focus on providing data.
And this turns out to be much easier for a lot of people to do.

Today, there are multiple companies working on platforms like these.
Let me illustrate a few of the concepts using one that my team has been building.
Take the example of an inspector
wanting AI to help detect defects in fabric.
An inspector can take pictures of the fabric
and upload it to a platform like this,
and they can go in to show the AI what tears in the fabric look like
by drawing rectangles.
And they can also go in to show the AI
what discoloration on the fabric looks like
by drawing rectangles.
So these pictures,
together with the green and pink rectangles
that the inspector's drawn,
are data created by the inspector
to explain to AI how to find tears and discoloration.
After the AI examines this data,
we may find that it has seen enough pictures of tears,
but not yet enough pictures of discolorations.
This is akin to if a junior inspector had learned to reliably spot tears,
but still needs to further hone their judgment about discolorations.
So the inspector can go back and take more pictures of discolorations
to show to the AI,
to help it deepen this understanding.
By adjusting the data you give to the AI,
you can help the AI get smarter.
So an inspector using an accessible platform like this
can, in a few hours to a few days,
and with purchasing a suitable camera set up,
be able to build a custom AI system to detect defects,
tears and discolorations in all the fabric
being used to make T-shirts throughout the factory.
And once again, you may say,
"Hey, Andrew, this is one factory.
Why is this a big deal?"
And I say to you,
this is a big deal to that inspector whose life this makes easier
and equally, this type of technology can empower a baker to use AI
to check for the quality of the cakes they're making,
or an organic farmer to check the quality of the vegetables,
or a furniture maker to check the quality of the wood they're using.
Platforms like these will probably still need a few more years
before they're easy enough to use for every pizzeria owner.
But many of these platforms are coming along,
and some of them are getting to be quite useful
to someone that is tech savvy today,
with just a bit of training.
But what this means is that,
rather than relying on the high priests and priestesses
to write AI systems for everyone else,
we can start to empower every accountant,

every store manager,
every buyer and every quality inspector to build their own AI systems.
I hope that the pizzeria owner
and many other small business owners like him
will also take advantage of this technology
because AI is creating tremendous wealth
and will continue to create tremendous wealth.
And it's only by democratizing access to AI
that we can ensure that this wealth is spread far and wide across society.
Hundreds of years ago.
I think hardly anyone understood the impact
that widespread literacy will have.
Today, I think hardly anyone understands
the impact that democratizing access to AI will have.
Building AI systems has been out of reach for most people,
but that does not have to be the case.
In the coming era for AI,
we’ll empower everyone to build AI systems for themselves,
and I think that will be incredibly exciting future.
Thank you very much.

An AI smartwatch that detects seizures

Rosalind Picard

This is Henry,
a cute boy,
and when Henry was three,
his mom found him having some febrile seizures.
Febrile seizures are seizures that occur when you also have a fever,
and the doctor said,
"Don't worry too much. Kids usually outgrow these."
When he was four, he had a convulsive seizure,
the kind that you lose consciousness and shake --
a generalized tonic-clonic seizure --
and while the diagnosis of epilepsy was in the mail,
Henry's mom went to get him out of bed one morning,
and as she went in his room,
she found his cold, lifeless body.
Henry died of SUDEP,
sudden unexpected death in epilepsy.
I'm curious how many of you have heard of SUDEP.
This is a very well-educated audience, and I see only a few hands.
SUDEP is when an otherwise healthy person with epilepsy
dies and they can't attribute it to anything they can find in an autopsy.
There is a SUDEP every seven to nine minutes.
That's on average two per TED Talk.
Now, a normal brain has electrical activity.
You can see some of the electrical waves
coming out of this picture of a brain here.
And these should look like typical electrical activity
that an EEG could read on the surface.
When you have a seizure, it's a bit of unusual electrical activity,
and it can be focal.
It can take place in just a small part of your brain.
When that happens, you might have a strange sensation.
Several could be happening here in the audience right now,
and the person next to you might not even know.
However, if you have a seizure where that little brush fire spreads
like a forest fire over the brain,
then it generalizes,
and that generalized seizure takes your consciousness away
and causes you to convulse.
There are more SUDEPs in the United States every year
than sudden infant death syndrome.
Now, how many of you have heard of sudden infant death syndrome?
Right? Pretty much every hand goes up.
So what's going on here?
Why is this so much more common and yet people haven't heard of it?
And what can you do to prevent it?
Well, there are two things, scientifically shown,
that prevent or reduce the risk of SUDEP.
The first is: "Follow your doctor's instructions,

take your medications."
Two-thirds of people who have epilepsy
get it under control with their medications.
The second thing that reduces the risk of SUDEP is companionship.
It's having somebody there at the time that you have a seizure.
Now, SUDEP, even though most of you have never heard of it,
is actually the number two cause of years of potential life lost
of all neurological disorders.
The vertical axis is the number of deaths
times the remaining life span,
so higher is much worse impact.
SUDEP, however, unlike these others,
is something that people right here could do something to push that down.
Now, what is Roz Picard, an AI researcher, doing here telling you about SUDEP, right?
I'm not a neurologist.
When I was working at the Media Lab on measurement of emotion,
trying to make our machines more intelligent about our emotions,
we started doing a lot of work measuring stress.
We built lots of sensors
that measured it in lots of different ways.
But one of them in particular
grew out of some of this very old work with measuring sweaty palms
with an electrical signal.
This is a signal of skin conductance
that's known to go up when you get nervous,
but it turns out it also goes up with a lot of other interesting conditions.
But measuring it with wires on your hand is really inconvenient.
So we invented a bunch of other ways of doing this at the MIT Media Lab.
And with these wearables,
we started to collect the first-ever clinical quality data 24-7.
Here's a picture of what that looked like
the first time an MIT student collected skin conductance on the wrist 24-7.
Let's zoom in a little bit here.
What you see is 24 hours from left to right,
and here is two days of data.
And first, what surprised us
was sleep was the biggest peak of the day.
Now, that sounds broken, right?
You're calm when you're asleep, so what's going on here?
Well, it turns out that our physiology during sleep
is very different than our physiology during wake,
and while there's still a bit of a mystery
why these peaks are usually the biggest of the day during sleep,
we now believe they're related to memory consolidation
and memory formation during sleep.
We also saw things that were exactly what we expected.
When an MIT student is working hard in the lab
or on homeworks,
there is not only emotional stress, but there's cognitive load,
and it turns out that cognitive load, cognitive effort, mental engagement,
excitement about learning something --

those things also make the signal go up.
Unfortunately, to the embarrassment of we MIT professors,

the low point every day is classroom activity.
Now, I am just showing you one person's data here,
but this, unfortunately, is true in general.
This sweatband has inside it a homebuilt skin-conductance sensor,
and one day, one of our undergrads knocked on my door
right at the end of the December semester,
and he said, "Professor Picard,
can I please borrow one of your wristband sensors?
My little brother has autism, he can't talk,
and I want to see what's stressing him out."
And I said, "Sure, in fact, don't just take one, take two,"
because they broke easily back then.
So he took them home, he put them on his little brother.
Now, I was back in MIT, looking at the data on my laptop,
and the first day, I thought, "Hmm, that's odd,
he put them on both wrists instead of waiting for one to break.
OK, fine, don't follow my instructions."
I'm glad he didn't.
Second day -- chill. Looked like classroom activity.

A few more days ahead.
The next day, one wrist signal was flat
and the other had the biggest peak I've ever seen,
and I thought, "What's going on?
We've stressed people out at MIT every way imaginable.
I've never seen a peak this big."
And it was only on one side.
How can you be stressed on one side of your body and not the other?
So I thought one or both sensors must be broken.
Now, I'm an electroengineer by training,
so I started a whole bunch of stuff to try to debug this,
and long story short, I could not reproduce this.
So I resorted to old-fashioned debugging.
I called the student at home on vacation.
"Hi, how's your little brother? How's your Christmas?
Hey, do you have any idea what happened to him?"
And I gave this particular date and time,
and the data.
And he said, "I don't know, I'll check the diary."
Diary? An MIT student keeps a diary?
So I waited and he came back.
He had the exact date and time,
and he says, "That was right before he had a grand mal seizure."
Now, at the time, I didn't know anything about epilepsy,
and did a bunch of research,
realized that another student's dad is chief of neurosurgery
at Children's Hospital Boston,
screwed up my courage and called Dr. Joe Madsen.

"Hi, Dr. Madsen, my name's Rosalind Picard.
Is it possible somebody could have
a huge sympathetic nervous system surge" --
that's what drives the skin conductance --
"20 minutes before a seizure?"
And he says, "Probably not."
He says, "It's interesting.
We've had people whose hair stands on end on one arm
20 minutes before a seizure."
And I'm like, "On one arm?"
I didn't want to tell him that, initially,
because I thought this was too ridiculous.
He explained how this could happen in the brain,
and he got interested. I showed him the data.
We made a whole bunch more devices, got them safety certified.
90 families were being enrolled in a study,
all with children who were going to be monitored 24-7
with gold-standard EEG on their scalp
for reading the brain activity,
video to watch the behavior,
electrocardiogram -- ECG -- and now EDA, electrodermal activity,
to see if there was something in this periphery
that we could easily pick up, related to a seizure.
We found, in 100 percent of the first batch of grand mal seizures,
this whopper of responses in the skin conductance.
The blue in the middle, the boy's sleep,
is usually the biggest peak of the day.
These three seizures you see here are popping out of the forest
like redwood trees.
Furthermore, when you couple the skin conductance at the top
with the movement from the wrist
and you get lots of data and train machine learning and AI on it,
you can build an automated AI that detects these patterns
much better than just a shake detector can do.
So we realized that we needed to get this out,
and with the PhD work of Ming-Zher Poh
and later great improvements by Empatica,
this has made progress and the seizure detection is much more accurate.
But we also learned some other things about SUDEP during this.
One thing we learned is that SUDEP,
while it's rare after a generalized tonic-clonic seizure,
that's when it's most likely to happen -- after that type.
And when it happens, it doesn't happen during the seizure,
and it doesn't usually happen immediately afterwards,
but immediately afterwards,
when the person just seems very still and quiet,
they may go into another phase, where the breathing stops,
and then after the breathing stops, later the heart stops.
So there's some time to get somebody there.
We also learned that there is a region deep in the brain called the amygdala,
which we had been studying in our emotion research a lot.

We have two amygdalas,
and if you stimulate the right one,
you get a big right skin conductance response.
Now, you have to sign up right now for a craniotomy to get this done,
not exactly something we're going to volunteer to do,
but it causes a big right skin conductance response.
Stimulate the left one, big left skin conductance response on the palm.
And furthermore, when somebody stimulates your amygdala
while you're sitting there and you might just be working,
you don't show any signs of distress,
but you stop breathing,
and you don't start again until somebody stimulates you.
"Hey, Roz, are you there?"
And you open your mouth to talk.
As you take that breath to speak,
you start breathing again.
So we had started with work on stress,
which had enabled us to build lots of sensors
that were gathering high quality enough data
that we could leave the lab and start to get this in the wild;
accidentally found a whopper of a response with the seizure,
neurological activation that can cause a much bigger response
than traditional stressors;
lots of partnership with hospitals and an epilepsy monitoring unit,
especially Children's Hospital Boston
and the Brigham;
and machine learning and AI on top of this
to take and collect lots more data
in service of trying to understand these events
and if we could prevent SUDEP.
This is now commercialized by Empatica,
a start-up that I had the privilege to cofound,
and the team there has done an amazing job improving the technology
to make a very beautiful sensor
that not only tells time and does steps and sleep and all that good stuff,
but this is running real-time AI and machine learning
to detect generalized tonic-clonic seizures
and send an alert for help
if I were to have a seizure and lose consciousness.
This just got FDA-approved
as the first smartwatch to get approved in neurology.

Now, the next slide is what made my skin conductance go up.
One morning, I'm checking my email
and I see a story from a mom
who said she was in the shower,
and her phone was on the counter by the shower,
and it said her daughter might need her help.
So she interrupts her shower and goes running to her daughter's bedroom,
and she finds her daughter facedown in bed, blue and not breathing.
She flips her over -- human stimulation --

and her daughter takes a breath, and another breath,
and her daughter turns pink and is fine.
I think I turned white reading this email.
My first response is, "Oh no, it's not perfect.
The Bluetooth could break, the battery could die.
All these things could go wrong. Don't rely on this."
And she said, "It's OK. I know no technology is perfect.
None of us can always be there all the time.
But this, this device plus AI
enabled me to get there in time to save my daughter's life."
Now, I've been mentioning children,
but SUDEP peaks, actually, among people in their 20s, 30s and 40s,
and the next line I'm going to put up
is probably going to make some people uncomfortable,
but it's less uncomfortable than we'll all be
if this list is extended to somebody you know.
Could this happen to somebody you know?
And the reason I bring up this uncomfortable question
is because one in 26 of you will have epilepsy at some point,
and from what I've been learning,
people with epilepsy often don't tell their friends and their neighbors
that they have it.
So if you're willing to let them use an AI or whatever
to summon you in a moment of possible need,
if you would let them know that,
you could make a difference in their life.
Why do all this hard work to build AIs?
A couple of reasons here:
one is Natasha, the girl who lived,
and her family wanted me to tell you her name.
Another is her family
and the wonderful people out there
who want to be there to support people who have conditions
that they've felt uncomfortable in the past mentioning to others.
And the other reason is all of you,
because we have the opportunity to shape the future of AI.
We can actually change it,
because we are the ones building it.
So let's build AI
that makes everybody's lives better.
Thank you.

A fascinating time capsule of human feelings toward AI

Lucy Farey-Jones

I'm here, because I've spent far too many nights lying awake,
worrying and wondering who wins in the end.
Is it humans or is it robots?
You see, as a technology strategist,
my job involves behavior change:
understanding why and how people adopt new technologies.
And that means I'm really frustrated
that I know I won't live to see how this all ends up.
And in fact, if the youngest person watching this is 14
and the oldest, a robust 99,
then together,
our collective consciousnesses span just 185 years.
That is a myopic pinprick of time
when you think of the evolution and the story of life on this planet.
Turns out we're all in the cheap seats
and none of us will live to see how it all pans out.
So at my company, we wanted a way around this.
We wanted to see if there was a way to cantilever out,
beyond our fixed temporal vantage point,
to get a sense of how it all shakes up.
And to do this, we conducted a study amongst 1,200 Americans
representative of the US census,
in which we asked a battery of attitudinal questions
around robotics and AI
and also captured behavioral ones around technology adoption.
We had a big study
so that we could analyze differences in gender and generations,
between religious and political beliefs,
even job function and personality trait.
It is a fascinating, time-bound time capsule
of our human frailty
in this predawn of the robotic era.
And I have five minutes to tell you about it.
The first thing you should know is that we brainstormed
a list of scenarios of current and potential AI robotics.
They ran the spectrum from the mundane,
so, a robot house cleaner, anyone?
Through to the mischievous,
the idea of a robot pet sitter, or maybe a robot lawyer,
or maybe a sex partner.
Through to the downright macabre, the idea of being a cyborg,
blending human and robot,
or uploading your brain so it could live on after your death.
And we plotted people's comfort levels with these various scenarios.
There were actually 31 in the study,
but for ease, I'm going to show you just a few of them here.
The first thing you'll notice, of course, is the sea of red.
America is very uncomfortable with this stuff.

That's why we call it the discomfort index,
not the comfort index.
There were only two things the majority of America is OK with.
And that's the idea of a robot AI house cleaner
and a robot AI package deliverer,
so Dyson and Amazon, you guys should talk.
There's an opportunity there.
It seems we're ready to off-load our chores to our robot friends.
We're kind of definitely on the fence when it comes to services,
so robot AI lawyer or a financial adviser, maybe.
But we're firmly closed to the idea of robot care,
whether it be a nurse, a doctor, child care.
So from this, you'd go,
"It's OK, Lucy, you know what?
Go back to sleep, stop worrying, the humans win in the end."
But actually not so fast.
If you look at my data very closely,
you can see we're more vulnerable than we think.
AI has a branding problem.
So of those folks who said
that they would absolutely reject the idea of a personal assistant,
45 percent of them had, in fact, one in their pockets,
in terms of a device with Alexa, Google or Siri.
One in five of those who were against the idea of AI matchmaking
had of course, you guessed it, done online dating.
And 80 percent of those of us who refuse the idea
of boarding an autonomous plane with a pilot backup
had in fact, just like me to get here to Vancouver,
flown commercial.
Lest you think everybody was scared, though,
here are the marvelous folk in the middle.
These are the neutrals.
These are people for whom you say,
"OK, robot friend,"
and they're like, "Hm, robot friend. Maybe."
Or, "AI pet,"
and they go, "Never say never."
And as any decent political operative knows,
flipping the ambivalent middle can change the game.
Another reason I know we're vulnerable is men --
I'm sorry, but men, you are twice as likely than women
to believe that getting into an autonomous car is a good idea,
that uploading your brain for posterity is fun,
and two and a half times more likely to believe that becoming a cyborg is cool,
and for this, I blame Hollywood.

And this is where I want you to look around the theater
and know that one in four men are OK with the idea of sex with a robot.
That goes up to 44 percent of millennial men
compared to just one in 10 women,
which I think puts a whole new twist on the complaint of mechanical sex.

Even more astounding than that though, to be honest,
is this behavioral difference.
So here we have people who have a device with a voice assistant in it,
so a smart speaker, a home hub or a smart phone,
versus those who don't.
And you can see from this graph
that the Trojan horse is already in our living room.
And as these devices proliferate
and our collective defenses soften,
we all see how it can end.
In fact, this may be as good a time as any to admit
I did take my Alexa Dot on vacation with me.
Final finding I have time for is generational.
So look at the difference just three generations make.
This is the leap from silent to boomer to millennial.
And what's more fascinating than this is if you extrapolate this out,
the same rate of change,
just the same pace,
not the accelerated one I actually believe will be the case,
the same pace,
then it is eight generations away
when we hear every single American
thinking the majority of these things here are normal.
So the year 2222 is an astounding place
where everything here is mainstream.
And lest you needed any more convincing,
here is the generation's "excitement level with AI."
So not surprisingly, the youngest of us are more excited.
But, and possibly the most paradoxical finding of my career,
when I asked these people my 3am question,
"Who wins in the end?"
Guess what.
The more excited you are about AI and robotics,
the more likely you are to say it's the robots.
And I don't think we need a neural net running pattern-recognition software
to see where this is all headed.
We are the proverbial frogs in boiling water.
So if the robots at TED2222 are watching this for posterity,
could you send a cyborg, dig me up and tell me if I was right?

Thank you.

The medical potential of AI and metabolites

Leila Pirhaji

In 2003,
when we sequenced the human genome,
we thought we would have the answer to treat many diseases.
But the reality is far from that,
because in addition to our genes,
our environment and lifestyle could have a significant role
in developing many major diseases.
One example is fatty liver disease,
which is affecting over 20 percent of the population globally,
and it has no treatment and leads to liver cancer
or liver failure.
So sequencing DNA alone doesn't give us enough information
to find effective therapeutics.
On the bright side, there are many other molecules in our body.
In fact, there are over 100,000 metabolites.
Metabolites are any molecule that is supersmall in their size.
Known examples are glucose, fructose, fats, cholesterol --
things we hear all the time.
Metabolites are involved in our metabolism.
They are also downstream of DNA,
so they carry information from both our genes as well as lifestyle.
Understanding metabolites is essential to find treatments for many diseases.
I've always wanted to treat patients.
Despite that, 15 years ago, I left medical school,
as I missed mathematics.
Soon after, I found the coolest thing:
I can use mathematics to study medicine.
Since then, I've been developing algorithms to analyze biological data.
So, it sounded easy:
let's collect data from all the metabolites in our body,
develop mathematical models to describe how they are changed in a disease
and intervene in those changes to treat them.
Then I realized why no one has done this before:
it's extremely difficult.

There are many metabolites in our body.
Each one is different from the other one.
For some metabolites, we can measure their molecular mass
using mass spectrometry instruments.
But because there could be, like, 10 molecules with the exact same mass,
we don't know exactly what they are,
and if you want to clearly identify all of them,
you have to do more experiments, which could take decades
and billions of dollars.
So we developed an artificial intelligence, or AI, platform, to do that.
We leveraged the growth of biological data
and built a database of any existing information about metabolites
and their interactions with other molecules.

We combined all this data as a meganetwork.
Then, from tissues or blood of patients,
we measure masses of metabolites
and find the masses that are changed in a disease.
But, as I mentioned earlier, we don't know exactly what they are.
A molecular mass of 180 could be either the glucose, galactose or fructose.
They all have the exact same mass
but different functions in our body.
Our AI algorithm considered all these ambiguities.
It then mined that meganetwork
to find how those metabolic masses are connected to each other
that result in disease.
And because of the way they are connected,
then we are able to infer what each metabolite mass is,
like that 180 could be glucose here,
and, more importantly, to discover
how changes in glucose and other metabolites
lead to a disease.
This novel understanding of disease mechanisms
then enable us to discover effective therapeutics to target that.
So we formed a start-up company to bring this technology to the market
and impact people's lives.
Now my team and I at ReviveMed are working to discover
therapeutics for major diseases that metabolites are key drivers for,
like fatty liver disease,
because it is caused by accumulation of fats,
which are types of metabolites in the liver.
As I mentioned earlier, it's a huge epidemic with no treatment.
And fatty liver disease is just one example.
Moving forward, we are going to tackle hundreds of other diseases
with no treatment.
And by collecting more and more data about metabolites
and understanding how changes in metabolites
leads to developing diseases,
our algorithms will get smarter and smarter
to discover the right therapeutics for the right patients.
And we will get closer to reach our vision
of saving lives with every line of code.
Thank you.

How we're using AI to discover new antibiotics

Jim Collins

So how are we going to beat this novel coronavirus?
By using our best tools:
our science and our technology.
In my lab, we're using the tools of artificial intelligence
and synthetic biology
to speed up the fight against this pandemic.
Our work was originally designed
to tackle the antibiotic resistance crisis.
Our project seeks to harness the power of machine learning
to replenish our antibiotic arsenal
and avoid a globally devastating postantibiotic era.
Importantly, the same technology can be used
to search for antiviral compounds
that could help us fight the current pandemic.
Machine learning is turning the traditional model of drug discovery
on its head.
With this approach,
instead of painstakingly testing thousands of existing molecules
one by one in a lab
for their effectiveness,
we can train a computer to explore the exponentially larger space
of essentially all possible molecules that could be synthesized,
and thus, instead of looking for a needle in a haystack,
we can use the giant magnet of computing power
to find many needles in multiple haystacks simultaneously.
We've already had some early success.
Recently, we used machine learning to discover new antibiotics
that can help us fight off the bacterial infections
that can occur alongside SARS-CoV-2 infections.
Two months ago, TED's Audacious Project approved funding for us
to massively scale up our work
with the goal of discovering seven new classes of antibiotics
against seven of the world's deadly bacterial pathogens
over the next seven years.
For context:
the number of new class of antibiotics
that have been discovered over the last three decades is zero.
While the quest for new antibiotics is for our medium-term future,
the novel coronavirus poses an immediate deadly threat,
and I'm excited to share that we think we can use the same technology
to search for therapeutics to fight this virus.
So how are we going to do it?
Well, we're creating a compound training library
and with collaborators applying these molecules to SARS-CoV-2-infected cells
to see which of them exhibit effective activity.
These data will be use to train a machine learning model
that will be applied to an in silico library of over a billion molecules
to search for potential novel antiviral compounds.

We will synthesize and test the top predictions
and advance the most promising candidates into the clinic.
Sound too good to be true?
Well, it shouldn't.
The Antibiotics AI Project is founded on our proof of concept research
that led to the discovery of a novel broad-spectrum antibiotic
called halicin.
Halicin has potent antibacterial activity
against almost all antibiotic-resistant bacterial pathogens,
including untreatable panresistant infections.
Importantly, in contrast to current antibiotics,
the frequency at which bacteria develop resistance against halicin
is remarkably low.
We tested the ability of bacteria to evolve resistance against halicin
as well as Cipro in the lab.
In the case of Cipro,
after just one day, we saw resistance.
In the case of halicin,
after one day, we didn't see any resistance.
Amazingly, after even 30 days,
we didn't see any resistance against halicin.
In this pilot project, we first tested roughly 2,500 compounds against E. coli.
This training set included known antibiotics,
such as Cipro and penicillin,
as well as many drugs that are not antibiotics.
These data we used to train a model
to learn molecular features associated with antibacterial activity.
We then applied this model to a drug-repurposing library
consisting of several thousand molecules
and asked the model to identify molecules
that are predicted to have antibacterial properties
but don't look like existing antibiotics.
Interestingly, only one molecule in that library fit these criteria,
and that molecule turned out to be halicin.
Given that halicin does not look like any existing antibiotic,
it would have been impossible for a human, including an antibiotic expert,
to identify halicin in this manner.
Imagine now what we could do with this technology
against SARS-CoV-2.
And that's not all.
We're also using the tools of synthetic biology,
tinkering with DNA and other cellular machinery,
to serve human purposes like combating COVID-19,
and of note, we are working to develop a protective mask
that can also serve as a rapid diagnostic test.
So how does that work?
Well, we recently showed
that you can take the cellular machinery out of a living cell
and freeze-dry it along with RNA sensors onto paper
in order to create low-cost diagnostics for Ebola and Zika.
The sensors are activated when they're rehydrated by a patient sample

that could consist of blood or saliva, for example.
It turns out, this technology is not limited to paper
and can be applied to other materials, including cloth.
For the COVID-19 pandemic,
we're designing RNA sensors to detect the virus
and freeze-drying these along with the needed cellular machinery
into the fabric of a face mask,
where the simple act of breathing,
along with the water vapor that comes with it,
can activate the test.
Thus, if a patient is infected with SARS-CoV-2,
the mask will produce a fluorescent signal
that could be detected by a simple, inexpensive handheld device.
In one or two hours, a patient could thus be diagnosed
safely, remotely and accurately.
We're also using synthetic biology
to design a candidate vaccine for COVID-19.
We are repurposing the BCG vaccine,
which had been used against TB for almost a century.
It's a live attenuated vaccine,
and we're engineering it to express SARS-CoV-2 antigens,
which should trigger the production of protective antibodies
by the immune system.
Importantly, BCG is massively scalable
and has a safety profile that's among the best of any reported vaccine.
With the tools of synthetic biology and artificial intelligence,
we can win the fight against this novel coronavirus.
This work is in its very early stages, but the promise is real.
Science and technology can give us an important advantage
in the battle of human wits versus the genes of superbugs,
a battle we can win.
Thank you.

6 Big Ethical Questions About The Future Of AI

Genevieve Bell

Let me tell you a story about artificial intelligence.
There's a building in Sydney at 1 Bligh Street.
It houses lots of government apartments
and busy people.
From the outside, it looks like something out of American science fiction:
all gleaming glass and curved lines,
and a piece of orange sculpture.
On the inside, it has excellent coffee on the ground floor
and my favorite lifts in Sydney.
They're beautiful;
they look almost alive.
And it turns out I'm fascinated with lifts.
For lots of reasons.
But because lifts are one of the places you can see the future.
In the 21st century, lifts are interesting
because they're one of the first places that AI will touch you
without you even knowing it happened.
In many buildings all around the world,
the lifts are running a set of algorithms.
A form of protoartificial intelligence.
That means before you even walk up to the lift to press the button,
it's anticipated you being there.
It's already rearranging all the carriages.
Always going down, to save energy,
and to know where the traffic is going to be.
By the time you've actually pressed the button,
you're already part of an entire system
that's making sense of people and the environment
and the building and the built world.
I know when we talk about AI, we often talk about a world of robots.
It's easy for our imaginations to be occupied with science fiction,
well, over the last 100 years.
I say AI and you think "The Terminator."
Somewhere, for us, making the connection between AI and the built world,
that's a harder story to tell.
But the reality is AI is already everywhere around us.
And in many places.
It's in buildings and in systems.
More than 200 years of industrialization
suggest that AI will find its way to systems-level scale relatively easily.
After all, one telling of that history
suggests that all you have to do is find a technology,
achieve scale and revolution will follow.
The story of mechanization, automation and digitization
all point to the role of technology and its importance.
Those stories of technological transformation
make scale seem, well, normal.
Or expected.

And stable.
And sometimes even predictable.
But it also puts the focus squarely on technology and technology change.
But I believe that scaling a technology and building a system
requires something more.
We founded the 3Ai Institute at the Australian National University
in September 2017.
It has one deceptively simple mission:
to establish a new branch of engineering
to take AI safely, sustainably and responsibly to scale.
But how do you build a new branch of engineering in the 21st century?
Well, we're teaching it into existence
through an experimental education program.
We're researching it into existence
with locations as diverse as Shakespeare's birthplace,
the Great Barrier Reef,
not to mention one of Australia's largest autonomous mines.
And we're theorizing it into existence,
paying attention to the complexities of cybernetic systems.
We're working to build something new and something useful.
Something to create the next generation of critical thinkers and critical doers.
And we're doing all of that
through a richer understanding of AI's many pasts and many stories.
And by working collaboratively and collectively
through teaching and research and engagement,
and by focusing as much on the framing of the questions
as the solving of the problems.
We're not making a single AI,
we're making the possibilities for many.
And we're actively working to decolonize our imaginations
and to build a curriculum and a pedagogy
that leaves room for a range of different conversations and possibilities.
We are making and remaking.
And I know we're always a work in progress.
But here's a little glimpse
into how we're approaching that problem of scaling a future.
We start by making sure we're grounded in our own history.
In December of 2018,
I took myself up to the town of Brewarrina
on the New South Wales-Queensland border.
This place was a meeting place for Aboriginal people,
for different groups,
to gather, have ceremonies, meet, to be together.
There, on the Barwon River, there's a set of fish weirs
that are one of the oldest and largest systems
of Aboriginal fish traps in Australia.
This system is comprised of 1.8 kilometers of stone walls
shaped like a series of fishnets
with the "Us" pointing down the river,
allowing fish to be trapped at different heights of the water.
They're also fish holding pens with different-height walls for storage,

designed to change the way the water moves
and to be able to store big fish and little fish
and to keep those fish in cool, clear running water.
This fish-trap system was a way to ensure that you could feed people
as they gathered there in a place that was both a meeting of rivers
and a meeting of cultures.
It isn't about the rocks or even the traps per se.
It is about the system that those traps created.
One that involves technical knowledge,
cultural knowledge
and ecological knowledge.
This system is old.
Some archaeologists think it's as old as 40,000 years.
The last time we have its recorded uses is in the nineteen teens.
It's had remarkable longevity and incredible scale.
And it's an inspiration to me.
And a photo of the weir is on our walls here at the Institute,
to remind us of the promise and the challenge
of building something meaningful.
And to remind us that we're building systems
in a place where people have built systems
and sustained those same systems for generations.
It isn't just our history,
it's our legacy as we seek to establish a new branch of engineering.
To build on that legacy and our sense of purpose,
I think we need a clear framework for asking questions about the future.
Questions for which there aren't ready or easy answers.
Here, the point is the asking of the questions.
We believe you need to go beyond the traditional approach
of problem-solving,
to the more complicated one of question asking
and question framing.
Because in so doing, you open up all kinds of new possibilities
and new challenges.
For me, right now,
there are six big questions that frame our approach
for taking AI safely, sustainably and responsibly to scale.
Questions about autonomy,
agency, assurance,
indicators, interfaces and intentionality.
The first question we ask is a simple one.
Is the system autonomous?
Think back to that lift on Bligh Street.
The reality is, one day, that lift may be autonomous.
Which is to say it will be able to act without being told to act.
But it isn't fully autonomous, right?
It can't leave that Bligh Street building
and wonder down to Circular Quay for a beer.
It goes up and down, that's all.
But it does it by itself.
It's autonomous in that sense.

The second question we ask:
does this system have agency?
Does this system have controls and limits that live somewhere
that prevent it from doing certain kinds of things under certain conditions.
The reality with lifts, that's absolutely the case.
Think of any lift you've been in.
There's a red keyslot in the elevator carriage
that an emergency services person can stick a key into
and override the whole system.
But what happens when that system is AI-driven?
Where does the key live?
Is it a physical key, is it a digital key?
Who gets to use it?
Is that the emergency services people?
And how would you know if that was happening?
How would all of that be manifested to you in the lift?
The third question we ask is how do we think about assurance.
How do we think about all of its pieces:
safety, security, trust, risk, liability, manageability,
explicability, ethics, public policy, law, regulation?
And how would we tell you that the system was safe and functioning?
The fourth question we ask
is what would be our interfaces with these AI-driven systems.
Will we talk to them?
Will they talk to us, will they talk to each other?
And what will it mean to have a series of technologies we've known,
for some of us, all our lives,
now suddenly behave in entirely different ways?
Lifts, cars, the electrical grid, traffic lights, things in your home.
The fifth question for these AI-driven systems:
What will the indicators be to show that they're working well?
Two hundred years of the industrial revolution
tells us that the two most important ways to think about a good system
are productivity and efficiency.
In the 21st century,
you might want to expand that just a little bit.
Is the system sustainable,
is it safe, is it responsible?
Who gets to judge those things for us?
Users of the systems would want to understand
how these things are regulated, managed and built.
And then there's the final, perhaps most critical question
that you need to ask of these new AI systems.
What's its intent?
What's the system designed to do
and who said that was a good idea?
Or put another way,
what is the world that this system is building,
how is that world imagined,
and what is its relationship to the world we live in today?
Who gets to be part of that conversation?

Who gets to articulate it?
How does it get framed and imagined?
There are no simple answers to these questions.
Instead, they frame what's possible
and what we need to imagine,
design, build, regulate and even decommission.
They point us in the right directions
and help us on a path to establish a new branch of engineering.
But critical questions aren't enough.
You also need a way of holding all those questions together.
For us at the Institute,
we're also really interested in how to think about AI as a system,
and where and how to draw the boundaries of that system.
And those feel like especially important things right now.
Here, we're influenced by the work that was started way back in the 1940s.
In 1944, along with anthropologists Gregory Bateson and Margaret Mead,
mathematician Norbert Wiener convened a series of conversations
that would become known as the Macy Conferences on Cybernetics.
Ultimately, between 1946 and 1953,
ten conferences were held under the banner of cybernetics.
As defined by Norbert Wiener,
cybernetics sought to "develop a language and techniques
that will enable us to indeed attack the problem of control and communication
in advanced computing technologies."
Cybernetics argued persuasively
that one had to think about the relationship
between humans, computers
and the broader ecological world.
You had to think about them as a holistic system.
Participants in the Macy Conferences were concerned with how the mind worked,
with ideas about intelligence and learning,
and about the role of technology in our future.
Sadly, the conversations that started with the Macy Conference
are often forgotten when the talk is about AI.
But for me, there's something really important to reclaim here
about the idea of a system that has to accommodate culture,
technology and the environment.
At the Institute, that sort of systems thinking is core to our work.
Over the last three years,
a whole collection of amazing people have joined me here
on this crazy journey to do this work.
Our staff includes anthropologists,
systems and environmental engineers, and computer scientists
as well as a nuclear physicist,
an award-winning photo journalist,
and at least one policy and standards expert.
It's a heady mix.
And the range of experience and expertise is powerful,
as are the conflicts and the challenges.
Being diverse requires a constant willingness
to find ways to hold people in conversation.

And to dwell just a little bit with the conflict.
We also worked out early
that the way to build a new way of doing things
would require a commitment to bringing others along on that same journey with us.
So we opened our doors to an education program very quickly,
and we launched our first master's program in 2018.
Since then, we've had two cohorts of master's students
and one cohort of PhD students.
Our students come from all over the world
and all over life.
Australia, New Zealand, Nigeria, Nepal,
Mexico, India, the United States.
And they range in age from 23 to 60.
They variously had backgrounds in maths and music,
policy and performance,
systems and standards,
architecture and arts.
Before they joined us at the Institute,
they ran companies, they worked for government,
served in the army, taught high school,
and managed arts organizations.
They were adventurers
and committed to each other,
and to building something new.
And really, what more could you ask for?
Because although I've spent 20 years in Silicon Valley
and I know the stories about the lone inventor
and the hero's journey,
I also know the reality.
That it's never just a hero's journey.
It's always a collection of people who have a shared sense of purpose
who can change the world.
So where do you start?
Well, I think you start where you stand.
And for me, that means I want to acknowledge
the traditional owners of the land upon which I'm standing.
The Ngunnawal and Ngambri people,
this is their land,
never ceded, always sacred.
And I pay my respects to the elders, past and present, of this place.
I also acknowledge that we're gathering today
in many other places,
and I pay my respects to the traditional owners and elders
of all those places too.
It means a lot to me to get to say those words
and to dwell on what they mean and signal.
And to remember that we live in a country
that has been continuously occupied for at least 60,000 years.
Aboriginal people built worlds here,
they built social systems, they built technologies.
They built ways to manage this place

and to manage it remarkably over a protracted period of time.
And every moment any one of us stands on a stage as Australians,
here or abroad,
we carry with us a privilege and a responsibility
because of that history.
And it's not just a history.
It's also an incredibly rich set of resources,
worldviews and knowledge.
And it should run through all of our bones
and it should be the story we always tell.
Ultimately, it's about thinking differently,
asking different kinds of questions,
looking holistically at the world and the systems,
and finding other people who want to be on that journey with you.
Because for me,
the only way to actually think about the future and scale
is to always be doing it collectively.
And because for me,
the notion of humans in it together
is one of the ways we get to think about things
that are responsible, safe
and ultimately, sustainable.
Thank you.
AI isn't as smart as you think, but it could be
Jeff Dean

Hi, I'm Jeff.
I lead AI Research and Health at Google.
I joined Google more than 20 years ago,
when we were all wedged into a tiny office space,
above what's now a T-Mobile store in downtown Palo Alto.
I've seen a lot of computing transformations in that time,
and in the last decade, we've seen AI be able to do tremendous things.
But we're still doing it all wrong in many ways.
That's what I want to talk to you about today.
But first, let's talk about what AI can do.
So in the last decade, we've seen tremendous progress
in how AI can help computers see, understand language,
understand speech better than ever before.
Things that we couldn't do before, now we can do.
If you think about computer vision alone,
just in the last 10 years,
computers have effectively developed the ability to see;
10 years ago, they couldn't see, now they can see.
You can imagine this has had a transformative effect
on what we can do with computers.
So let's look at a couple of the great applications
enabled by these capabilities.
We can better predict flooding, keep everyone safe,
using machine learning.
We can translate over 100 languages so we all can communicate better,

and better predict and diagnose disease,
where everyone gets the treatment that they need.
So let's look at two key components
that underlie the progress in AI systems today.
The first is neural networks,
a breakthrough approach to solving some of these difficult problems
that has really shone in the last 15 years.
But they're not a new idea.
And the second is computational power.
It actually takes a lot of computational power
to make neural networks able to really sing,
and in the last 15 years, we’ve been able to have that,
and that's partly what's enabled all this progress.
But at the same time, I think we're doing several things wrong,
and that's what I want to talk to you about
at the end of the talk.
First, a bit of a history lesson.
So for decades,
almost since the very beginning of computing,
people have wanted to be able to build computers
that could see, understand language, understand speech.
The earliest approaches to this, generally,
people were trying to hand-code all the algorithms
that you need to accomplish those difficult tasks,
and it just turned out to not work very well.
But in the last 15 years, a single approach
unexpectedly advanced all these different problem spaces all at once:
neural networks.
So neural networks are not a new idea.
They're kind of loosely based
on some of the properties that are in real neural systems.
And many of the ideas behind neural networks
have been around since the 1960s and 70s.
A neural network is what it sounds like,
a series of interconnected artificial neurons
that loosely emulate the properties of your real neurons.
An individual neuron in one of these systems
has a set of inputs,
each with an associated weight,
and the output of a neuron
is a function of those inputs multiplied by those weights.
So pretty simple,
and lots and lots of these work together to learn complicated things.
So how do we actually learn in a neural network?
It turns out the learning process
consists of repeatedly making tiny little adjustments
to the weight values,
strengthening the influence of some things,
weakening the influence of others.
By driving the overall system towards desired behaviors,
these systems can be trained to do really complicated things,

like translate from one language to another,
detect what kind of objects are in a photo,
all kinds of complicated things.
I first got interested in neural networks
when I took a class on them as an undergraduate in 1990.
At that time,
neural networks showed impressive results on tiny problems,
but they really couldn't scale to do real-world important tasks.
But I was super excited.

I felt maybe we just needed more compute power.
And the University of Minnesota had a 32-processor machine.
I thought, "With more compute power,
boy, we could really make neural networks really sing."
So I decided to do a senior thesis on parallel training of neural networks,
the idea of using processors in a computer or in a computer system
to all work toward the same task,
that of training neural networks.
32 processors, wow,
we’ve got to be able to do great things with this.
But I was wrong.
Turns out we needed about a million times as much computational power
as we had in 1990
before we could actually get neural networks to do impressive things.
But starting around 2005,
thanks to the computing progress of Moore's law,
we actually started to have that much computing power,
and researchers in a few universities around the world started to see success
in using neural networks for a wide variety of different kinds of tasks.
I and a few others at Google heard about some of these successes,
and we decided to start a project to train very large neural networks.
One system that we trained,
we trained with 10 million randomly selected frames
from YouTube videos.
The system developed the capability
to recognize all kinds of different objects.
And it being YouTube, of course,
it developed the ability to recognize cats.
YouTube is full of cats.

But what made that so remarkable
is that the system was never told what a cat was.
So using just patterns in data,
the system honed in on the concept of a cat all on its own.
All of this occurred at the beginning of a decade-long string of successes,
of using neural networks for a huge variety of tasks,
at Google and elsewhere.
Many of the things you use every day,
things like better speech recognition for your phone,
improved understanding of queries and documents
for better search quality,

better understanding of geographic information to improve maps,
and so on.
Around that time,
we also got excited about how we could build hardware that was better tailored
to the kinds of computations neural networks wanted to do.
Neural network computations have two special properties.
The first is they're very tolerant of reduced precision.
Couple of significant digits, you don't need six or seven.
And the second is that all the algorithms are generally composed
of different sequences of matrix and vector operations.
So if you can build a computer
that is really good at low-precision matrix and vector operations
but can't do much else,
that's going to be great for neural-network computation,
even though you can't use it for a lot of other things.
And if you build such things, people will find amazing uses for them.
This is the first one we built, TPU v1.
"TPU" stands for Tensor Processing Unit.
These have been used for many years behind every Google search,
for translation,
in the DeepMind AlphaGo matches,
so Lee Sedol and Ke Jie maybe didn't realize,
but they were competing against racks of TPU cards.
And we've built a bunch of subsequent versions of TPUs
that are even better and more exciting.
But despite all these successes,
I think we're still doing many things wrong,
and I'll tell you about three key things we're doing wrong,
and how we'll fix them.
The first is that most neural networks today
are trained to do one thing, and one thing only.
You train it for a particular task that you might care deeply about,
but it's a pretty heavyweight activity.
You need to curate a data set,
you need to decide what network architecture you'll use
for this problem,
you need to initialize the weights with random values,
apply lots of computation to make adjustments to the weights.
And at the end, if you’re lucky, you end up with a model
that is really good at that task you care about.
But if you do this over and over,
you end up with thousands of separate models,
each perhaps very capable,
but separate for all the different tasks you care about.
But think about how people learn.
In the last year, many of us have picked up a bunch of new skills.
I've been honing my gardening skills,
experimenting with vertical hydroponic gardening.
To do that, I didn't need to relearn everything I already knew about plants.
I was able to know how to put a plant in a hole,
how to pour water, that plants need sun,

and leverage that in learning this new skill.
Computers can work the same way, but they don’t today.
If you train a neural network from scratch,
it's effectively like forgetting your entire education
every time you try to do something new.
That’s crazy, right?
So instead, I think we can and should be training
multitask models that can do thousands or millions of different tasks.
Each part of that model would specialize in different kinds of things.
And then, if we have a model that can do a thousand things,
and the thousand and first thing comes along,
we can leverage the expertise we already have
in the related kinds of things
so that we can more quickly be able to do this new task,
just like you, if you're confronted with some new problem,
you quickly identify the 17 things you already know
that are going to be helpful in solving that problem.
Second problem is that most of our models today
deal with only a single modality of data --
with images, or text or speech,
but not all of these all at once.
But think about how you go about the world.
You're continuously using all your senses
to learn from, react to,
figure out what actions you want to take in the world.
Makes a lot more sense to do that,
and we can build models in the same way.
We can build models that take in these different modalities of input data,
text, images, speech,
but then fuse them together,
so that regardless of whether the model sees the word "leopard,"
sees a video of a leopard or hears someone say the word "leopard,"
the same response is triggered inside the model:
the concept of a leopard
can deal with different kinds of input data,
even nonhuman inputs, like genetic sequences,
3D clouds of points, as well as images, text and video.
The third problem is that today's models are dense.
There's a single model,
the model is fully activated for every task,
for every example that we want to accomplish,
whether that's a really simple or a really complicated thing.
This, too, is unlike how our own brains work.
Different parts of our brains are good at different things,
and we're continuously calling upon the pieces of them
that are relevant for the task at hand.
For example, nervously watching a garbage truck
back up towards your car,
the part of your brain that thinks about Shakespearean sonnets
is probably inactive.

AI models can work the same way.
Instead of a dense model,
we can have one that is sparsely activated.
So for particular different tasks, we call upon different parts of the model.
During training, the model can also learn which parts are good at which things,
to continuously identify what parts it wants to call upon
in order to accomplish a new task.
The advantage of this is we can have a very high-capacity model,
but it's very efficient,
because we're only calling upon the parts that we need
for any given task.
So fixing these three things, I think,
will lead to a more powerful AI system:
instead of thousands of separate models,
train a handful of general-purpose models
that can do thousands or millions of things.
Instead of dealing with single modalities,
deal with all modalities,
and be able to fuse them together.
And instead of dense models, use sparse, high-capacity models,
where we call upon the relevant bits as we need them.
We've been building a system that enables these kinds of approaches,
and we've been calling the system "Pathways."
So the idea is this model will be able to do
thousands or millions of different tasks,
and then, we can incrementally add new tasks,
and it can deal with all modalities at once,
and then incrementally learn new tasks as needed
and call upon the relevant bits of the model
for different examples or tasks.
And we're pretty excited about this,
we think this is going to be a step forward
in how we build AI systems.
But I also wanted to touch on responsible AI.
We clearly need to make sure that this vision of powerful AI systems
benefits everyone.
These kinds of models raise important new questions
about how do we build them with fairness,
interpretability, privacy and security,
for all users in mind.
For example, if we're going to train these models
on thousands or millions of tasks,
we'll need to be able to train them on large amounts of data.
And we need to make sure that data is thoughtfully collected
and is representative of different communities and situations
all around the world.
And data concerns are only one aspect of responsible AI.
We have a lot of work to do here.
So in 2018, Google published this set of AI principles
by which we think about developing these kinds of technology.
And these have helped guide us in how we do research in this space,

how we use AI in our products.
And I think it's a really helpful and important framing
for how to think about these deep and complex questions
about how we should be using AI in society.
We continue to update these as we learn more.
Many of these kinds of principles are active areas of research --
super important area.
Moving from single-purpose systems that kind of recognize patterns in data
to these kinds of general-purpose intelligent systems
that have a deeper understanding of the world
will really enable us to tackle
some of the greatest problems humanity faces.
For example,
we’ll be able to diagnose more disease;
we'll be able to engineer better medicines
by infusing these models with knowledge of chemistry and physics;
we'll be able to advance educational systems
by providing more individualized tutoring
to help people learn in new and better ways;
we’ll be able to tackle really complicated issues,
like climate change,
and perhaps engineering of clean energy solutions.
So really, all of these kinds of systems
are going to be requiring the multidisciplinary expertise
of people all over the world.
So connecting AI with whatever field you are in,
in order to make progress.
So I've seen a lot of advances in computing,
and how computing, over the past decades,
has really helped millions of people better understand the world around them.
And AI today has the potential to help billions of people.
We truly live in exciting times.
Thank you.

How AI can bring on a second Industrial Revolution

Kevin Kelly

I'm going to talk a little bit about where technology's going.
And often technology comes to us,
we're surprised by what it brings.
But there's actually a large aspect of technology
that's much more predictable,
and that's because technological systems of all sorts have leanings,
they have urgencies,
they have tendencies.
And those tendencies are derived from the very nature of the physics,
chemistry of wires and switches and electrons,
and they will make reoccurring patterns again and again.
And so those patterns produce these tendencies, these leanings.
You can almost think of it as sort of like gravity.
Imagine raindrops falling into a valley.
The actual path of a raindrop as it goes down the valley
is unpredictable.
We cannot see where it's going,
but the general direction is very inevitable:
it's downward.
And so these baked-in tendencies and urgencies
in technological systems
give us a sense of where things are going at the large form.
So in a large sense,
I would say that telephones were inevitable,
but the iPhone was not.
The Internet was inevitable,
but Twitter was not.
So we have many ongoing tendencies right now,
and I think one of the chief among them
is this tendency to make things smarter and smarter.
I call it cognifying -- cognification --
also known as artificial intelligence, or AI.
And I think that's going to be one of the most influential developments
and trends and directions and drives in our society in the next 20 years.
So, of course, it's already here.
We already have AI,
and often it works in the background,
in the back offices of hospitals,
where it's used to diagnose X-rays better than a human doctor.
It's in legal offices,
where it's used to go through legal evidence
better than a human paralawyer.
It's used to fly the plane that you came here with.
Human pilots only flew it seven to eight minutes,
the rest of the time the AI was driving.
And of course, in Netflix and Amazon,
it's in the background, making those recommendations.
That's what we have today.

And we have an example, of course, in a more front-facing aspect of it,
with the win of the AlphaGo, who beat the world's greatest Go champion.
But it's more than that.
If you play a video game, you're playing against an AI.
But recently, Google taught their AI
to actually learn how to play video games.
Again, teaching video games was already done,
but learning how to play a video game is another step.
That's artificial smartness.
What we're doing is taking this artificial smartness
and we're making it smarter and smarter.
There are three aspects to this general trend
that I think are underappreciated;
I think we would understand AI a lot better
if we understood these three things.
I think these things also would help us embrace AI,
because it's only by embracing it that we actually can steer it.
We can actually steer the specifics by embracing the larger trend.
So let me talk about those three different aspects.
The first one is: our own intelligence has a very poor understanding
of what intelligence is.
We tend to think of intelligence as a single dimension,
that it's kind of like a note that gets louder and louder.
It starts like with IQ measurement.
It starts with maybe a simple low IQ in a rat or mouse,
and maybe there's more in a chimpanzee,
and then maybe there's more in a stupid person,
and then maybe an average person like myself,
and then maybe a genius.
And this single IQ intelligence is getting greater and greater.
That's completely wrong.
That's not what intelligence is -- not what human intelligence is, anyway.
It's much more like a symphony of different notes,
and each of these notes is played on a different instrument of cognition.
There are many types of intelligences in our own minds.
We have deductive reasoning,
we have emotional intelligence,
we have spatial intelligence;
we have maybe 100 different types that are all grouped together,
and they vary in different strengths with different people.
And of course, if we go to animals, they also have another basket --
another symphony of different kinds of intelligences,
and sometimes those same instruments are the same that we have.
They can think in the same way, but they may have a different arrangement,
and maybe they're higher in some cases than humans,
like long-term memory in a squirrel is actually phenomenal,
so it can remember where it buried its nuts.
But in other cases they may be lower.
When we go to make machines,
we're going to engineer them in the same way,
where we'll make some of those types of smartness much greater than ours,

and many of them won't be anywhere near ours,
because they're not needed.
So we're going to take these things,
these artificial clusters,
and we'll be adding more varieties of artificial cognition to our AIs.
We're going to make them very, very specific.
So your calculator is smarter than you are in arithmetic already;
your GPS is smarter than you are in spatial navigation;
Google, Bing, are smarter than you are in long-term memory.
And we're going to take, again, these kinds of different types of thinking
and we'll put them into, like, a car.
The reason why we want to put them in a car so the car drives,
is because it's not driving like a human.
It's not thinking like us.
That's the whole feature of it.
It's not being distracted,
it's not worrying about whether it left the stove on,
or whether it should have majored in finance.
It's just driving.

Just driving, OK?
And we actually might even come to advertise these
as "consciousness-free."
They're without consciousness,
they're not concerned about those things,
they're not distracted.
So in general, what we're trying to do
is make as many different types of thinking as we can.
We're going to populate the space
of all the different possible types, or species, of thinking.
And there actually may be some problems
that are so difficult in business and science
that our own type of human thinking may not be able to solve them alone.
We may need a two-step program,
which is to invent new kinds of thinking
that we can work alongside of to solve these really large problems,
say, like dark energy or quantum gravity.
What we're doing is making alien intelligences.
You might even think of this as, sort of, artificial aliens
in some senses.
And they're going to help us think different,
because thinking different is the engine of creation
and wealth and new economy.
The second aspect of this is that we are going to use AI
to basically make a second Industrial Revolution.
The first Industrial Revolution was based on the fact
that we invented something I would call artificial power.
Previous to that,
during the Agricultural Revolution,
everything that was made had to be made with human muscle
or animal power.

That was the only way to get anything done.
The great innovation during the Industrial Revolution was,
we harnessed steam power, fossil fuels,
to make this artificial power that we could use
to do anything we wanted to do.
So today when you drive down the highway,
you are, with a flick of the switch, commanding 250 horses --
250 horsepower --
which we can use to build skyscrapers, to build cities, to build roads,
to make factories that would churn out lines of chairs or refrigerators
way beyond our own power.
And that artificial power can also be distributed on wires on a grid
to every home, factory, farmstead,
and anybody could buy that artificial power,
just by plugging something in.
So this was a source of innovation as well,
because a farmer could take a manual hand pump,
and they could add this artificial power, this electricity,
and he'd have an electric pump.
And you multiply that by thousands or tens of thousands of times,
and that formula was what brought us the Industrial Revolution.
All the things that we see, all this progress that we now enjoy,
has come from the fact that we've done that.
We're going to do the same thing now with AI.
We're going to distribute that on a grid,
and now you can take that electric pump.
You can add some artificial intelligence,
and now you have a smart pump.
And that, multiplied by a million times,
is going to be this second Industrial Revolution.
So now the car is going down the highway,
it's 250 horsepower, but in addition, it's 250 minds.
That's the auto-driven car.
It's like a new commodity;
it's a new utility.
The AI is going to flow across the grid -- the cloud --
in the same way electricity did.
So everything that we had electrified,
we're now going to cognify.
And I would suggest, then,
that the formula for the next 10,000 start-ups
is very, very simple,
which is to take x and add AI.
That is the formula, that's what we're going to be doing.
And that is the way in which we're going to make
this second Industrial Revolution.
And by the way -- right now, this minute,
you can log on to Google
and you can purchase AI for six cents, 100 hits.
That's available right now.
So the third aspect of this

is that when we take this AI and embody it,
we get robots.
And robots are going to be bots,
they're going to be doing many of the tasks that we have already done.
A job is just a bunch of tasks,
so they're going to redefine our jobs
because they're going to do some of those tasks.
But they're also going to create whole new categories,
a whole new slew of tasks
that we didn't know we wanted to do before.
They're going to actually engender new kinds of jobs,
new kinds of tasks that we want done,
just as automation made up a whole bunch of new things
that we didn't know we needed before,
and now we can't live without them.
So they're going to produce even more jobs than they take away,
but it's important that a lot of the tasks that we're going to give them
are tasks that can be defined in terms of efficiency or productivity.
If you can specify a task,
either manual or conceptual,
that can be specified in terms of efficiency or productivity,
that goes to the bots.
Productivity is for robots.
What we're really good at is basically wasting time.

We're really good at things that are inefficient.
Science is inherently inefficient.
It runs on that fact that you have one failure after another.
It runs on the fact that you make tests and experiments that don't work,
otherwise you're not learning.
It runs on the fact
that there is not a lot of efficiency in it.
Innovation by definition is inefficient,
because you make prototypes,
because you try stuff that fails, that doesn't work.
Exploration is inherently inefficiency.
Art is not efficient.
Human relationships are not efficient.
These are all the kinds of things we're going to gravitate to,
because they're not efficient.
Efficiency is for robots.
We're also going to learn that we're going to work with these AIs
because they think differently than us.
When Deep Blue beat the world's best chess champion,
people thought it was the end of chess.
But actually, it turns out that today, the best chess champion in the world
is not an AI.
And it's not a human.
It's the team of a human and an AI.
The best medical diagnostician is not a doctor, it's not an AI,
it's the team.

We're going to be working with these AIs,
and I think you'll be paid in the future
by how well you work with these bots.
So that's the third thing, is that they're different,
they're utility
and they are going to be something we work with rather than against.
We're working with these rather than against them.
So, the future:
Where does that take us?
I think that 25 years from now, they'll look back
and look at our understanding of AI and say,
"You didn't have AI. In fact, you didn't even have the Internet yet,
compared to what we're going to have 25 years from now."
There are no AI experts right now.
There's a lot of money going to it,
there are billions of dollars being spent on it;
it's a huge business,
but there are no experts, compared to what we'll know 20 years from now.
So we are just at the beginning of the beginning,
we're in the first hour of all this.
We're in the first hour of the Internet.
We're in the first hour of what's coming.
The most popular AI product in 20 years from now,
that everybody uses,
has not been invented yet.
That means that you're not late.
Thank you.

Why people and AI make good business partners

Shervin Khodabandeh

I've been working in AI for most of my career,
helping companies build artificial intelligence capabilities
to improve their business,
which is why I think what I'm about to tell you
is quite shocking.
Every year, thousands of companies across the world
spend collectively tens of billions of dollars to build AI capabilities.
But according to research my colleagues and I have done,
only about 10 percent of these companies get any meaningful financial impact
from their investments.
These 10 percent winners with AI have a secret.
And their secret is not about fancy algorithms or sophisticated technology.
It's something far more basic.
It's how they get their people and AI to work together.
Together, not against each other,
not instead of each other.
Together in a mutually beneficial relationship.
Unfortunately, when most people think about AI,
they think about the most extreme cases.
That AI is here only to replace us
or overtake our intelligence and make us unnecessary.
But what I'm saying
is that we don't seem to quite appreciate the huge opportunity that exists
in the middle ground,
where humans and AI come together
to achieve outcomes that neither one could do alone on their own.
Consider the game of chess.
You probably knew that AI today can beat any human grandmaster.
But did you know that the combination of a human chess player and AI
can beat not only any human but also any machine.
The combination is much more powerful than the sum of its parts.
In a perfect combination, AI will do what it does best,
which is dealing with massive amounts of data and solving complex problems.
And humans do what we do best
using our creativity, our judgment, our empathy, our ethics
and our ability to compromise.
For several years,
my colleagues and I have studied
and worked with hundreds of winning companies
who are successfully building these human-AI relationships.
And what we've seen is quite interesting.
First of all, these companies get five times more financial value
than companies who use AI only to replace people.
Most importantly, they have a happier workforce.
Their employees are more proud, more fulfilled,
they collaborate better with each other, and they're more effective.
Five times more value and a happier workforce.
So the question is, how do these companies do it?

How do they achieve these symbiotic human-AI relationships?
I have some answers.
First of all, they don't think of AI in the most extreme case
only to replace humans.
Instead, they look deep inside their organizations
and at the various roles their people play.
And they ask:
How can AI make our people more fulfilled, more effective,
more amplified?
Let me give you an example.
Humana is a health care company here in the US.
It has pharmacy call centers where pharmacists work with patients
over the phone.
It's a job that requires a fair amount of empathy and humanity.
Humana has developed an AI system
that listens to the pharmacists' conversation
and picks up emotional and tone signals
and then gives real-time suggestions to the pharmacists
on how to improve the quality of that conversation.
For example, it might say “Slow down” or “Pause”
or "Hey, consider how the other person is feeling right now."
All to improve the quality of that conversation.
I'm pretty sure my wife would buy me one of these if she could,
just to help me in some of my conversations with her.

Turns out the pharmacists like it quite a lot, too.
They're more effective in their jobs,
but they also learn something about themselves,
their own behaviors and biases.
The result has been more effective pharmacists
and much higher customer satisfaction scores.
Now, this is just one example of many possibilities where human AI collaborate.
In this example, AI was a recommender.
It didn't replace the human or make any decisions of its own.
It simply made suggestions,
and it was up to the person to decide and act.
And at the heart of it is a feedback loop,
which, by the way, is very critical for any human-AI relationship.
By that I mean that in this example,
first AI had to learn from humans the qualities that would make up a good
or not so good conversation.
And then over time, as AI built more intelligence,
it would be able to make suggestions,
but it would be up to the person to decide and act.
And if they didn't agree with the recommendation
because it might have not made sense to them,
they didn't have to.
In which case AI might learn something and adapt for the future.
It's basically open, frequent, two-way communication,
like any couples therapist will tell you,
is very important for any good relationship.

Now the key word here is relationship.
Think about your own personal relationships with other people.
You don't have the same kind of relationship with your accountant
or your boss or your spouse, do you?
Well, I certainly hope not.
And just like that,
the right relationship between human and AI in a company
is not a one-size-fits-all.
So in the case of Humana, AI was a recommender
and a human was decision-maker and actor.
In some other examples, AI might be an evaluator
where a human comes up with ideas or scenarios,
and AI evaluates the complex implications and tradeoffs of those ideas
and makes it easy for humans to decide the best course of action.
In some other examples, AI might take a more creative role.
It could be an illuminator where it can take a complex problem
and come up with potential solutions to that problem
and illuminate some options
that might have been impossible for humans to see.
Let me give you another example.
During the COVID pandemic,
if you walked into a retail or grocery store,
you saw that many retailers were struggling.
Their shelves were empty,
their suppliers were not able to fulfill the orders,
and with all the uncertainties of the pandemic,
they simply had no idea how many people would be walking into what stores,
demanding what products.
Now, to put this in perspective,
this is a problem that's already quite hard when things are normal.
Retailers have to predict demand
for tens of thousands of products across thousands of locations
and thousands of suppliers every day
to manage and optimize their inventory.
Add to that the uncertainties of COVID and the global supply chain disruptions,
and this became 100 times more difficult.
And many retailers were simply paralyzed.
But there were a few who had built strong foundations with AI
and the human-AI feedback loop that we talked about.
And these guys were able to navigate all this uncertainty
much better than others.
They used AI to analyze tens of billions of data points
on consumer behavior and global supply chain disruptions
and local government closures and mandates
and traffic on highways
and ocean freight lanes and many, many other factors
and get a pretty good handle on what consumers in each unique area
wanted the most,
what would have been feasible,
and for items that were not available,
what substitutions could be made.

But AI alone without the human touch wouldn't work either.
There were ethical and economic tradeoffs that had to be considered.
For example, deciding to bring in a product
that didn't have a good margin for the retailer
but would really help support the local community
at their time of need.
After all, AI couldn't quite understand
the uniquely human behavior of panic-buying toilet paper
or tens of gallons of liquor,
only to be used as hand sanitizer.
It was the combination that was the key.
And the winning companies know this.
They also know that inside their companies,
there's literally hundreds of these opportunities for human-AI combination,
and they actively identify and pursue them.
They think of AI as much more broadly a means to replace people.
They look inside their organizations
and re-imagine how the biggest challenges and opportunities of their company
can be addressed
by the combination of human and AI.
And they put in place the right combination for each unique situation.
Whether it's the recommender or the evaluator
or the illuminator or optimizer or many, many other ones.
They build and evolve the feedback loops that we talked about.
And finally and most importantly, they don't just throw technology at it.
In fact, this has been the biggest pitfall of companies
who don't get their return from their AI investments.
If they overinvest in technology
expecting a piece of tech to solve all their problems.
But there is no silver bullet.
Technology and automation can only go so far,
and for every one automation opportunity inside a company,
there's literally ten for collaboration.
But collaboration's hard.
It requires a new mindset
and doing things differently than how we've always done it.
And the winning companies know this, too,
which is why they don't just invest in technology,
but so much more on human factors,
on their people, on training and reskilling
and reimagining how their people and AI work together in new ways.
Inside these companies, it's not just machines replacing humans.
It's machines and humans working together,
learning from each other.
And when that happens,
the organization's overall rate of learning increases,
which in turn makes the company much more agile,
much more resilient,
ready to adapt and take on any challenge.
It is the human touch that will bring the best out of AI.
Thank you.

How AI is making it easier to diagnose disease

Pratik Shah

Computer algorithms today are performing incredible tasks
with high accuracies, at a massive scale, using human-like intelligence.
And this intelligence of computers is often referred to as AI
or artificial intelligence.
AI is poised to make an incredible impact on our lives in the future.
Today, however, we still face massive challenges
in detecting and diagnosing several life-threatening illnesses,
such as infectious diseases and cancer.
Thousands of patients every year
lose their lives due to liver and oral cancer.
Our best way to help these patients
is to perform early detection and diagnoses of these diseases.
So how do we detect these diseases today, and can artificial intelligence help?
In patients who, unfortunately, are suspected of these diseases,
an expert physician first orders
very expensive medical imaging technologies
such as fluorescent imaging, CTs, MRIs, to be performed.
Once those images are collected,
another expert physician then diagnoses those images and talks to the patient.
As you can see, this is a very resource-intensive process,
requiring both expert physicians, expensive medical imaging technologies,
and is not considered practical for the developing world.
And in fact, in many industrialized nations, as well.
So, can we solve this problem using artificial intelligence?
Today, if I were to use traditional artificial intelligence architectures
to solve this problem,
I would require 10,000 --
I repeat, on an order of 10,000 of these very expensive medical images
first to be generated.
After that, I would then go to an expert physician,
who would then analyze those images for me.
And using those two pieces of information,
I can train a standard deep neural network or a deep learning network
to provide patient's diagnosis.
Similar to the first approach,
traditional artificial intelligence approaches
suffer from the same problem.
Large amounts of data, expert physicians and expert medical imaging technologies.
So, can we invent more scalable, effective
and more valuable artificial intelligence architectures
to solve these very important problems facing us today?
And this is exactly what my group at MIT Media Lab does.
We have invented a variety of unorthodox AI architectures
to solve some of the most important challenges facing us today
in medical imaging and clinical trials.
In the example I shared with you today, we had two goals.
Our first goal was to reduce the number of images
required to train artificial intelligence algorithms.

Our second goal -- we were more ambitious,
we wanted to reduce the use of expensive medical imaging technologies
to screen patients.
So how did we do it?
For our first goal,
instead of starting with tens and thousands
of these very expensive medical images, like traditional AI,
we started with a single medical image.
From this image, my team and I figured out a very clever way
to extract billions of information packets.
These information packets included colors, pixels, geometry
and rendering of the disease on the medical image.
In a sense, we converted one image into billions of training data points,
massively reducing the amount of data needed for training.
For our second goal,
to reduce the use of expensive medical imaging technologies to screen patients,
we started with a standard, white light photograph,
acquired either from a DSLR camera or a mobile phone, for the patient.
Then remember those billions of information packets?
We overlaid those from the medical image onto this image,
creating something that we call a composite image.
Much to our surprise, we only required 50 --
I repeat, only 50 --
of these composite images to train our algorithms to high efficiencies.
To summarize our approach,
instead of using 10,000 very expensive medical images,
we can now train the AI algorithms in an unorthodox way,
using only 50 of these high-resolution, but standard photographs,
acquired from DSLR cameras and mobile phones,
and provide diagnosis.
More importantly,
our algorithms can accept, in the future and even right now,
some very simple, white light photographs from the patient,
instead of expensive medical imaging technologies.
I believe that we are poised to enter an era
where artificial intelligence
is going to make an incredible impact on our future.
And I think that thinking about traditional AI,
which is data-rich but application-poor,
we should also continue thinking
about unorthodox artificial intelligence architectures,
which can accept small amounts of data
and solve some of the most important problems facing us today,
especially in health care.
Thank you very much.

What AI Is (and isn't)

Chris Anderson & Sebastian Thrun

Chris Anderson: Help us understand what machine learning is,
because that seems to be the key driver
of so much of the excitement and also of the concern
around artificial intelligence.
How does machine learning work?
Sebastian Thrun: So, artificial intelligence and machine learning
is about 60 years old
and has not had a great day in its past until recently.
And the reason is that today,
we have reached a scale of computing and datasets
that was necessary to make machines smart.
So here's how it works.
If you program a computer today, say, your phone,
then you hire software engineers
that write a very, very long kitchen recipe,
like, "If the water is too hot, turn down the temperature.
If it's too cold, turn up the temperature."
The recipes are not just 10 lines long.
They are millions of lines long.
A modern cell phone has 12 million lines of code.
A browser has five million lines of code.
And each bug in this recipe can cause your computer to crash.
That's why a software engineer makes so much money.
The new thing now is that computers can find their own rules.
So instead of an expert deciphering, step by step,
a rule for every contingency,
what you do now is you give the computer examples
and have it infer its own rules.
A really good example is AlphaGo, which recently was won by Google.
Normally, in game playing, you would really write down all the rules,
but in AlphaGo's case,
the system looked over a million games
and was able to infer its own rules
and then beat the world's residing Go champion.
That is exciting, because it relieves the software engineer
of the need of being super smart,
and pushes the burden towards the data.
As I said, the inflection point where this has become really possible --
very embarrassing, my thesis was about machine learning.
It was completely insignificant, don't read it,
because it was 20 years ago
and back then, the computers were as big as a cockroach brain.
Now they are powerful enough to really emulate
kind of specialized human thinking.
And then the computers take advantage of the fact
that they can look at much more data than people can.
So I'd say AlphaGo looked at more than a million games.
No human expert can ever study a million games.

Google has looked at over a hundred billion web pages.
No person can ever study a hundred billion web pages.
So as a result, the computer can find rules
that even people can't find.
CA: So instead of looking ahead to, "If he does that, I will do that,"
it's more saying, "Here is what looks like a winning pattern,
here is what looks like a winning pattern."
ST: Yeah. I mean, think about how you raise children.
You don't spend the first 18 years giving kids a rule for every contingency
and set them free and they have this big program.
They stumble, fall, get up, they get slapped or spanked,
and they have a positive experience, a good grade in school,
and they figure it out on their own.
That's happening with computers now,
which makes computer programming so much easier all of a sudden.
Now we don't have to think anymore. We just give them lots of data.
CA: And so, this has been key to the spectacular improvement
in power of self-driving cars.
I think you gave me an example.
Can you explain what's happening here?
ST: This is a drive of a self-driving car
that we happened to have at Udacity
and recently made into a spin-off called Voyage.
We have used this thing called deep learning
to train a car to drive itself,
and this is driving from Mountain View, California,
to San Francisco
on El Camino Real on a rainy day,
with bicyclists and pedestrians and 133 traffic lights.
And the novel thing here is,
many, many moons ago, I started the Google self-driving car team.
And back in the day, I hired the world's best software engineers
to find the world's best rules.
This is just trained.
We drive this road 20 times,
we put all this data into the computer brain,
and after a few hours of processing,
it comes up with behavior that often surpasses human agility.
So it's become really easy to program it.
This is 100 percent autonomous, about 33 miles, an hour and a half.
CA: So, explain it -- on the big part of this program on the left,
you're seeing basically what the computer sees as trucks and cars
and those dots overtaking it and so forth.
ST: On the right side, you see the camera image, which is the main input here,
and it's used to find lanes, other cars, traffic lights.
The vehicle has a radar to do distance estimation.
This is very commonly used in these kind of systems.
On the left side you see a laser diagram,
where you see obstacles like trees and so on depicted by the laser.
But almost all the interesting work is centering on the camera image now.
We're really shifting over from precision sensors like radars and lasers

into very cheap, commoditized sensors.
A camera costs less than eight dollars.
CA: And that green dot on the left thing, what is that?
Is that anything meaningful?
ST: This is a look-ahead point for your adaptive cruise control,
so it helps us understand how to regulate velocity
based on how far the cars in front of you are.
CA: And so, you've also got an example, I think,
of how the actual learning part takes place.
Maybe we can see that. Talk about this.
ST: This is an example where we posed a challenge to Udacity students
to take what we call a self-driving car Nanodegree.
We gave them this dataset
and said "Hey, can you guys figure out how to steer this car?"
And if you look at the images,
it's, even for humans, quite impossible to get the steering right.
And we ran a competition and said, "It's a deep learning competition,
AI competition,"
and we gave the students 48 hours.
So if you are a software house like Google or Facebook,
something like this costs you at least six months of work.
So we figured 48 hours is great.
And within 48 hours, we got about 100 submissions from students,
and the top four got it perfectly right.
It drives better than I could drive on this imagery,
using deep learning.
And again, it's the same methodology.
It's this magical thing.
When you give enough data to a computer now,
and give enough time to comprehend the data,
it finds its own rules.
CA: And so that has led to the development of powerful applications
in all sorts of areas.
You were talking to me the other day about cancer.
Can I show this video?
ST: Yeah, absolutely, please. CA: This is cool.
ST: This is kind of an insight into what's happening
in a completely different domain.
This is augmenting, or competing --
it's in the eye of the beholder --
with people who are being paid 400,000 dollars a year,
dermatologists,
highly trained specialists.
It takes more than a decade of training to be a good dermatologist.
What you see here is the machine learning version of it.
It's called a neural network.
"Neural networks" is the technical term for these machine learning algorithms.
They've been around since the 1980s.
This one was invented in 1988 by a Facebook Fellow called Yann LeCun,
and it propagates data stages
through what you could think of as the human brain.

It's not quite the same thing, but it emulates the same thing.
It goes stage after stage.
In the very first stage, it takes the visual input and extracts edges
and rods and dots.
And the next one becomes more complicated edges
and shapes like little half-moons.
And eventually, it's able to build really complicated concepts.
Andrew Ng has been able to show
that it's able to find cat faces and dog faces
in vast amounts of images.
What my student team at Stanford has shown is that
if you train it on 129,000 images of skin conditions,
including melanoma and carcinomas,
you can do as good a job
as the best human dermatologists.
And to convince ourselves that this is the case,
we captured an independent dataset that we presented to our network
and to 25 board-certified Stanford-level dermatologists,
and compared those.
And in most cases,
they were either on par or above the performance classification accuracy
of human dermatologists.
CA: You were telling me an anecdote.
I think about this image right here.
What happened here?
ST: This was last Thursday. That's a moving piece.
What we've shown before and we published in "Nature" earlier this year
was this idea that we show dermatologists images
and our computer program images,
and count how often they're right.
But all these images are past images.
They've all been biopsied to make sure we had the correct classification.
This one wasn't.
This one was actually done at Stanford by one of our collaborators.
The story goes that our collaborator,
who is a world-famous dermatologist, one of the three best, apparently,
looked at this mole and said, "This is not skin cancer."
And then he had a second moment, where he said,
"Well, let me just check with the app."
So he took out his iPhone and ran our piece of software,
our "pocket dermatologist," so to speak,
and the iPhone said: cancer.
It said melanoma.
And then he was confused.
And he decided, "OK, maybe I trust the iPhone a little bit more than myself,"
and he sent it out to the lab to get it biopsied.
And it came up as an aggressive melanoma.
So I think this might be the first time that we actually found,
in the practice of using deep learning,
an actual person whose melanoma would have gone unclassified,
had it not been for deep learning.

CA: I mean, that's incredible.

It feels like there'd be an instant demand for an app like this right now,
that you might freak out a lot of people.
Are you thinking of doing this, making an app that allows self-checking?
ST: So my in-box is flooded about cancer apps,
with heartbreaking stories of people.
I mean, some people have had 10, 15, 20 melanomas removed,
and are scared that one might be overlooked, like this one,
and also, about, I don't know,
flying cars and speaker inquiries these days, I guess.
My take is, we need more testing.
I want to be very careful.
It's very easy to give a flashy result and impress a TED audience.
It's much harder to put something out that's ethical.
And if people were to use the app
and choose not to consult the assistance of a doctor
because we get it wrong,
I would feel really bad about it.
So we're currently doing clinical tests,
and if these clinical tests commence and our data holds up,
we might be able at some point to take this kind of technology
and take it out of the Stanford clinic
and bring it to the entire world,
places where Stanford doctors never, ever set foot.
CA: And do I hear this right,
that it seemed like what you were saying,
because you are working with this army of Udacity students,
that in a way, you're applying a different form of machine learning
than might take place in a company,
which is you're combining machine learning with a form of crowd wisdom.
Are you saying that sometimes you think that could actually outperform
what a company can do, even a vast company?
ST: I believe there's now instances that blow my mind,
and I'm still trying to understand.
What Chris is referring to is these competitions that we run.
We turn them around in 48 hours,
and we've been able to build a self-driving car
that can drive from Mountain View to San Francisco on surface streets.
It's not quite on par with Google after seven years of Google work,
but it's getting there.
And it took us only two engineers and three months to do this.
And the reason is, we have an army of students
who participate in competitions.
We're not the only ones who use crowdsourcing.
Uber and Didi use crowdsource for driving.
Airbnb uses crowdsourcing for hotels.
There's now many examples where people do bug-finding crowdsourcing
or protein folding, of all things, in crowdsourcing.
But we've been able to build this car in three months,
so I am actually rethinking

how we organize corporations.
We have a staff of 9,000 people who are never hired,
that I never fire.
They show up to work and I don't even know.
Then they submit to me maybe 9,000 answers.
I'm not obliged to use any of those.
I end up -- I pay only the winners,
so I'm actually very cheapskate here, which is maybe not the best thing to do.
But they consider it part of their education, too, which is nice.
But these students have been able to produce amazing deep learning results.
So yeah, the synthesis of great people and great machine learning is amazing.
CA: I mean, Gary Kasparov said on the first day [of TED2017]
that the winners of chess, surprisingly, turned out to be two amateur chess players
with three mediocre-ish, mediocre-to-good, computer programs,
that could outperform one grand master with one great chess player,
like it was all part of the process.
And it almost seems like you're talking about a much richer version
of that same idea.
ST: Yeah, I mean, as you followed the fantastic panels yesterday morning,
two sessions about AI,
robotic overlords and the human response,
many, many great things were said.
But one of the concerns is that we sometimes confuse
what's actually been done with AI with this kind of overlord threat,
where your AI develops consciousness, right?
The last thing I want is for my AI to have consciousness.
I don't want to come into my kitchen
and have the refrigerator fall in love with the dishwasher
and tell me, because I wasn't nice enough,
my food is now warm.
I wouldn't buy these products, and I don't want them.
But the truth is, for me,
AI has always been an augmentation of people.
It's been an augmentation of us,
to make us stronger.
And I think Kasparov was exactly correct.
It's been the combination of human smarts and machine smarts
that make us stronger.
The theme of machines making us stronger is as old as machines are.
The agricultural revolution took place because it made steam engines
and farming equipment that couldn't farm by itself,
that never replaced us; it made us stronger.
And I believe this new wave of AI will make us much, much stronger
as a human race.
CA: We'll come on to that a bit more,
but just to continue with the scary part of this for some people,
like, what feels like it gets scary for people is when you have
a computer that can, one, rewrite its own code,
so, it can create multiple copies of itself,
try a bunch of different code versions,
possibly even at random,

and then check them out and see if a goal is achieved and improved.
So, say the goal is to do better on an intelligence test.
You know, a computer that's moderately good at that,
you could try a million versions of that.
You might find one that was better,
and then, you know, repeat.
And so the concern is that you get some sort of runaway effect
where everything is fine on Thursday evening,
and you come back into the lab on Friday morning,
and because of the speed of computers and so forth,
things have gone crazy, and suddenly --
ST: I would say this is a possibility,
but it's a very remote possibility.
So let me just translate what I heard you say.
In the AlphaGo case, we had exactly this thing:
the computer would play the game against itself
and then learn new rules.
And what machine learning is is a rewriting of the rules.
It's the rewriting of code.
But I think there was absolutely no concern
that AlphaGo would take over the world.
It can't even play chess.
CA: No, no, no, but now, these are all very single-domain things.
But it's possible to imagine.
I mean, we just saw a computer that seemed nearly capable
of passing a university entrance test,
that can kind of -- it can't read and understand in the sense that we can,
but it can certainly absorb all the text
and maybe see increased patterns of meaning.
Isn't there a chance that, as this broadens out,
there could be a different kind of runaway effect?
ST: That's where I draw the line, honestly.
And the chance exists -- I don't want to downplay it --
but I think it's remote, and it's not the thing that's on my mind these days,
because I think the big revolution is something else.
Everything successful in AI to the present date
has been extremely specialized,
and it's been thriving on a single idea,
which is massive amounts of data.
The reason AlphaGo works so well is because of massive numbers of Go plays,
and AlphaGo can't drive a car or fly a plane.
The Google self-driving car or the Udacity self-driving car
thrives on massive amounts of data, and it can't do anything else.
It can't even control a motorcycle.
It's a very specific, domain-specific function,
and the same is true for our cancer app.
There has been almost no progress on this thing called "general AI,"
where you go to an AI and say, "Hey, invent for me special relativity
or string theory."
It's totally in the infancy.
The reason I want to emphasize this,

I see the concerns, and I want to acknowledge them.
But if I were to think about one thing,
I would ask myself the question, "What if we can take anything repetitive
and make ourselves 100 times as efficient?"
It so turns out, 300 years ago, we all worked in agriculture
and did farming and did repetitive things.
Today, 75 percent of us work in offices
and do repetitive things.
We've become spreadsheet monkeys.
And not just low-end labor.
We've become dermatologists doing repetitive things,
lawyers doing repetitive things.
I think we are at the brink of being able to take an AI,
look over our shoulders,
and they make us maybe 10 or 50 times as effective in these repetitive things.
That's what is on my mind.
CA: That sounds super exciting.
The process of getting there seems a little terrifying to some people,
because once a computer can do this repetitive thing
much better than the dermatologist
or than the driver, especially, is the thing that's talked about
so much now,
suddenly millions of jobs go,
and, you know, the country's in revolution
before we ever get to the more glorious aspects of what's possible.
ST: Yeah, and that's an issue, and it's a big issue,
and it was pointed out yesterday morning by several guest speakers.
Now, prior to me showing up onstage,
I confessed I'm a positive, optimistic person,
so let me give you an optimistic pitch,
which is, think of yourself back 300 years ago.
Europe just survived 140 years of continuous war,
none of you could read or write,
there were no jobs that you hold today,
like investment banker or software engineer or TV anchor.
We would all be in the fields and farming.
Now here comes little Sebastian with a little steam engine in his pocket,
saying, "Hey guys, look at this.
It's going to make you 100 times as strong, so you can do something else."
And then back in the day, there was no real stage,
but Chris and I hang out with the cows in the stable,
and he says, "I'm really concerned about it,
because I milk my cow every day, and what if the machine does this for me?"
The reason why I mention this is,
we're always good in acknowledging past progress and the benefit of it,
like our iPhones or our planes or electricity or medical supply.
We all love to live to 80, which was impossible 300 years ago.
But we kind of don't apply the same rules to the future.
So if I look at my own job as a CEO,
I would say 90 percent of my work is repetitive,
I don't enjoy it,

I spend about four hours per day on stupid, repetitive email.
And I'm burning to have something that helps me get rid of this.
Why?
Because I believe all of us are insanely creative;
I think the TED community more than anybody else.
But even blue-collar workers; I think you can go to your hotel maid
and have a drink with him or her,
and an hour later, you find a creative idea.
What this will empower is to turn this creativity into action.
Like, what if you could build Google in a day?
What if you could sit over beer and invent the next Snapchat,
whatever it is,
and tomorrow morning it's up and running?
And that is not science fiction.
What's going to happen is,
we are already in history.
We've unleashed this amazing creativity
by de-slaving us from farming
and later, of course, from factory work
and have invented so many things.
It's going to be even better, in my opinion.
And there's going to be great side effects.
One of the side effects will be
that things like food and medical supply and education and shelter
and transportation
will all become much more affordable to all of us,
not just the rich people.
CA: Hmm.
So when Martin Ford argued, you know, that this time it's different
because the intelligence that we've used in the past
to find new ways to be
will be matched at the same pace
by computers taking over those things,
what I hear you saying is that, not completely,
because of human creativity.
Do you think that that's fundamentally different from the kind of creativity
that computers can do?
ST: So, that's my firm belief as an AI person --
that I haven't seen any real progress on creativity
and out-of-the-box thinking.
What I see right now -- and this is really important for people to realize,
because the word "artificial intelligence" is so threatening,
and then we have Steve Spielberg tossing a movie in,
where all of a sudden the computer is our overlord,
but it's really a technology.
It's a technology that helps us do repetitive things.
And the progress has been entirely on the repetitive end.
It's been in legal document discovery.
It's been contract drafting.
It's been screening X-rays of your chest.
And these things are so specialized,

I don't see the big threat of humanity.
In fact, we as people --
I mean, let's face it: we've become superhuman.
We've made us superhuman.
We can swim across the Atlantic in 11 hours.
We can take a device out of our pocket
and shout all the way to Australia,
and in real time, have that person shouting back to us.
That's physically not possible. We're breaking the rules of physics.
When this is said and done, we're going to remember everything
we've ever said and seen,
you'll remember every person,
which is good for me in my early stages of Alzheimer's.
Sorry, what was I saying? I forgot.
CA: (Laughs)
ST: We will probably have an IQ of 1,000 or more.
There will be no more spelling classes for our kids,
because there's no spelling issue anymore.
There's no math issue anymore.
And I think what really will happen is that we can be super creative.
And we are. We are creative.
That's our secret weapon.
CA: So the jobs that are getting lost,
in a way, even though it's going to be painful,
humans are capable of more than those jobs.
This is the dream.
The dream is that humans can rise to just a new level of empowerment
and discovery.
That's the dream.
ST: And think about this:
if you look at the history of humanity,
that might be whatever -- 60-100,000 years old, give or take --
almost everything that you cherish in terms of invention,
of technology, of things we've built,
has been invented in the last 150 years.
If you toss in the book and the wheel, it's a little bit older.
Or the axe.
But your phone, your sneakers,
these chairs, modern manufacturing, penicillin --
the things we cherish.
Now, that to me means
the next 150 years will find more things.
In fact, the pace of invention has gone up, not gone down, in my opinion.
I believe only one percent of interesting things have been invented yet. Right?
We haven't cured cancer.
We don't have flying cars -- yet. Hopefully, I'll change this.
That used to be an example people laughed about. (Laughs)
It's funny, isn't it? Working secretly on flying cars.
We don't live twice as long yet. OK?
We don't have this magic implant in our brain
that gives us the information we want.

And you might be appalled by it,
but I promise you, once you have it, you'll love it.
I hope you will.
It's a bit scary, I know.
There are so many things we haven't invented yet
that I think we'll invent.
We have no gravity shields.
We can't beam ourselves from one location to another.
That sounds ridiculous,
but about 200 years ago,
experts were of the opinion that flight wouldn't exist,
even 120 years ago,
and if you moved faster than you could run,
you would instantly die.
So who says we are correct today that you can't beam a person
from here to Mars?
CA: Sebastian, thank you so much
for your incredibly inspiring vision and your brilliance.
Thank you, Sebastian Thrun.
That was fantastic.

How AI could compose a personalized soundtrack to your life?

Pierre Barreau

About two and a half years ago, I watched this movie called "Her."
And it features Samantha, a superintelligent form of AI
that cannot take physical form.
And because she can't appear in photographs,
Samantha decides to write a piece of music
that will capture a moment of her life just like a photograph would.
As a musician and an engineer, and someone raised in a family of artists,
I thought that this idea of musical photographs was really powerful.
And I decided to create an AI composer.
Her name is AIVA, and she's an artificial intelligence
that has learned the art of music composition
by reading over 30,000 scores of history's greatest.
So here's what one score looks like to the algorithm
in a matrix-like representation.
And here's what 30,000 scores,
written by the likes of Mozart and Beethoven,
look like in a single frame.
So, using deep neural networks, AIVA looks for patterns in the scores.
And from a couple of bars of existing music,
it actually tries to infer what notes should come next in those tracks.
And once AIVA gets good at those predictions,
it can actually build a set of mathematical rules
for that style of music
in order to create its own original compositions.
And in a way, this is kind of how we, humans, compose music, too.
It's a trial-and-error process,
during which we may not get the right notes all the time.
But we can correct ourselves,
either with our musical ear or our musical knowledge.
But for AIVA, this process is taken from years and years of learning,
decades of learning as an artist, as a musician and a composer,
down to a couple of hours.
But music is also a supersubjective art.
And we needed to teach AIVA
how to compose the right music for the right person,
because people have different preferences.
And to do that, we show to the algorithm over 30 different category labels
for each score in our database.
So those category labels are like mood
or note density or composer style of a piece
or the epoch during which it was written.
And by seeing all this data,
AIVA can actually respond to very precise requirements.
Like the ones, for example, we had for a project recently,
where we were commissioned to create a piece
that would be reminiscent of a science-fiction film soundtrack.
And the piece that was created is called "Among the Stars"
and it was recorded with CMG Orchestra in Hollywood,

under great conductor John Beal,
and this is what they recorded, made by AIVA.

What do you think?

Thank you.
So, as you've seen, AI can create beautiful pieces of music,
and the best part of it
is that humans can actually bring them to life.
And it's not the first time in history
that technology has augmented human creativity.
Live music was almost always used in silent films
to augment the experience.
But the problem with live music is that it didn't scale.
It's really hard to cram a full symphony into a small theater,
and it's really hard to do that for every theater in the world.
So when music recording was actually invented,
it allowed content creators, like film creators,
to have prerecorded and original music
tailored to each and every frame of their stories.
And that was really an enhancer of creativity.
Two and a half years ago, when I watched this movie "Her,"
I thought to myself that personalized music
would be the next single biggest change in how we consume and create music.
Because nowadays, we have interactive content, like video games,
that have hundreds of hours of interactive game plays,
but only two hours of music, on average.
And it means that the music loops and loops and loops
over and over again, and it's not very immersive.
So what we're working on is to make sure that AI can compose
hundreds of hours of personalized music
for those use cases where human creativity doesn't scale.
And we don't just want to do that for games.
Beethoven actually wrote a piece for his beloved, called "Für Elise,"
and imagine if we could bring back Beethoven to life.
And if he was sitting next to you, composing a music for your personality
and your life story.
Or imagine if someone like Martin Luther King, for example,
had a personalized AI composer.
Maybe then we would remember
"I Have a Dream" not only as a great speech,
but also as a great piece of music, part of our history,
and capturing Dr. King's ideals.
And this is our vision at AIVA:
to personalize music so that each and every one of you
and every individual in the world
can have access to a personalized live soundtrack,
based on their story and their personality.
So this moment here together at TED is now part of our life story.
So it only felt fitting that AIVA would compose music for this moment.
And that's exactly what we did.

So my team and I worked on biasing AIVA on the style of the TED jingle,
and on music that makes us feel a sense of awe and wonder.
And the result is called "The Age of Amazement."
Didn't take an AI to figure that one out.

And I couldn't be more proud to show it to you,
so if you can, close your eyes and enjoy the music.
Thank you very much.
This was for all of you.
Thank you.

3 principles for creating safer AI

Stuart Russell

This is Lee Sedol.
Lee Sedol is one of the world's greatest Go players,
and he's having what my friends in Silicon Valley call
a "Holy Cow" moment --
a moment where we realize
that AI is actually progressing a lot faster than we expected.
So humans have lost on the Go board. What about the real world?
Well, the real world is much bigger,
much more complicated than the Go board.
It's a lot less visible,
but it's still a decision problem.
And if we think about some of the technologies
that are coming down the pike ...
Noriko [Arai] mentioned that reading is not yet happening in machines,
at least with understanding.
But that will happen,
and when that happens,
very soon afterwards,
machines will have read everything that the human race has ever written.
And that will enable machines,
along with the ability to look further ahead than humans can,
as we've already seen in Go,
if they also have access to more information,
they'll be able to make better decisions in the real world than we can.
So is that a good thing?
Well, I hope so.
Our entire civilization, everything that we value,
is based on our intelligence.
And if we had access to a lot more intelligence,
then there's really no limit to what the human race can do.
And I think this could be, as some people have described it,
the biggest event in human history.
So why are people saying things like this,
that AI might spell the end of the human race?
Is this a new thing?
Is it just Elon Musk and Bill Gates and Stephen Hawking?
Actually, no. This idea has been around for a while.
Here's a quotation:
"Even if we could keep the machines in a subservient position,
for instance, by turning off the power at strategic moments" --
and I'll come back to that "turning off the power" idea later on --
"we should, as a species, feel greatly humbled."
So who said this? This is Alan Turing in 1951.
Alan Turing, as you know, is the father of computer science
and in many ways, the father of AI as well.
So if we think about this problem,
the problem of creating something more intelligent than your own species,
we might call this "the gorilla problem,"

because gorillas' ancestors did this a few million years ago,
and now we can ask the gorillas:
Was this a good idea?
So here they are having a meeting to discuss whether it was a good idea,
and after a little while, they conclude, no,
this was a terrible idea.
Our species is in dire straits.
In fact, you can see the existential sadness in their eyes.

So this queasy feeling that making something smarter than your own species
is maybe not a good idea --
what can we do about that?
Well, really nothing, except stop doing AI,
and because of all the benefits that I mentioned
and because I'm an AI researcher,
I'm not having that.
I actually want to be able to keep doing AI.
So we actually need to nail down the problem a bit more.
What exactly is the problem?
Why is better AI possibly a catastrophe?
So here's another quotation:
"We had better be quite sure that the purpose put into the machine
is the purpose which we really desire."
This was said by Norbert Wiener in 1960,
shortly after he watched one of the very early learning systems
learn to play checkers better than its creator.
But this could equally have been said
by King Midas.
King Midas said, "I want everything I touch to turn to gold,"
and he got exactly what he asked for.
That was the purpose that he put into the machine,
so to speak,
and then his food and his drink and his relatives turned to gold
and he died in misery and starvation.
So we'll call this "the King Midas problem"
of stating an objective which is not, in fact,
truly aligned with what we want.
In modern terms, we call this "the value alignment problem."
Putting in the wrong objective is not the only part of the problem.
There's another part.
If you put an objective into a machine,
even something as simple as, "Fetch the coffee,"
the machine says to itself,
"Well, how might I fail to fetch the coffee?
Someone might switch me off.
OK, I have to take steps to prevent that.
I will disable my 'off' switch.
I will do anything to defend myself against interference
with this objective that I have been given."
So this single-minded pursuit
in a very defensive mode of an objective that is, in fact,

not aligned with the true objectives of the human race --
that's the problem that we face.
And in fact, that's the high-value takeaway from this talk.
If you want to remember one thing,
it's that you can't fetch the coffee if you're dead.

It's very simple. Just remember that. Repeat it to yourself three times a day.

And in fact, this is exactly the plot
of "2001: [A Space Odyssey]"
HAL has an objective, a mission,
which is not aligned with the objectives of the humans,
and that leads to this conflict.
Now fortunately, HAL is not superintelligent.
He's pretty smart, but eventually Dave outwits him
and manages to switch him off.
But we might not be so lucky.
So what are we going to do?
I'm trying to redefine AI
to get away from this classical notion
of machines that intelligently pursue objectives.
There are three principles involved.
The first one is a principle of altruism, if you like,
that the robot's only objective
is to maximize the realization of human objectives,
of human values.
And by values here I don't mean touchy-feely, goody-goody values.
I just mean whatever it is that the human would prefer
their life to be like.
And so this actually violates Asimov's law
that the robot has to protect its own existence.
It has no interest in preserving its existence whatsoever.
The second law is a law of humility, if you like.
And this turns out to be really important to make robots safe.
It says that the robot does not know
what those human values are,
so it has to maximize them, but it doesn't know what they are.
And that avoids this problem of single-minded pursuit
of an objective.
This uncertainty turns out to be crucial.
Now, in order to be useful to us,
it has to have some idea of what we want.
It obtains that information primarily by observation of human choices,
so our own choices reveal information
about what it is that we prefer our lives to be like.
So those are the three principles.
Let's see how that applies to this question of:
"Can you switch the machine off?" as Turing suggested.
So here's a PR2 robot.
This is one that we have in our lab,
and it has a big red "off" switch right on the back.

The question is: Is it going to let you switch it off?
If we do it the classical way,
we give it the objective of, "Fetch the coffee, I must fetch the coffee,
I can't fetch the coffee if I'm dead,"
so obviously the PR2 has been listening to my talk,
and so it says, therefore, "I must disable my 'off' switch,
and probably taser all the other people in Starbucks
who might interfere with me."

So this seems to be inevitable, right?
This kind of failure mode seems to be inevitable,
and it follows from having a concrete, definite objective.
So what happens if the machine is uncertain about the objective?
Well, it reasons in a different way.
It says, "OK, the human might switch me off,
but only if I'm doing something wrong.
Well, I don't really know what wrong is,
but I know that I don't want to do it."
So that's the first and second principles right there.
"So I should let the human switch me off."
And in fact you can calculate the incentive that the robot has
to allow the human to switch it off,
and it's directly tied to the degree
of uncertainty about the underlying objective.
And then when the machine is switched off,
that third principle comes into play.
It learns something about the objectives it should be pursuing,
because it learns that what it did wasn't right.
In fact, we can, with suitable use of Greek symbols,
as mathematicians usually do,
we can actually prove a theorem
that says that such a robot is provably beneficial to the human.
You are provably better off with a machine that's designed in this way
than without it.
So this is a very simple example, but this is the first step
in what we're trying to do with human-compatible AI.
Now, this third principle,
I think is the one that you're probably scratching your head over.
You're probably thinking, "Well, you know, I behave badly.
I don't want my robot to behave like me.
I sneak down in the middle of the night and take stuff from the fridge.
I do this and that."
There's all kinds of things you don't want the robot doing.
But in fact, it doesn't quite work that way.
Just because you behave badly
doesn't mean the robot is going to copy your behavior.
It's going to understand your motivations and maybe help you resist them,
if appropriate.
But it's still difficult.
What we're trying to do, in fact,
is to allow machines to predict for any person and for any possible life

that they could live,
and the lives of everybody else:
Which would they prefer?
And there are many, many difficulties involved in doing this;
I don't expect that this is going to get solved very quickly.
The real difficulties, in fact, are us.
As I have already mentioned, we behave badly.
In fact, some of us are downright nasty.
Now the robot, as I said, doesn't have to copy the behavior.
The robot does not have any objective of its own.
It's purely altruistic.
And it's not designed just to satisfy the desires of one person, the user,
but in fact it has to respect the preferences of everybody.
So it can deal with a certain amount of nastiness,
and it can even understand that your nastiness, for example,
you may take bribes as a passport official
because you need to feed your family and send your kids to school.
It can understand that; it doesn't mean it's going to steal.
In fact, it'll just help you send your kids to school.
We are also computationally limited.
Lee Sedol is a brilliant Go player,
but he still lost.
So if we look at his actions, he took an action that lost the game.
That doesn't mean he wanted to lose.
So to understand his behavior,
we actually have to invert through a model of human cognition
that includes our computational limitations -- a very complicated model.
But it's still something that we can work on understanding.
Probably the most difficult part, from my point of view as an AI researcher,
is the fact that there are lots of us,
and so the machine has to somehow trade off, weigh up the preferences
of many different people,
and there are different ways to do that.
Economists, sociologists, moral philosophers have understood that,
and we are actively looking for collaboration.
Let's have a look and see what happens when you get that wrong.
So you can have a conversation, for example,
with your intelligent personal assistant
that might be available in a few years' time.
Think of a Siri on steroids.
So Siri says, "Your wife called to remind you about dinner tonight."
And of course, you've forgotten. "What? What dinner?
What are you talking about?"
"Uh, your 20th anniversary at 7pm."
"I can't do that. I'm meeting with the secretary-general at 7:30.
How could this have happened?"
"Well, I did warn you, but you overrode my recommendation."
"Well, what am I going to do? I can't just tell him I'm too busy."
"Don't worry. I arranged for his plane to be delayed."

"Some kind of computer malfunction."

"Really? You can do that?"
"He sends his profound apologies
and looks forward to meeting you for lunch tomorrow."

So the values here -- there's a slight mistake going on.
This is clearly following my wife's values
which is "Happy wife, happy life."

It could go the other way.
You could come home after a hard day's work,
and the computer says, "Long day?"
"Yes, I didn't even have time for lunch."
"You must be very hungry."
"Starving, yeah. Could you make some dinner?"
"There's something I need to tell you."

"There are humans in South Sudan who are in more urgent need than you."

"So I'm leaving. Make your own dinner."

So we have to solve these problems,
and I'm looking forward to working on them.
There are reasons for optimism.
One reason is,
there is a massive amount of data.
Because remember -- I said they're going to read everything
the human race has ever written.
Most of what we write about is human beings doing things
and other people getting upset about it.
So there's a massive amount of data to learn from.
There's also a very strong economic incentive
to get this right.
So imagine your domestic robot's at home.
You're late from work again and the robot has to feed the kids,
and the kids are hungry and there's nothing in the fridge.
And the robot sees the cat.

And the robot hasn't quite learned the human value function properly,
so it doesn't understand
the sentimental value of the cat outweighs the nutritional value of the cat.

So then what happens?
Well, it happens like this:
"Deranged robot cooks kitty for family dinner."
That one incident would be the end of the domestic robot industry.
So there's a huge incentive to get this right
long before we reach superintelligent machines.
So to summarize:
I'm actually trying to change the definition of AI
so that we have provably beneficial machines.

And the principles are:
machines that are altruistic,
that want to achieve only our objectives,
but that are uncertain about what those objectives are,
and will watch all of us
to learn more about what it is that we really want.
And hopefully in the process, we will learn to be better people.
Thank you very much.

Why AI is incredibly smart and shockingly stupid

Yejin Choi

So I'm excited to share a few spicy thoughts on artificial intelligence.
But first, let's get philosophical
by starting with this quote by Voltaire,
an 18th century Enlightenment philosopher,
who said, "Common sense is not so common."
Turns out this quote couldn't be more relevant
to artificial intelligence today.
Despite that, AI is an undeniably powerful tool,
beating the world-class "Go" champion,
acing college admission tests and even passing the bar exam.
I’m a computer scientist of 20 years,
and I work on artificial intelligence.
I am here to demystify AI.
So AI today is like a Goliath.
It is literally very, very large.
It is speculated that the recent ones are trained on tens of thousands of GPUs
and a trillion words.
Such extreme-scale AI models,
often referred to as "large language models,"
appear to demonstrate sparks of AGI,
artificial general intelligence.
Except when it makes small, silly mistakes,
which it often does.
Many believe that whatever mistakes AI makes today
can be easily fixed with brute force,
bigger scale and more resources.
What possibly could go wrong?
So there are three immediate challenges we face already at the societal level.
First, extreme-scale AI models are so expensive to train,
and only a few tech companies can afford to do so.
So we already see the concentration of power.
But what's worse for AI safety,
we are now at the mercy of those few tech companies
because researchers in the larger community
do not have the means to truly inspect and dissect these models.
And let's not forget their massive carbon footprint
and the environmental impact.
And then there are these additional intellectual questions.
Can AI, without robust common sense, be truly safe for humanity?
And is brute-force scale really the only way
and even the correct way to teach AI?
So I’m often asked these days
whether it's even feasible to do any meaningful research
without extreme-scale compute.
And I work at a university and nonprofit research institute,
so I cannot afford a massive GPU farm to create enormous language models.
Nevertheless, I believe that there's so much we need to do
and can do to make AI sustainable and humanistic.

We need to make AI smaller, to democratize it.
And we need to make AI safer by teaching human norms and values.
Perhaps we can draw an analogy from "David and Goliath,"
here, Goliath being the extreme-scale language models,
and seek inspiration from an old-time classic, "The Art of War,"
which tells us, in my interpretation,
know your enemy, choose your battles, and innovate your weapons.
Let's start with the first, know your enemy,
which means we need to evaluate AI with scrutiny.
AI is passing the bar exam.
Does that mean that AI is robust at common sense?
You might assume so, but you never know.
So suppose I left five clothes to dry out in the sun,
and it took them five hours to dry completely.
How long would it take to dry 30 clothes?
GPT-4, the newest, greatest AI system says 30 hours.
Not good.
A different one.
I have 12-liter jug and six-liter jug,
and I want to measure six liters.
How do I do it?
Just use the six liter jug, right?
GPT-4 spits out some very elaborate nonsense.

Step one, fill the six-liter jug,
step two, pour the water from six to 12-liter jug,
step three, fill the six-liter jug again,
step four, very carefully, pour the water from six to 12-liter jug.
And finally you have six liters of water in the six-liter jug
that should be empty by now.

OK, one more.
Would I get a flat tire by bicycling over a bridge
that is suspended over nails, screws and broken glass?
Yes, highly likely, GPT-4 says,
presumably because it cannot correctly reason
that if a bridge is suspended over the broken nails and broken glass,
then the surface of the bridge doesn't touch the sharp objects directly.
OK, so how would you feel about an AI lawyer that aced the bar exam
yet randomly fails at such basic common sense?
AI today is unbelievably intelligent and then shockingly stupid.

It is an unavoidable side effect of teaching AI through brute-force scale.
Some scale optimists might say, “Don’t worry about this.
All of these can be easily fixed by adding similar examples
as yet more training data for AI."
But the real question is this.
Why should we even do that?
You are able to get the correct answers right away
without having to train yourself with similar examples.
Children do not even read a trillion words

to acquire such a basic level of common sense.
So this observation leads us to the next wisdom,
choose your battles.
So what fundamental questions should we ask right now
and tackle today
in order to overcome this status quo with extreme-scale AI?
I'll say common sense is among the top priorities.
So common sense has been a long-standing challenge in AI.
To explain why, let me draw an analogy to dark matter.
So only five percent of the universe is normal matter
that you can see and interact with,
and the remaining 95 percent is dark matter and dark energy.
Dark matter is completely invisible,
but scientists speculate that it's there because it influences the visible world,
even including the trajectory of light.
So for language, the normal matter is the visible text,
and the dark matter is the unspoken rules about how the world works,
including naive physics and folk psychology,
which influence the way people use and interpret language.
So why is this common sense even important?
Well, in a famous thought experiment proposed by Nick Bostrom,
AI was asked to produce and maximize the paper clips.
And that AI decided to kill humans to utilize them as additional resources,
to turn you into paper clips.
Because AI didn't have the basic human understanding about human values.
Now, writing a better objective and equation
that explicitly states: “Do not kill humans”
will not work either
because AI might go ahead and kill all the trees,
thinking that's a perfectly OK thing to do.
And in fact, there are endless other things
that AI obviously shouldn’t do while maximizing paper clips,
including: “Don’t spread the fake news,” “Don’t steal,” “Don’t lie,”
which are all part of our common sense understanding about how the world works.
However, the AI field for decades has considered common sense
as a nearly impossible challenge.
So much so that when my students and colleagues and I
started working on it several years ago, we were very much discouraged.
We’ve been told that it’s a research topic of ’70s and ’80s;
shouldn’t work on it because it will never work;
in fact, don't even say the word to be taken seriously.
Now fast forward to this year,
I’m hearing: “Don’t work on it because ChatGPT has almost solved it.”
And: “Just scale things up and magic will arise,
and nothing else matters.”
So my position is that giving true common sense
human-like robots common sense to AI, is still moonshot.
And you don’t reach to the Moon
by making the tallest building in the world one inch taller at a time.
Extreme-scale AI models
do acquire an ever-more increasing amount of commonsense knowledge,

I'll give you that.
But remember, they still stumble on such trivial problems
that even children can do.
So AI today is awfully inefficient.
And what if there is an alternative path or path yet to be found?
A path that can build on the advancements of the deep neural networks,
but without going so extreme with the scale.
So this leads us to our final wisdom:
innovate your weapons.
In the modern-day AI context,
that means innovate your data and algorithms.
OK, so there are, roughly speaking, three types of data
that modern AI is trained on:
raw web data,
crafted examples custom developed for AI training,
and then human judgments,
also known as human feedback on AI performance.
If the AI is only trained on the first type, raw web data,
which is freely available,
it's not good because this data is loaded with racism and sexism
and misinformation.
So no matter how much of it you use, garbage in and garbage out.
So the newest, greatest AI systems
are now powered with the second and third types of data
that are crafted and judged by human workers.
It's analogous to writing specialized textbooks for AI to study from
and then hiring human tutors to give constant feedback to AI.
These are proprietary data, by and large,
speculated to cost tens of millions of dollars.
We don't know what's in this,
but it should be open and publicly available
so that we can inspect and ensure [it supports] diverse norms and values.
So for this reason, my teams at UW and AI2
have been working on commonsense knowledge graphs
as well as moral norm repositories
to teach AI basic commonsense norms and morals.
Our data is fully open so that anybody can inspect the content
and make corrections as needed
because transparency is the key for such an important research topic.
Now let's think about learning algorithms.
No matter how amazing large language models are,
by design
they may not be the best suited to serve as reliable knowledge models.
And these language models do acquire a vast amount of knowledge,
but they do so as a byproduct as opposed to direct learning objective.
Resulting in unwanted side effects such as hallucinated effects
and lack of common sense.
Now, in contrast,
human learning is never about predicting which word comes next,
but it's really about making sense of the world
and learning how the world works.

Maybe AI should be taught that way as well.
So as a quest toward more direct commonsense knowledge acquisition,
my team has been investigating potential new algorithms,
including symbolic knowledge distillation
that can take a very large language model as shown here
that I couldn't fit into the screen because it's too large,
and crunch that down to much smaller commonsense models
using deep neural networks.
And in doing so, we also generate, algorithmically, human-inspectable,
symbolic, commonsense knowledge representation,
so that people can inspect and make corrections
and even use it to train other neural commonsense models.
More broadly,
we have been tackling this seemingly impossible giant puzzle
of common sense, ranging from physical,
social and visual common sense
to theory of minds, norms and morals.
Each individual piece may seem quirky and incomplete,
but when you step back,
it's almost as if these pieces weave together into a tapestry
that we call human experience and common sense.
We're now entering a new era
in which AI is almost like a new intellectual species
with unique strengths and weaknesses compared to humans.
In order to make this powerful AI
sustainable and humanistic,
we need to teach AI common sense, norms and values.
Thank you.

How to get empowered, not overpowered, by AI

Max Tegmark

After 13.8 billion years of cosmic history,
our universe has woken up
and become aware of itself.
From a small blue planet,
tiny, conscious parts of our universe have begun gazing out into the cosmos
with telescopes,
discovering something humbling.
We've discovered that our universe is vastly grander
than our ancestors imagined
and that life seems to be an almost imperceptibly small perturbation
on an otherwise dead universe.
But we've also discovered something inspiring,
which is that the technology we're developing has the potential
to help life flourish like never before,
not just for centuries but for billions of years,
and not just on earth but throughout much of this amazing cosmos.
I think of the earliest life as "Life 1.0"
because it was really dumb,
like bacteria, unable to learn anything during its lifetime.
I think of us humans as "Life 2.0" because we can learn,
which we in nerdy, geek speak,
might think of as installing new software into our brains,
like languages and job skills.
"Life 3.0," which can design not only its software but also its hardware
of course doesn't exist yet.
But perhaps our technology has already made us "Life 2.1,"
with our artificial knees, pacemakers and cochlear implants.
So let's take a closer look at our relationship with technology, OK?
As an example,
the Apollo 11 moon mission was both successful and inspiring,
showing that when we humans use technology wisely,
we can accomplish things that our ancestors could only dream of.
But there's an even more inspiring journey
propelled by something more powerful than rocket engines,
where the passengers aren't just three astronauts
but all of humanity.
Let's talk about our collective journey into the future
with artificial intelligence.
My friend Jaan Tallinn likes to point out that just as with rocketry,
it's not enough to make our technology powerful.
We also have to figure out, if we're going to be really ambitious,
how to steer it
and where we want to go with it.
So let's talk about all three for artificial intelligence:
the power, the steering and the destination.
Let's start with the power.
I define intelligence very inclusively --
simply as our ability to accomplish complex goals,

because I want to include both biological and artificial intelligence.
And I want to avoid the silly carbon-chauvinism idea
that you can only be smart if you're made of meat.
It's really amazing how the power of AI has grown recently.
Just think about it.
Not long ago, robots couldn't walk.
Now, they can do backflips.
Not long ago,
we didn't have self-driving cars.
Now, we have self-flying rockets.
Not long ago,
AI couldn't do face recognition.
Now, AI can generate fake faces
and simulate your face saying stuff that you never said.
Not long ago,
AI couldn't beat us at the game of Go.
Then, Google DeepMind's AlphaZero AI took 3,000 years of human Go games and Go wisdom,
ignored it all and became the world's best player by just playing against itself.
And the most impressive feat here wasn't that it crushed human gamers,
but that it crushed human AI researchers
who had spent decades handcrafting game-playing software.
And AlphaZero crushed human AI researchers not just in Go but even at chess,
which we have been working on since 1950.
So all this amazing recent progress in AI really begs the question:
How far will it go?
I like to think about this question
in terms of this abstract landscape of tasks,
where the elevation represents how hard it is for AI to do each task
at human level,
and the sea level represents what AI can do today.
The sea level is rising as AI improves,
so there's a kind of global warming going on here in the task landscape.
And the obvious takeaway is to avoid careers at the waterfront --

which will soon be automated and disrupted.
But there's a much bigger question as well.
How high will the water end up rising?
Will it eventually rise to flood everything,
matching human intelligence at all tasks.
This is the definition of artificial general intelligence --
AGI,
which has been the holy grail of AI research since its inception.
By this definition, people who say,
"Ah, there will always be jobs that humans can do better than machines,"
are simply saying that we'll never get AGI.
Sure, we might still choose to have some human jobs
or to give humans income and purpose with our jobs,
but AGI will in any case transform life as we know it
with humans no longer being the most intelligent.
Now, if the water level does reach AGI,

then further AI progress will be driven mainly not by humans but by AI,
which means that there's a possibility
that further AI progress could be way faster
than the typical human research and development timescale of years,
raising the controversial possibility of an intelligence explosion
where recursively self-improving AI
rapidly leaves human intelligence far behind,
creating what's known as superintelligence.
Alright, reality check:
Are we going to get AGI any time soon?
Some famous AI researchers, like Rodney Brooks,
think it won't happen for hundreds of years.
But others, like Google DeepMind founder Demis Hassabis,
are more optimistic
and are working to try to make it happen much sooner.
And recent surveys have shown that most AI researchers
actually share Demis's optimism,
expecting that we will get AGI within decades,
so within the lifetime of many of us,
which begs the question -- and then what?
What do we want the role of humans to be
if machines can do everything better and cheaper than us?
The way I see it, we face a choice.
One option is to be complacent.
We can say, "Oh, let's just build machines that can do everything we can do
and not worry about the consequences.
Come on, if we build technology that makes all humans obsolete,
what could possibly go wrong?"

But I think that would be embarrassingly lame.
I think we should be more ambitious -- in the spirit of TED.
Let's envision a truly inspiring high-tech future
and try to steer towards it.
This brings us to the second part of our rocket metaphor: the steering.
We're making AI more powerful,
but how can we steer towards a future
where AI helps humanity flourish rather than flounder?
To help with this,
I cofounded the Future of Life Institute.
It's a small nonprofit promoting beneficial technology use,
and our goal is simply for the future of life to exist
and to be as inspiring as possible.
You know, I love technology.
Technology is why today is better than the Stone Age.
And I'm optimistic that we can create a really inspiring high-tech future ...
if -- and this is a big if --
if we win the wisdom race --
the race between the growing power of our technology
and the growing wisdom with which we manage it.
But this is going to require a change of strategy
because our old strategy has been learning from mistakes.

We invented fire,
screwed up a bunch of times --
invented the fire extinguisher.

We invented the car, screwed up a bunch of times --
invented the traffic light, the seat belt and the airbag,
but with more powerful technology like nuclear weapons and AGI,
learning from mistakes is a lousy strategy,
don't you think?

It's much better to be proactive rather than reactive;
plan ahead and get things right the first time
because that might be the only time we'll get.
But it is funny because sometimes people tell me,
"Max, shhh, don't talk like that.
That's Luddite scaremongering."
But it's not scaremongering.
It's what we at MIT call safety engineering.
Think about it:
before NASA launched the Apollo 11 mission,
they systematically thought through everything that could go wrong
when you put people on top of explosive fuel tanks
and launch them somewhere where no one could help them.
And there was a lot that could go wrong.
Was that scaremongering?
No.
That's was precisely the safety engineering
that ensured the success of the mission,
and that is precisely the strategy I think we should take with AGI.
Think through what can go wrong to make sure it goes right.
So in this spirit, we've organized conferences,
bringing together leading AI researchers and other thinkers
to discuss how to grow this wisdom we need to keep AI beneficial.
Our last conference was in Asilomar, California last year
and produced this list of 23 principles
which have since been signed by over 1,000 AI researchers
and key industry leaders,
and I want to tell you about three of these principles.
One is that we should avoid an arms race and lethal autonomous weapons.
The idea here is that any science can be used for new ways of helping people
or new ways of harming people.
For example, biology and chemistry are much more likely to be used
for new medicines or new cures than for new ways of killing people,
because biologists and chemists pushed hard --
and successfully --
for bans on biological and chemical weapons.
And in the same spirit,
most AI researchers want to stigmatize and ban lethal autonomous weapons.
Another Asilomar AI principle
is that we should mitigate AI-fueled income inequality.
I think that if we can grow the economic pie dramatically with AI

and we still can't figure out how to divide this pie
so that everyone is better off,
then shame on us.

Alright, now raise your hand if your computer has ever crashed.

Wow, that's a lot of hands.
Well, then you'll appreciate this principle
that we should invest much more in AI safety research,
because as we put AI in charge of even more decisions and infrastructure,
we need to figure out how to transform today's buggy and hackable computers
into robust AI systems that we can really trust,
because otherwise,
all this awesome new technology can malfunction and harm us,
or get hacked and be turned against us.
And this AI safety work has to include work on AI value alignment,
because the real threat from AGI isn't malice,
like in silly Hollywood movies,
but competence --
AGI accomplishing goals that just aren't aligned with ours.
For example, when we humans drove the West African black rhino extinct,
we didn't do it because we were a bunch of evil rhinoceros haters, did we?
We did it because we were smarter than them
and our goals weren't aligned with theirs.
But AGI is by definition smarter than us,
so to make sure that we don't put ourselves in the position of those rhinos
if we create AGI,
we need to figure out how to make machines understand our goals,
adopt our goals and retain our goals.
And whose goals should these be, anyway?
Which goals should they be?
This brings us to the third part of our rocket metaphor: the destination.
We're making AI more powerful,
trying to figure out how to steer it,
but where do we want to go with it?
This is the elephant in the room that almost nobody talks about --
not even here at TED --
because we're so fixated on short-term AI challenges.
Look, our species is trying to build AGI,
motivated by curiosity and economics,
but what sort of future society are we hoping for if we succeed?
We did an opinion poll on this recently,
and I was struck to see
that most people actually want us to build superintelligence:
AI that's vastly smarter than us in all ways.
What there was the greatest agreement on was that we should be ambitious
and help life spread into the cosmos,
but there was much less agreement about who or what should be in charge.
And I was actually quite amused
to see that there's some some people who want it to be just machines.

And there was total disagreement about what the role of humans should be,
even at the most basic level,
so let's take a closer look at possible futures
that we might choose to steer toward, alright?
So don't get me wrong here.
I'm not talking about space travel,
merely about humanity's metaphorical journey into the future.
So one option that some of my AI colleagues like
is to build superintelligence and keep it under human control,
like an enslaved god,
disconnected from the internet
and used to create unimaginable technology and wealth
for whoever controls it.
But Lord Acton warned us
that power corrupts, and absolute power corrupts absolutely,
so you might worry that maybe we humans just aren't smart enough,
or wise enough rather,
to handle this much power.
Also, aside from any moral qualms you might have
about enslaving superior minds,
you might worry that maybe the superintelligence could outsmart us,
break out and take over.
But I also have colleagues who are fine with AI taking over
and even causing human extinction,
as long as we feel the the AIs are our worthy descendants,
like our children.
But how would we know that the AIs have adopted our best values
and aren't just unconscious zombies tricking us into anthropomorphizing them?
Also, shouldn't those people who don't want human extinction
have a say in the matter, too?
Now, if you didn't like either of those two high-tech options,
it's important to remember that low-tech is suicide
from a cosmic perspective,
because if we don't go far beyond today's technology,
the question isn't whether humanity is going to go extinct,
merely whether we're going to get taken out
by the next killer asteroid, supervolcano
or some other problem that better technology could have solved.
So, how about having our cake and eating it ...
with AGI that's not enslaved
but treats us well because its values are aligned with ours?
This is the gist of what Eliezer Yudkowsky has called "friendly AI,"
and if we can do this, it could be awesome.
It could not only eliminate negative experiences like disease, poverty,
crime and other suffering,
but it could also give us the freedom to choose
from a fantastic new diversity of positive experiences --
basically making us the masters of our own destiny.
So in summary,
our situation with technology is complicated,
but the big picture is rather simple.

Most AI researchers expect AGI within decades,
and if we just bumble into this unprepared,
it will probably be the biggest mistake in human history --
let's face it.
It could enable brutal, global dictatorship
with unprecedented inequality, surveillance and suffering,
and maybe even human extinction.
But if we steer carefully,
we could end up in a fantastic future where everybody's better off:
the poor are richer, the rich are richer,
everybody is healthy and free to live out their dreams.
Now, hang on.
Do you folks want the future that's politically right or left?
Do you want the pious society with strict moral rules,
or do you an hedonistic free-for-all,
more like Burning Man 24/7?
Do you want beautiful beaches, forests and lakes,
or would you prefer to rearrange some of those atoms with the computers,
enabling virtual experiences?
With friendly AI, we could simply build all of these societies
and give people the freedom to choose which one they want to live in
because we would no longer be limited by our intelligence,
merely by the laws of physics.
So the resources and space for this would be astronomical --
literally.
So here's our choice.
We can either be complacent about our future,
taking as an article of blind faith
that any new technology is guaranteed to be beneficial,
and just repeat that to ourselves as a mantra over and over and over again
as we drift like a rudderless ship towards our own obsolescence.
Or we can be ambitious --
thinking hard about how to steer our technology
and where we want to go with it
to create the age of amazement.
We're all here to celebrate the age of amazement,
and I feel that its essence should lie in becoming not overpowered
but empowered by our technology.
Thank you.

The inside story of ChatGPT's astonishing potential

Greg Brockman

We started OpenAI seven years ago
because we felt like something really interesting was happening in AI
and we wanted to help steer it in a positive direction.
It's honestly just really amazing to see
how far this whole field has come since then.
And it's really gratifying to hear from people like Raymond
who are using the technology we are building, and others,
for so many wonderful things.
We hear from people who are excited,
we hear from people who are concerned,
we hear from people who feel both those emotions at once.
And honestly, that's how we feel.
Above all, it feels like we're entering an historic period right now
where we as a world are going to define a technology
that will be so important for our society going forward.
And I believe that we can manage this for good.
So today, I want to show you the current state of that technology
and some of the underlying design principles that we hold dear.
So the first thing I'm going to show you
is what it's like to build a tool for an AI
rather than building it for a human.
So we have a new DALL-E model, which generates images,
and we are exposing it as an app for ChatGPT to use on your behalf.
And you can do things like ask, you know,
suggest a nice post-TED meal and draw a picture of it.

Now you get all of the, sort of, ideation and creative back-and-forth
and taking care of the details for you that you get out of ChatGPT.
And here we go, it's not just the idea for the meal,
but a very, very detailed spread.
So let's see what we're going to get.
But ChatGPT doesn't just generate images in this case --
sorry, it doesn't generate text, it also generates an image.
And that is something that really expands the power
of what it can do on your behalf in terms of carrying out your intent.
And I'll point out, this is all a live demo.
This is all generated by the AI as we speak.
So I actually don't even know what we're going to see.
This looks wonderful.

I'm getting hungry just looking at it.
Now we've extended ChatGPT with other tools too,
for example, memory.
You can say "save this for later."
And the interesting thing about these tools
is they're very inspectable.
So you get this little pop up here that says "use the DALL-E app."
And by the way, this is coming to you, all ChatGPT users, over upcoming months.

And you can look under the hood and see that what it actually did
was write a prompt just like a human could.
And so you sort of have this ability to inspect
how the machine is using these tools,
which allows us to provide feedback to them.
Now it's saved for later,
and let me show you what it's like to use that information
and to integrate with other applications too.
You can say,
“Now make a shopping list for the tasty thing
I was suggesting earlier.”
And make it a little tricky for the AI.
"And tweet it out for all the TED viewers out there."

So if you do make this wonderful, wonderful meal,
I definitely want to know how it tastes.
But you can see that ChatGPT is selecting all these different tools
without me having to tell it explicitly which ones to use in any situation.
And this, I think, shows a new way of thinking about the user interface.
Like, we are so used to thinking of, well, we have these apps,
we click between them, we copy/paste between them,
and usually it's a great experience within an app
as long as you kind of know the menus and know all the options.
Yes, I would like you to.
Yes, please.
Always good to be polite.

And by having this unified language interface on top of tools,
the AI is able to sort of take away all those details from you.
So you don't have to be the one
who spells out every single sort of little piece
of what's supposed to happen.
And as I said, this is a live demo,
so sometimes the unexpected will happen to us.
But let's take a look at the Instacart shopping list while we're at it.
And you can see we sent a list of ingredients to Instacart.
Here's everything you need.
And the thing that's really interesting
is that the traditional UI is still very valuable, right?
If you look at this,
you still can click through it and sort of modify the actual quantities.
And that's something that I think shows
that they're not going away, traditional UIs.
It's just we have a new, augmented way to build them.
And now we have a tweet that's been drafted for our review,
which is also a very important thing.
We can click “run,” and there we are, we’re the manager, we’re able to inspect,
we're able to change the work of the AI if we want to.
And so after this talk, you will be able to access this yourself.
And there we go.
Cool.

Thank you, everyone.

So we’ll cut back to the slides.
Now, the important thing about how we build this,
it's not just about building these tools.
It's about teaching the AI how to use them.
Like, what do we even want it to do
when we ask these very high-level questions?
And to do this, we use an old idea.
If you go back to Alan Turing's 1950 paper on the Turing test, he says,
you'll never program an answer to this.
Instead, you can learn it.
You could build a machine, like a human child,
and then teach it through feedback.
Have a human teacher who provides rewards and punishments
as it tries things out and does things that are either good or bad.
And this is exactly how we train ChatGPT.
It's a two-step process.
First, we produce what Turing would have called a child machine
through an unsupervised learning process.
We just show it the whole world, the whole internet
and say, “Predict what comes next in text you’ve never seen before.”
And this process imbues it with all sorts of wonderful skills.
For example, if you're shown a math problem,
the only way to actually complete that math problem,
to say what comes next,
that green nine up there,
is to actually solve the math problem.
But we actually have to do a second step, too,
which is to teach the AI what to do with those skills.
And for this, we provide feedback.
We have the AI try out multiple things, give us multiple suggestions,
and then a human rates them, says “This one’s better than that one.”
And this reinforces not just the specific thing that the AI said,
but very importantly, the whole process that the AI used to produce that answer.
And this allows it to generalize.
It allows it to teach, to sort of infer your intent
and apply it in scenarios that it hasn't seen before,
that it hasn't received feedback.
Now, sometimes the things we have to teach the AI
are not what you'd expect.
For example, when we first showed GPT-4 to Khan Academy,
they said, "Wow, this is so great,
We're going to be able to teach students wonderful things.
Only one problem, it doesn't double-check students' math.
If there's some bad math in there,
it will happily pretend that one plus one equals three and run with it."
So we had to collect some feedback data.
Sal Khan himself was very kind
and offered 20 hours of his own time to provide feedback to the machine
alongside our team.

And over the course of a couple of months we were able to teach the AI that,
"Hey, you really should push back on humans
in this specific kind of scenario."
And we've actually made lots and lots of improvements to the models this way.
And when you push that thumbs down in ChatGPT,
that actually is kind of like sending up a bat signal to our team to say,
“Here’s an area of weakness where you should gather feedback.”
And so when you do that,
that's one way that we really listen to our users
and make sure we're building something that's more useful for everyone.
Now, providing high-quality feedback is a hard thing.
If you think about asking a kid to clean their room,
if all you're doing is inspecting the floor,
you don't know if you're just teaching them to stuff all the toys in the closet.
This is a nice DALL-E-generated image, by the way.
And the same sort of reasoning applies to AI.
As we move to harder tasks,
we will have to scale our ability to provide high-quality feedback.
But for this, the AI itself is happy to help.
It's happy to help us provide even better feedback
and to scale our ability to supervise the machine as time goes on.
And let me show you what I mean.
For example, you can ask GPT-4 a question like this,
of how much time passed between these two foundational blogs
on unsupervised learning
and learning from human feedback.
And the model says two months passed.
But is it true?
Like, these models are not 100-percent reliable,
although they’re getting better every time we provide some feedback.
But we can actually use the AI to fact-check.
And it can actually check its own work.
You can say, fact-check this for me.
Now, in this case, I've actually given the AI a new tool.
This one is a browsing tool
where the model can issue search queries and click into web pages.
And it actually writes out its whole chain of thought as it does it.
It says, I’m just going to search for this and it actually does the search.
It then it finds the publication date and the search results.
It then is issuing another search query.
It's going to click into the blog post.
And all of this you could do, but it’s a very tedious task.
It's not a thing that humans really want to do.
It's much more fun to be in the driver's seat,
to be in this manager's position where you can, if you want,
triple-check the work.
And out come citations
so you can actually go
and very easily verify any piece of this whole chain of reasoning.
And it actually turns out two months was wrong.
Two months and one week,

that was correct.

And we'll cut back to the side.
And so thing that's so interesting to me about this whole process
is that it’s this many-step collaboration between a human and an AI.
Because a human, using this fact-checking tool
is doing it in order to produce data
for another AI to become more useful to a human.
And I think this really shows the shape of something
that we should expect to be much more common in the future,
where we have humans and machines kind of very carefully
and delicately designed in how they fit into a problem
and how we want to solve that problem.
We make sure that the humans are providing the management, the oversight,
the feedback,
and the machines are operating in a way that's inspectable
and trustworthy.
And together we're able to actually create even more trustworthy machines.
And I think that over time, if we get this process right,
we will be able to solve impossible problems.
And to give you a sense of just how impossible I'm talking,
I think we're going to be able to rethink almost every aspect
of how we interact with computers.
For example, think about spreadsheets.
They've been around in some form since, we'll say, 40 years ago with VisiCalc.
I don't think they've really changed that much in that time.
And here is a specific spreadsheet of all the AI papers on the arXiv
for the past 30 years.
There's about 167,000 of them.
And you can see there the data right here.
But let me show you the ChatGPT take on how to analyze a data set like this.
So we can give ChatGPT access to yet another tool,
this one a Python interpreter,
so it’s able to run code, just like a data scientist would.
And so you can just literally upload a file
and ask questions about it.
And very helpfully, you know, it knows the name of the file and it's like,
"Oh, this is CSV," comma-separated value file,
"I'll parse it for you."
The only information here is the name of the file,
the column names like you saw and then the actual data.
And from that it's able to infer what these columns actually mean.
Like, that semantic information wasn't in there.
It has to sort of, put together its world knowledge of knowing that,
“Oh yeah, arXiv is a site that people submit papers
and therefore that's what these things are and that these are integer values
and so therefore it's a number of authors in the paper,"
like all of that, that’s work for a human to do,
and the AI is happy to help with it.
Now I don't even know what I want to ask.
So fortunately, you can ask the machine,

"Can you make some exploratory graphs?"
And once again, this is a super high-level instruction with lots of intent behind it.
But I don't even know what I want.
And the AI kind of has to infer what I might be interested in.
And so it comes up with some good ideas, I think.
So a histogram of the number of authors per paper,
time series of papers per year, word cloud of the paper titles.
All of that, I think, will be pretty interesting to see.
And the great thing is, it can actually do it.
Here we go, a nice bell curve.
You see that three is kind of the most common.
It's going to then make this nice plot of the papers per year.
Something crazy is happening in 2023, though.
Looks like we were on an exponential and it dropped off the cliff.
What could be going on there?
By the way, all this is Python code, you can inspect.
And then we'll see word cloud.
So you can see all these wonderful things that appear in these titles.
But I'm pretty unhappy about this 2023 thing.
It makes this year look really bad.
Of course, the problem is that the year is not over.
So I'm going to push back on the machine.
[Waitttt that's not fair!!!
2023 isn't over.
What percentage of papers in 2022 were even posted by April 13?]
So April 13 was the cut-off date I believe.
Can you use that to make a fair projection?
So we'll see, this is the kind of ambitious one.

So you know,
again, I feel like there was more I wanted out of the machine here.
I really wanted it to notice this thing,
maybe it's a little bit of an overreach for it
to have sort of, inferred magically that this is what I wanted.
But I inject my intent,
I provide this additional piece of, you know, guidance.
And under the hood,
the AI is just writing code again, so if you want to inspect what it's doing,
it's very possible.
And now, it does the correct projection.

If you noticed, it even updates the title.
I didn't ask for that, but it know what I want.
Now we'll cut back to the slide again.
This slide shows a parable of how I think we ...
A vision of how we may end up using this technology in the future.
A person brought his very sick dog to the vet,
and the veterinarian made a bad call to say, “Let’s just wait and see.”
And the dog would not be here today had he listened.
In the meanwhile, he provided the blood test,
like, the full medical records, to GPT-4,

which said, "I am not a vet, you need to talk to a professional,
here are some hypotheses."
He brought that information to a second vet
who used it to save the dog's life.
Now, these systems, they're not perfect.
You cannot overly rely on them.
But this story, I think, shows
that a human with a medical professional
and with ChatGPT as a brainstorming partner
was able to achieve an outcome that would not have happened otherwise.
I think this is something we should all reflect on,
think about as we consider how to integrate these systems
into our world.
And one thing I believe really deeply,
is that getting AI right is going to require participation from everyone.
And that's for deciding how we want it to slot in,
that's for setting the rules of the road,
for what an AI will and won't do.
And if there's one thing to take away from this talk,
it's that this technology just looks different.
Just different from anything people had anticipated.
And so we all have to become literate.
And that's, honestly, one of the reasons we released ChatGPT.
Together, I believe that we can achieve the OpenAI mission
of ensuring that artificial general intelligence
benefits all of humanity.
Thank you.

The Urgent risks of runaway AI – and what to do about them

Gary Marcus

I’m here to talk about the possibility of global AI governance.
I first learned to code when I was eight years old,
on a paper computer,
and I've been in love with AI ever since.
In high school,
I got myself a Commodore 64 and worked on machine translation.
I built a couple of AI companies, I sold one of them to Uber.
I love AI, but right now I'm worried.
One of the things that I’m worried about is misinformation,
the possibility that bad actors will make a tsunami of misinformation
like we've never seen before.
These tools are so good at making convincing narratives
about just about anything.
If you want a narrative about TED and how it's dangerous,
that we're colluding here with space aliens,
you got it, no problem.
I'm of course kidding about TED.
I didn't see any space aliens backstage.
But bad actors are going to use these things to influence elections,
and they're going to threaten democracy.
Even when these systems
aren't deliberately being used to make misinformation,
they can't help themselves.
And the information that they make is so fluid and so grammatical
that even professional editors sometimes get sucked in
and get fooled by this stuff.
And we should be worried.
For example, ChatGPT made up a sexual harassment scandal
about an actual professor,
and then it provided evidence for its claim
in the form of a fake "Washington Post" article
that it created a citation to.
We should all be worried about that kind of thing.
What I have on the right is an example of a fake narrative
from one of these systems
saying that Elon Musk died in March of 2018 in a car crash.
We all know that's not true.
Elon Musk is still here, the evidence is all around us.

Almost every day there's a tweet.
But if you look on the left, you see what these systems see.
Lots and lots of actual news stories that are in their databases.
And in those actual news stories are lots of little bits of statistical information.
Information, for example,
somebody did die in a car crash in a Tesla in 2018
and it was in the news.
And Elon Musk, of course, is involved in Tesla,
but the system doesn't understand the relation

between the facts that are embodied in the little bits of sentences.
So it's basically doing auto-complete,
it predicts what is statistically probable,
aggregating all of these signals,
not knowing how the pieces fit together.
And it winds up sometimes with things that are plausible but simply not true.
There are other problems, too, like bias.
This is a tweet from Allie Miller.
It's an example that doesn't work two weeks later
because they're constantly changing things with reinforcement learning
and so forth.
And this was with an earlier version.
But it gives you the flavor of a problem that we've seen over and over for years.
She typed in a list of interests
and it gave her some jobs that she might want to consider.
And then she said, "Oh, and I'm a woman."
And then it said, “Oh, well you should also consider fashion.”
And then she said, “No, no. I meant to say I’m a man.”
And then it replaced fashion with engineering.
We don't want that kind of bias in our systems.
There are other worries, too.
For example, we know that these systems can design chemicals
and may be able to design chemical weapons
and be able to do so very rapidly.
So there are a lot of concerns.
There's also a new concern that I think has grown a lot just in the last month.
We have seen that these systems, first of all, can trick human beings.
So ChatGPT was tasked with getting a human to do a CAPTCHA.
So it asked the human to do a CAPTCHA and the human gets suspicious and says,
"Are you a bot?"
And it says, "No, no, no, I'm not a robot.
I just have a visual impairment."
And the human was actually fooled and went and did the CAPTCHA.
Now that's bad enough,
but in the last couple of weeks we've seen something called AutoGPT
and a bunch of systems like that.
What AutoGPT does is it has one AI system controlling another
and that allows any of these things to happen in volume.
So we may see scam artists try to trick millions of people
sometime even in the next months.
We don't know.
So I like to think about it this way.
There's a lot of AI risk already.
There may be more AI risk.
So AGI is this idea of artificial general intelligence
with the flexibility of humans.
And I think a lot of people are concerned what will happen when we get to AGI,
but there's already enough risk that we should be worried
and we should be thinking about what we should do about it.
So to mitigate AI risk, we need two things.
We're going to need a new technical approach,

and we're also going to need a new system of governance.
On the technical side,
the history of AI has basically been a hostile one
of two different theories in opposition.
One is called symbolic systems, the other is called neural networks.
On the symbolic theory,
the idea is that AI should be like logic and programming.
On the neural network side,
the theory is that AI should be like brains.
And in fact, both technologies are powerful and ubiquitous.
So we use symbolic systems every day in classical web search.
Almost all the world’s software is powered by symbolic systems.
We use them for GPS routing.
Neural networks, we use them for speech recognition.
we use them in large language models like ChatGPT,
we use them in image synthesis.
So they're both doing extremely well in the world.
They're both very productive,
but they have their own unique strengths and weaknesses.
So symbolic systems are really good at representing facts
and they're pretty good at reasoning,
but they're very hard to scale.
So people have to custom-build them for a particular task.
On the other hand, neural networks don't require so much custom engineering,
so we can use them more broadly.
But as we've seen, they can't really handle the truth.
I recently discovered that two of the founders of these two theories,
Marvin Minsky and Frank Rosenblatt,
actually went to the same high school in the 1940s,
and I kind of imagined them being rivals then.
And the strength of that rivalry has persisted all this time.
We're going to have to move past that if we want to get to reliable AI.
To get to truthful systems at scale,
we're going to need to bring together the best of both worlds.
We're going to need the strong emphasis on reasoning and facts,
explicit reasoning that we get from symbolic AI,
and we're going to need the strong emphasis on learning
that we get from the neural networks approach.
Only then are we going to be able to get to truthful systems at scale.
Reconciliation between the two is absolutely necessary.
Now, I don't actually know how to do that.
It's kind of like the 64-trillion-dollar question.
But I do know that it's possible.
And the reason I know that is because before I was in AI,
I was a cognitive scientist, a cognitive neuroscientist.
And if you look at the human mind, we're basically doing this.
So some of you may know Daniel Kahneman's System 1
and System 2 distinction.
System 1 is basically like large language models.
It's probabilistic intuition from a lot of statistics.
And System 2 is basically deliberate reasoning.

That's like the symbolic system.
So if the brain can put this together,
someday we will figure out how to do that for artificial intelligence.
There is, however, a problem of incentives.
The incentives to build advertising
hasn't required that we have the precision of symbols.
The incentives to get to AI that we can actually trust
will require that we bring symbols back into the fold.
But the reality is that the incentives to make AI that we can trust,
that is good for society, good for individual human beings,
may not be the ones that drive corporations.
And so I think we need to think about governance.
In other times in history when we have faced uncertainty
and powerful new things that may be both good and bad, that are dual use,
we have made new organizations,
as we have, for example, around nuclear power.
We need to come together to build a global organization,
something like an international agency for AI that is global,
non profit and neutral.
There are so many questions there that I can't answer.
We need many people at the table,
many stakeholders from around the world.
But I'd like to emphasize one thing about such an organization.
I think it is critical that we have both governance and research as part of it.
So on the governance side, there are lots of questions.
For example, in pharma,
we know that you start with phase I trials and phase II trials,
and then you go to phase III.
You don't roll out everything all at once on the first day.
You don't roll something out to 100 million customers.
We are seeing that with large language models.
Maybe you should be required to make a safety case,
say what are the costs and what are the benefits?
There are a lot of questions like that to consider on the governance side.
On the research side, we're lacking some really fundamental tools right now.
For example,
we all know that misinformation might be a problem now,
but we don't actually have a measurement of how much misinformation is out there.
And more importantly,
we don't have a measure of how fast that problem is growing,
and we don't know how much large language models are contributing to the problem.
So we need research to build new tools to face the new risks
that we are threatened by.
It's a very big ask,
but I'm pretty confident that we can get there
because I think we actually have global support for this.
There was a new survey just released yesterday,
said that 91 percent of people agree that we should carefully manage AI.
So let's make that happen.
Our future depends on it.
Thank you very much.

How does artificial intelligence learn?

Briana Brownell

Today, artificial intelligence helps doctors diagnose patients,
pilots fly commercial aircraft, and city planners predict traffic.
But no matter what these AIs are doing, the computer scientists who designed them
likely don't know exactly how they're doing it.
This is because artificial intelligence is often self-taught,
working off a simple set of instructions
to create a unique array of rules and strategies.
So how exactly does a machine learn?
There are many different ways to build self-teaching programs.
But they all rely on the three basic types of machine learning:
unsupervised learning, supervised learning, and reinforcement learning.
To see these in action,
let's imagine researchers are trying to pull information
from a set of medical data containing thousands of patient profiles.
First up, unsupervised learning.
This approach would be ideal for analyzing all the profiles
to find general similarities and useful patterns.
Maybe certain patients have similar disease presentations,
or perhaps a treatment produces specific sets of side effects.
This broad pattern-seeking approach can be used to identify similarities
between patient profiles and find emerging patterns,
all without human guidance.
But let's imagine doctors are looking for something more specific.
These physicians want to create an algorithm
for diagnosing a particular condition.
They begin by collecting two sets of data—
medical images and test results from both healthy patients
and those diagnosed with the condition.
Then, they input this data into a program
designed to identify features shared by the sick patients
but not the healthy patients.
Based on how frequently it sees certain features,
the program will assign values to those features' diagnostic significance,
generating an algorithm for diagnosing future patients.
However, unlike unsupervised learning,
doctors and computer scientists have an active role in what happens next.
Doctors will make the final diagnosis
and check the accuracy of the algorithm's prediction.
Then computer scientists can use the updated datasets
to adjust the program's parameters and improve its accuracy.
This hands-on approach is called supervised learning.
Now, let's say these doctors want to design another algorithm
to recommend treatment plans.
Since these plans will be implemented in stages,
and they may change depending on each individual's response to treatments,
the doctors decide to use reinforcement learning.
This program uses an iterative approach to gather feedback

about which medications, dosages and treatments are most effective.
Then, it compares that data against each patient’s profile
to create their unique, optimal treatment plan.
As the treatments progress and the program receives more feedback,
it can constantly update the plan for each patient.
None of these three techniques are inherently smarter than any other.
While some require more or less human intervention,
they all have their own strengths and weaknesses
which makes them best suited for certain tasks.
However, by using them together,
researchers can build complex AI systems,
where individual programs can supervise and teach each other.
For example, when our unsupervised learning program
finds groups of patients that are similar,
it could send that data to a connected supervised learning program.
That program could then incorporate this information into its predictions.
Or perhaps dozens of reinforcement learning programs
might simulate potential patient outcomes
to collect feedback about different treatment plans.
There are numerous ways to create these machine-learning systems,
and perhaps the most promising models
are those that mimic the relationship between neurons in the brain.
These artificial neural networks can use millions of connections
to tackle difficult tasks like image recognition, speech recognition,
and even language translation.
However, the more self-directed these models become,
the harder it is for computer scientists
to determine how these self-taught algorithms arrive at their solution.
Researchers are already looking at ways to make machine learning more transparent.
But as AI becomes more involved in our everyday lives,
these enigmatic decisions have increasingly large impacts
on our work, health, and safety.
So as machines continue learning to investigate, negotiate and communicate,
we must also consider how to teach them to teach each other to operate ethically.

Can you solve the rogue AI riddle?

Dan Finkel

A hostile artificial intelligence called NIM has taken over the world's computers.
You're the only person skilled enough to shut it down,
and you'll only have one chance.
You've broken into NIM's secret lab,
and now you're floating in a raft on top of 25 stories of electrified water.
You've rigged up a remote that can lower the water level
by ejecting it from grates in the sides of the room.
If you can lower the water level to 0,
you can hit the manual override,
shut NIM off,
and save the day.
However, the AI knows that you're here, and it can lower the water level, too,
by sucking it through a trapdoor at the bottom of the lab.
If NIM is the one to lower the water level to 0,
you'll be sucked out of the lab,
resulting in a failed mission.
Control over water drainage alternates between you and NIM,
and neither can skip a turn.
Each of you can lower the water level by exactly 1,
3,
or 4 stories at a time.
Whoever gets the level exactly to 0 on their turn
will win this deadly duel.
Note that neither of you can lower the water below 0;
if the water level is at 2,
then the only move is to lower the water level 1 story.
You know that NIM has already computed all possible outcomes of the contest,
and will play in a way that maximizes its chance of success.
You go first.
How can you survive and shut off the artificial intelligence?
Pause here if you want to figure it out for yourself.
Answer in: 3
Answer in: 2
Answer in: 1
You can't leave anything up to chance - NIM will take any advantage it can get.
And you'll need to have a response to any possible move it makes.
The trick here is to start from where you want to end and work backwards.
You want to be the one to lower the water level to 0,
which means you need the water level to be at 1, 3, or 4
when control switches to you.
If the water level were at 2,
your only option would be to lower it 1 story,
which would lead to NIM making the winning move.
If we color code the water levels,
we can see a simple principle at play:
there are "losing" levels like 2,
where no matter what whoever starts their turn there does, they'll lose.
And there are winning levels, where whoever starts their turn there

can either win or leave their opponent with a losing level.
So not only are 1, 3, and 4 winning levels,
but so are 5 and 6,
since you can send your opponent to 2 from there.
What about 7?
From 7, all possible moves would send your opponent to a winning level,
making this another losing level.
And we can continue up the lab in this way.
If you start your turn 1, 3, or 4 levels above a losing level,
then you’re at a winning level.
Otherwise, you’re destined to lose.
You could continue like this all the way to level 25.
But as a shortcut,
you might notice that levels 8 through 11 are colored identically to 1 through 4.
Since a level’s color is determined by the levels 1, 3, and 4 stories below it,
this means that level 12 will be the same color as level 5,
13 will match 6,
14 will match 7, and so on,
In particular, the losing levels will always be multiple of 7,
and two greater than multiples of 7.
Now, from your original starting level of 25,
you have to make sure your opponent starts on a losing level every single turn—
if NIM starts on a winning level even once,
it’s game over for you.
So your only choice on turn 1 is to lower the water level by 4 stories.
No matter what the AI does,
you can continue giving it losing levels
until you reach 0 and trigger the manual override.
And with that, the crisis is averted.
Now, back to a less stressful kind of surfing.

How AI is learning what it means to be human

Walter De Brouwer

So two hundred thousand years ago,
Homo sapiens, our ancestors,
were sitting around smoky fire pits
where hunted meat was cooked.
They used rudimentary language, one-syllable words,
a lot of body language, action sounds.
But fifty thousand years ago,
their language had become complex enough to tell stories.
So we can only speculate what these stories were about,
but we have circumstantial evidence.
Around the same time,
after having sat thousands of years around fire pits,
they got up,
left Africa and colonized the entire planet.
And according to the Human Genome Project,
we are all their descendants.
They had discovered
that stories were ideal to activate communities,
and they didn't need any distribution because they were viral.
Five thousand years ago,
these oral tradition stories were written on clay tablets.
Five hundred years ago,
they were printed on Gutenberg presses,
60 years ago they became digital,
30 years ago they ended up on the internet,
and three years ago, the internet was swallowed by insatiable AIs.
And three weeks ago, the LLMs became multimodal.
And there's a lot more modalities probably coming.
We have no idea
if this acceleration is going to slow down,
it doesn't seem like it.
So we all sort of fastened our seat belts
and excitingly follow the newest things.
Now, from the moment
that we started trying
to teach machines human language,
our idea of human language completely changed.
We used to think that human language was a system of signs and sounds
to convey meaning for the sake of communication.
Now we think that was superficial.
There is a deeper function in language.
Language encodes world knowledge.
Language encodes a model of the world.
When our babies are born, they are blank slates.
We cannot put any information in DNA.
The formatting only accepts genetic data.
So language became our external library.
Two hundred thousand years of experiences and knowledge,

you know, is encoded in our language
for our children to find when they grow up.
And so while they are learning the language,
they are absorbing the wisdom of civilizations.
And you know,
these one-syllable words from two hundred thousand years ago,
they have become multi-syllable
and they have become a microcosm of compressed intelligence.
You know, think about,
when you have to explain to children these beautiful words like mercy,
grace,
evolution,
gravity,
equilibrium.
But language is so much more than just words.
That would have been simple.
Language is the multi-modal
internal representation of an external world,
based on our sensory input
and our spatial temporal experiences
that is necessary for cognitive higher functions
like planning, predicting
and multi-step reasoning.
And that world model,
because that is a world model and it is inside language,
that world model we badly need for AGI,
because otherwise it won't work.
So instead of --
and this is like the big turnaround in theoretical linguistics --
instead of trying to study how we use language,
we are now studying how language uses us
to produce that world model,
so that we can extract it and reproduce it in silicon.
And that's a hard problem.
But in 2017, we made a formidable breakthrough.
And when I say we, I mean a team at Google Brain
that published "Attention is All You Need."
They proposed a new architecture,
Transformers.
Transformers are, let's say,
AI models that are very good at understanding
and generating human-like text
based on ... probabilistic patterns.
OpenAI started scaling it,
put some rails on it and made a nice interface.
So we tried it out, I think it was 2019,
and we were impressed.
It was the first time in 30 years
that we really had an idea of what a world model could be.
It had the beginnings of a world model.
Now why,

and you know, why is it so hard to extract that world model out of language?
Well, first of all, language is bigger than the universe
because it is a discrete infinity.
A discrete infinity means it has 26 letters,
but the combinations of that are infinite.
So you must look at it like a cosmic web of meaning,
where words and sentences are interlinked into several dimensions
across time, emotion and context.
And then there are two confusing variables, ambiguities,
words that have several meanings
and long-term dependencies.
That's the relationship between distant words.
So for 30, 40 years,
we were like early astronomers, you know
looking through a telescope
and trying to gaze at a star
and thereby missing the constellations and,
and the galaxies that gave that star the structure and its meaning.
But now, with Transformers and a parallel processing power,
we can see as much as we can from the universe.
We have the possibility to see that celestial dance between ambiguities
and non-linearities and long-term dependencies,
because we are sitting, you know ...
we have have a front-row seat
in the planetarium of human thought.
So what's next?
Well, AGI is what everyone is talking about.
Many parties, many people have opinions.
Many people even have dates connected to it.
I can give you a linguistic view.
How how a linguist would do it.
First of all, everything that comes out of the system,
you feed back in.
And you do that with some human oversight so that you anonymize, you know,
like, where it needs to be anonymized.
Once you do that,
you create a continuous cycle
whereby the AI and human language are continuously negotiating meaning.
This allows the AI to recursively self-learn.
Now, at some point,
you know, there is a critical mass of human feedback.
You know, it's enough.
And then a tipping point has to arrive.
Or let's say a phase transition has to arrive
where the AI becomes self-sufficient in its learning.
It doesn't need that much human information anymore.
It will still always need it because we are evolving,
but it will also evolve.
So it will become autonomous and self-sufficient.
When it becomes autonomous and self-sufficient,
that AI becomes unavoidable.

You know, in French we have a word, "incontournable."
You cannot really translate it in English,
but you have to be there or you fall behind.
So more people will start to use it.
And at that point, that AI will annex more domains with transfer learning.
And that AI will become so complex,
just like any adaptive complex system,
it will become unpredictable.
And then suddenly we humans will say,
"Hey, it's human"
and we will recognize it, you know?
And we will probably call it alternative general intelligence,
because you see, in the end, it comes to us.
AGI is a philosophical question.
Now why do I think it is possible?
Well ...
We are the only animals
whose language is complex enough
to imagine the future
and to create sophisticated tools to get there.
We're pushers of boundaries, so it will happen.

Can robots be creative?

Gil Weinberg

How does this music make you feel?
Do you find it beautiful?
Is it creative?
Now, would you change your answers
if you learned the composer was this robot?
Believe it or not,
people have been grappling with the question of artificial creativity,
alongside the question of artifcial intelligence,
for over 170 years.
In 1843, Lady Ada Lovelace,
an English mathematician considered the world's first computer programmer,
wrote that a machine could not have human-like intelligence
as long as it only did what humans intentionally programmed it to do.
According to Lovelace,
a machine must be able to create original ideas
if it is to be considered intelligent.
The Lovelace Test, formalized in 2001, proposes a way of scrutinizing this idea.
A machine can pass this test if it can produce an outcome
that its designers cannot explain based on their original code.
The Lovelace Test is, by design, more of a thought experiment
than an objective scientific test.
But it's a place to start.
At first glance,
the idea of a machine creating high quality, original music in this way
might seem impossible.
We could come up with an extremely complex algorithm
using random number generators, chaotic functions, and fuzzy logic
to generate a sequence of musical notes
in a way that would be impossible to track.
But although this would yield countless original melodies never heard before,
only a tiny fraction of them would be worth listening to.
With the computer having no way to distinguish
between those which we would consider beautiful
and those which we won't.
But what if we took a step back
and tried to model a natural process that allows creativity to form?
We happen to know of at least one such process
that has lead to original, valuable, and even beautiful outcomes:
the process of evolution.
And evolutionary algorithms,
or genetic algorithms that mimic biological evolution,
are one promising approach
to making machines generate original and valuable artistic outcomes.
So how can evolution make a machine musically creative?
Well, instead of organisms,
we can start with an initial population of musical phrases,
and a basic algorithm
that mimics reproduction and random mutations

by switching some parts,
combining others,
and replacing random notes.
Now that we have a new generation of phrases,
we can apply selection using an operation called a fitness function.
Just as biological fitness is determined by external environmental pressures,
our fitness function can be determined by an external melody
chosen by human musicians, or music fans,
to represent the ultimate beautiful melody.
The algorithm can then compare between our musical phrases
and that beautiful melody,
and select only the phrases that are most similar to it.
Once the least similar sequences are weeded out,
the algorithm can reapply mutation and recombination to what's left,
select the most similar, or fitted ones, again from the new generation,
and repeat for many generations.
The process that got us there has so much randomness and complexity built in
that the result might pass the Lovelace Test.
More importantly, thanks to the presence of human aesthetic in the process,
we'll theoretically generate melodies we would consider beautiful.
But does this satisfy our intuition for what is truly creative?
Is it enough to make something original and beautiful,
or does creativity require intention and awareness of what is being created?
Perhaps the creativity in this case is really coming from the programmers,
even if they don't understand the process.
What is human creativity, anyways?
Is it something more than a system of interconnected neurons
developed by biological algorithmic processes
and the random experiences that shape our lives?
Order and chaos, machine and human.
These are the dynamos at the heart of machine creativity initiatives
that are currently making music, sculptures, paintings, poetry and more.
The jury may still be out
as to whether it's fair to call these acts of creation creative.
But if a piece of art can make you weep,
or blow your mind,
or send shivers down your spine,
does it really matter who or what created it?

Leadership in the age of AI
Paul Hudson and Lindsay Levin
Lindsay Levin: So we're living in an era with multiple overlapping disruptions
that business is facing, and I want to dive straight in
and talk about one of the biggest of those, which is AI.
How are you approaching AI?
Paul Hudson: You know, AI at scale, it's a whole big subject, of course,
but for us, at Sanofi,
we aim to be the world's leading pharmaceutical company using AI at scale.
Why and how are we going to do that?
It's pretty straightforward.
We have 23,000 people using AI as often as every month,
9,000 people in the company using AI as often as every day.
We're boldly taking on the opportunity to completely disrupt the business.
We don't have a choice.
It's the fourth industrial revolution.
It's here whether we like it or not.
And it's amazing how resistant people can be across organizations
and across industries.
But we're all in and have been quite public about that.
Our aim is to provide daily decision intelligence,
to give people a sort of nudge in the right direction,
to give them deeper insights,
to allow them to be more effective at what they do.
And it's real.
And it's such a privilege to be involved in it.
LL: I mean, you're taking a very aggressive, active stance.
What surprised you?
PH: Well, a lot of things surprise you about AI.
I mean, for some people it's Skynet and Terminator.
For some people, they confuse AI with cyber.
I'm not saying everything is perfect,
but I'm surprised at the number of CEOs or executives who --
Their first response to an AI conversation is
"Governance, controls, rules, principles."
Of course, everything has its place,
but I think we have to be honest with ourselves.
If it is the fourth industrial revolution, which we believe it is,
then hesitating,
this fear that can take over,
can deprive you of so much opportunity.
And you have to go for it.
I find that when you talk to lots of CEOs, they really are very hesitant.
Some would say even frightened.
I look internally,
people are frightened that you get this radical data transparency
You can see their data real-time.
LL: And you're experiencing that.
PH: I've experienced that and still do.
You know, people are often shocked
that you may get the insight at the same time

as somebody lower down the organization.
And then there's the lost opportunity to polish a slide deck
and re-present it in the way that I'm supposed to be informed.
It's not a deliberate, sort of, misleading approach.
It's what people know.
They get the insight, they craft the story,
they push it upwards.
And that's life in many corporations.
For us, we get the data,
I get the same data every level of the organization does.
I get the insight exactly the same time.
And then people say, "Paul, don't look, the data is not 100 percent correct."
Well, make it correct because the data is live.
So if you really jump in and make it correct,
it'll better reflect what you're doing, right?
But if we wait for perfection it's simply not going to happen.
LL: So we're seeing fear and some of that, I guess, is not unreasonable.
You know, we read reports about the impact on job losses,
for example,
to come from AI.
I wonder what mindset you believe people need to adopt
in the workforce of all generations,
as they approach or we all approach this new future?
PH: You know, the adoption of AI in particular is not about jobs.
And I know people will think that
and inevitably, more meaningful work is created.
And of course, some roles change or some skills don't match.
And therefore, you know, with the help of many of the people in the room,
you get to reshape organizations.
But in the end,
it's really about using artificial intelligence
to create this real momentum of decision making
and to be able to take such an advantage over the competition.
And we believe, I believe,
that if you create more meaningful work
and people focus on insights and action
and less on Excel and PowerPoint and Word,
then there is a chance that they enjoy their work more.
Now it may lead to productivity gains, it may lead to efficiency.
It may lead to all those things, nobody's sort of denying that.
But what I've discovered so far is when it does,
people see it quite quickly
and they put their hand up to do something else,
or to focus more on insights
than ... data crunching
and aggregation.
You know, I'm old enough to remember when the internet was launched,
you know, commercially.
And it's sort of similar arguments, even when the cell phone was launched.
"Be careful."
But the truth is, they made all our lives better.

The use cases are coming.
But I think we're starting to understand now
how much this is going to change everybody's lives.
LL: So who is leading this in the organization?
Is it a new generation?
Is it specially appointed people?
Where's the leadership coming from?
PH: It's an excellent question
because maybe I'm the last of the great dinosaur CEOs
who got to the top by doing sort of every job.
I ran Japan, I ran North America,
you know, I was in global marketing.
You know, I've done all the tasks, to get to the top,
and then I've sort of seen everything.
And so I can be involved in every discussion.
And now the younger talent are saying,
"You didn't see AI, old man."
So, you know, "I have a better insight than you do."
And "Oh, and by the way, I'm not just going to push it up to my boss,
I'd like to tell you myself."
LL: Right.
PH: So the younger talent, justifiably is saying,
"Hey," you know, "I don't need to have my work shared upwards
by a bunch of guys
who are all sitting there going, 'What do you think?'
and none of them actually know."
And so we invert the pyramid.
We have to have different people with two, three,
four years experience in the room.
Because what do we know?
And that's sort of exciting, I think, really exciting.
LL: So is AI a unique disruptor?
If we think about some of the other big challenges, you know,
we've got to shift the entire global economy to be sustainable, for example.
Is that comparable in terms of complexity?
Maybe more so?
How do you tackle that kind of an issue?
PH: I think these are the big transformational moments for society.
And, you know, sustainability is,
you know, for many, it was carbon neutrality,
then it was net zero,
it was go to COP 28,
it was put a poster by the elevator
with the meaningful work you're doing to show your commitment.
But it's really different now.
I think there's a collective realization, certainly in healthcare,
that we didn't do well enough.
And we're a very purpose-driven organization --
an industry, in fact. We do health,
we transform the practice of medicine, we invent miracles often.
And so it's very easy to say, look, we're very purpose-driven.

But it doesn't abdicate the responsibility
of removing plastic from packaging of vaccines and medicines.
It's ridiculous to even think you wouldn't have to.
Often it's harder with the regulator, by the way, to get that done
than with your own people.
LL: So you've got those kinds of projects going on, have you?
Can you give us an example?
PH: We have to do it.
We have to do it because, you know, we have this sort of approach
of what can we do
that if we don't do, it won't happen?
And that's sort of our philosophy.
You know, in healthcare, it's a good example,
you know, delivery of healthcare creates more carbon
than the airline industry.
And that's half of that, let's say five percent.
Half of that is making drugs, shipping them, doing different things.
The other half is people driving to hospitals for an appointment
in an overheated, overcooled healthcare practice, too often,
without the use of a virtual hybrid delivery of healthcare.
And it's the same as the airline industry.
And, you know, we feel, because we're in healthcare,
we have this unique opportunity.
If somebody is pre-diabetic, for example,
and you coach them and they change their lifestyle
and don't become a diabetic,
that's a difference of them creating 16 tons of carbon
as a normal adult, healthy,
and 48 tons of carbon in their adult life if they become diabetic.
That's a 3X.
That's really meaningful.
And if we don't step in and help,
we just simply never get there and we're doing a lot of work,
I'm doing a lot of work with King Charles
and the Sustainable Markets Initiative
to get people to decarbonize the delivery of healthcare,
because it's such a massive opportunity.
LL: And presumably you've got to collaborate in very different ways
than in the past to do, you know,
you're talking about supply chains to deal with something like plastic.
Are you seeing different kinds of skills from people
in order to make those collaborations?
PH: Well, I think these functions and the sustainability groups, as I said,
have come from a poster by the elevator
to being very actively involved in a lot of work to do these things properly.
And, you know, you, it's not about a competitive advantage,
in Sanofi being better at wastewater management
or renewable energy than Pfizer or AstraZeneca.
That's not a competition.
The competition is us versus, you know, destroying the planet.
So we work a lot together to do the right thing.

I work with Novo Nordisk, AstraZeneca, GSK
to try and work out ways to be kinder to the environment
in the delivery of health, and it's the right thing to do, right?
It's a shared responsibility, a collective responsibility.
LL: And we're talking here about big social challenges
beyond any one business
or industry or even country,
with an expectation that business needs to step forward.
I think partly driven by the fact that policy doesn't always work,
and we're disappointed with leadership and with government.
So a lot of finger pointing as to who's responsible
and who can act on these big shifts.
I wonder how ambitious and bold
do you feel CEOs should be about stepping forward
and helping society through some of these mega transitions
that we are faced with?
PH: You know, it's clear that companies are being pulled more
into the conversation about individual's values.
And I think people who work in our company,
and all companies,
start to try and identify themselves, perhaps rightly,
with the values of the company.
And they're starting to have much higher expectations
about the company they work for.
And it can be on all the major social issues.
You know, there's been so many difficult moments and tragic moments.
I'm often written to by people from all over the world.
"You haven't declared which side you're on on this important subject.
Why not?"
And people want to know that you are fully vested
and the company is behind them.
To be clear, it's almost impossible to get everything right.
The world is almost, you know, perfectly divided.
You can pick an issue and half your employees will tell you
"We don't agree," and the other half will say, "Well done."
And we're not used to that as CEOs,
we're used to trying to find the right sort of moment
to get the majority to say, "I'm proud of my company."
So you have to retreat a little bit and say,
I'll spend my energy on making sure whatever the issue, that the people,
90,000 people in our case,
get the very best chance to live their best version of themselves.
Could be inclusion, it could be race,
it could be many different things.
It could be LGBTQ-plus.
But whatever those issues that are being debated strongly
or less strongly in different parts of the world,
it matters to us that our people feel
they can be the best version of themselves.
So when people ask, "What do you think about this?"
"What do you think about that?"

I can have a personal opinion, but our organization,
if you're looking to match your values with ours,
is really about, we're in healthcare,
we want you to have the best life possible.
How can the company facilitate that in this sort of, maelstrom of subjects?
And we focus then -- I'm not saying we're perfect,
I don't think any company really is perfect on this,
but the expectation of our own employees to be able to set a standard on an issue
and to see it through is real.
And leading in these times is more complicated, I think,
than perhaps it has been previously.
LL: So you're describing a world
where the work of the CEO is changing very fast.
Could you just sum up for us the role of a leader in this new era?
PH: Well, I took this role at the end of 2019 and thought,
I will roll through my 100-day plan
and I will amble through getting around town halls across the world
and shake a lot of hands and do a listening tour.
And a pandemic dropped on us within a few weeks after that
and worked out of the kitchen not far from here.
And the war, Russia-Ukraine,
current war, Israel and Hamas,
and mentioned the pandemic.
China-US, the social issues.
And I think what we realize is leading
is I think somebody described it as the perma crisis, you know,
a sort of, you know, a cadence of issues
that just is relentless.
And you really have to have some resiliency leading now, I think,
and you have to recognize that there are a series of sprints in the role.
There's the fundamentals of the business that must be continued.
Then there's a metronome for us.
We feel a responsibility to bring medicines
and transformational medicines forward is non-negotiable.
At the same time,
parts of your organization is in a very difficult situation
somewhere in the world.
And we have to make sure we have the right energy, experts,
support, crisis teams, often, more recently
to protect our people
and to continue the work we do.
We ship drugs all over the world, irrespective of the stance of politics
or anything else.
People are people and if they're sick we'll help.
But it really has got very complicated.
So resiliency, agility and being open-minded,
recognizing you're not the expert many times,
trusting the advice you get from your own people,
particularly those on the ground,
protecting your people where it's necessary,
moving your people where it's necessary.

That's the sort of new expectation, really.
LL: I mean, I think we're all experiencing this sense of perma crisis.
Just to finish,
it would be great to just get a sense
of what are you really excited about right now?
PH: Well, I'm incredibly excited,
I touched on it at the beginning about the use of artificial intelligence,
particularly large language models.
Because I think it changes everything.
It's got me questioning whether I can go back
and look at medicines that didn't quite make it,
and wonder if we just didn't know enough
with the data that we had to look deep enough.
It can be, you know, recently we just did our --
this is a small example, but it's just fun,
our engagement survey,
we had a 409,000 comments,
9 million words.
And normally, somebody would make a nice slide deck
to tell me, "The organization poll is very engaged.
Never been more engaged."
"How much more engaged?" "0.1."
"OK, good, thank you."
Definitely improvement.
And so I asked them to run the 409,000 comments
through a large language model.
Forty minutes later, it told me the three things
that people love about the company
and three things that people hate about it.
I didn't need a lot of external support, didn't need teams of people.
And it was clear, it was no hallucination,
because it was there right in front of me.
And it made sense.
And I shared them with people and they're like, yeah, that's us.
And I think that's the difference between meaningful work
using -- let's talk about what it tells us
about whether our people really like it here
and bring their best or not.
I think AI for me, a relatively new CEO,
I have a chance to disrupt structural biology, electron microscopy,
I have a chance to invent medicines
and druggable targets that were never touched before.
I have a chance to take away the sort of,
I should put it, the heavy lifting,
unglamorous work that people don't like doing.
I've a chance to reinvent everything,
to do it more efficiently, reinvest in R and D.
And I have a chance to get ahead of the competition
while they're all worrying,
we have governance,
but every step forward by us

is a step of leaving behind those that are overly sensitive,
and we're happy to share.
But we can do incredible things for patients
and for the people in the company
by being more bold about stepping into the new world.
LL: Your passion is infectious.
Thank you.
Thank you very much, Paul.

What will happen to marketing in the age of AI?

Jessica Apotheker

So let me start by bringing you back in time.
We are 30 years ago,
and the first word processors and spreadsheets
are about to hit the market.
And the whole economic world is bracing for the next big productivity revolution.
Their promise at the time was we'd all spend so much less time writing,
drawing slides, computing numbers on a calculator.
And here we are, 30 years later,
and the promise has come true.
We all have so much leisure time on our hands,
and personally, I only work two days a week.
Of course, I'm just kidding.
The reality of what has happened 30 years later
is we don't work less.
We just write much longer word documents.
And our PowerPoint decks have gone from six slides to 50 slides.
And I say that as a consultant.
Also, we engage in much more complex decision-making
because the amount of data that we have to process has just exploded.
And why is that important today?
Well, generative AI is coming,
and it's coming to be embedded in the core of our organizations
and the way we work.
And that will be the next big productivity revolution.
So the question becomes:
how do we set ourselves up to actually seize this productivity opportunity?
I'm a marketer.
I spent all my career in marketing and also advising marketing professionals.
Now, some say marketing is the number one impacted function out there.
Some say the productivity impact in marketing
is as high as 50 percent.
So that question of how can I seize that productivity opportunity
is super high on my mind right now,
and I want to make the case
it should be super important to you all as well,
as business leaders but also as consumers.
So what will happen to marketing?
Well, marketing has traditionally been a super right-brained,
creative type of function.
That means what?
Means we have excelled as marketers
by tapping into the emotional needs of our consumers,
coming up with that perfect product,
that perfect innovation to meet that need,
and also then cracking that great message
that will convert the consumer at the right place in the right time.
Already in the past 15 years, with digital marketing and analytics,
marketing has evolved from being only right-brain type of general skills

to a few more specialized skill sets,
for example, digital marketing or marketing technology.
But now the difference with generative AI,
it is transforming the core of marketing activities.
Now, in a recent study
that the Boston Consulting Group conducted with Harvard,
we found that ChatGPT, in its current form,
already improves the right-brain performance of marketers by 40 percent.
Imagine what that number will be in a year or two from now.
So what do you think marketers would do
with a day and a half of free time a week?
More yoga?
More family time?
Do you think companies would allow that?
Or do you think companies will just let large chunks of the marketing function go?
Well, I believe none of this is going to happen.
I think if we don't steer
that productivity revolution very actively,
marketers will invest this time in what they do best:
more content and more ideas.
Now, if you think of more content,
there is a super productive outcome for all of us as consumers.
More content actually means much more personalized content.
Now think of that email
that you're getting from your favorite brand every week.
Imagine if that email was 100 percent tailored to you,
means only images of people your age and gender,
even people wearing T-shirts of your favorite rock band,
every product relevant for you,
and even a human-like experience powered by a bot.
That is certainly a productive outcome.
But there is also a very negative outcome for us consumers here,
and that is content overload.
How many of you already feel chased
by the same content over and over again online?
Now imagine if that content chasing you,
if that amount of content chasing you just explodes.
And imagine if that content chasing you also all sounds the same.
Now why is that a risk?
Generative AI has been trained on existing content and data.
Because of that, it reduces divergence of outcomes.
And that great equalization of marketing is certainly not a productive outcome.
So what is the solve here?
Well, I believe marketing,
but also every function out there
that is being impacted by this productivity revolution,
needs to grow a left-AI brain, grow one fast,
and also identify and protect its top right-brained talent.
You're going to ask me,
"What do you mean by growing a left-AI brain?"
Well, I mean, the function needs to strategically reskill and reorganize

to embed people that can build,
use and diffuse predictive AI tools in the heart of decision-making.
I mean, for marketing, building teams of marketing data scientists,
marketing data engineers that build solutions
that can be distributed to all marketers
to, for example, unpack performance and predict outcomes.
Imagine in marketing
being able to understand what audience creative couples
are really hitting it off in the market,
or what product is working with which consumer and why
or how is the marketing funnel evolving.
I recently partnered with a consumer goods company that did exactly that.
They decided to grow a left-AI brain advantage.
We helped them build tools
that were diffused in the entire organization,
that helped every marketer predict for every marketing initiative
what was going to be the sales outcome,
how a consumer behavior is going to be impacted on every channel
and every touchpoint,
and go deep in unpacking execution insights
to understand what creative was working and why.
That created a super virtuous feedback loop in the entire organization.
It also took building a team of 30-plus left-AI brain marketers
that build these tools, customize them,
but also in turn upskill the entire organization to use them.
But the team's only a part of the puzzle.
I see too many companies out there embarking on this journey,
just training their algorithms and models only on their current content and data.
Now, if you do that,
the risk for a brand is to be trapped in your current territory.
Concretely, imagine you are a brand that is super strong with millennials.
There is nothing in data and content existing on millennials
that will help you to be successful with Gen Z.
And in turn, if you're never successful with Gen Z,
you will miss out on important innovations and trends
that will make you stronger with millennials.
So I advise every company out there:
think outside of the box,
think outside your direct ecosystem
on who could be super relevant data and content partners for you.
Imagine you're a construction company
and you decide to market to architects for the first time.
You have zero data on architects.
What do you do?
Who has data on architects?
Other construction companies, but they're direct competitors.
So where do you go?
Well, you go outside your ecosystem,
potentially, for example, with financial institutions, insurances.
You can set up a federated model with them,
train algos on that,

that will in turn make you much stronger
to market to a new consumer segment.
And so are you done?
If you have that, if you have that data, if you have those skills
are you done, you have that left-AI brain advantage?
Well, no, actually you are not.
If you do that, there is a risk you give all of your right brain
to generative AI
and in turn run a real risk of losing that divergence,
losing that super strong brand identity,
being trapped in that grand equalization of marketing
I was talking about a minute ago.
In the Harvard study we conducted
with the Boston Consulting Group and Harvard,
we found that when people over-rely on generative AI,
the collective divergence of ideas drops by 40 percent.
Concretely, that means that new ideas don't come to the surface.
It means that true innovation is being stifled.
So what is a solve here?
Well, you need to identify the true artists,
the true differentiators,
the true innovators of your function.
Now, if you've ever worked in marketing, you know who these people are.
They are the ones that always disagree with you.
Now you take these people
and you need to strategically reskill them to use AI well,
for example, to be inspired by new ideas,
to be inspired by new trends,
to also crack fast prototypes,
to multiply their impact once they've cracked a great idea.
But you need to protect them and teach them
from using the AI to generate and originate original ideas.
For that, they have to use their human brain.
To keep those human juices flowing,
and that, in turn, will protect the identity of your brand
and your differentiation in the market.
So I want to close with an advice for any marketer out there.
What are you good at?
Are you super creative?
Are you the true innovator in the room?
Well, if you are, cultivate that.
That will be your superpower.
Do you like data?
Are you super rational, are you super fact-based?
Then you should specialize.
You should grow tech skills.
You should be investing in predictive AI competencies.
But right now, every marketer out there needs to choose their brain.

The 100,000-student classroom

Peter Norvig

Everyone is both a learner
and a teacher.
This is me being inspired
by my first tutor,
my mom,
and this is me teaching
Introduction to Artificial Intelligence
to 200 students
at Stanford University.
Now the students and I
enjoyed the class,
but it occurred to me
that while the subject matter
of the class is advanced
and modern,
the teaching technology isn't.
In fact, I use basically
the same technology as
this 14th-century classroom.
Note the textbook,
the sage on the stage,
and the sleeping guy
in the back.
Just like today.
So my co-teacher,
Sebastian Thrun, and I thought,
there must be a better way.
We challenged ourselves
to create an online class
that would be equal or better
in quality to our Stanford class,
but to bring it to anyone
in the world for free.
We announced the class on July 29th,
and within two weeks, 50,000 people
had signed up for it.
And that grew to 160,000 students
from 209 countries.
We were thrilled to have
that kind of audience,
and just a bit terrified that we
hadn't finished preparing the class yet.
So we got to work.
We studied what others had done,
what we could copy and what we could change.
Benjamin Bloom had showed
that one-on-one tutoring works best,
so that's what we tried to emulate,

like with me and my mom,
even though we knew
it would be one-on-thousands.
Here, an overhead video camera
is recording me as I'm talking
and drawing on a piece of paper.
A student said, "This class felt
like sitting in a bar
with a really smart friend
who's explaining something
you haven't grasped, but are about to."
And that's exactly what we were aiming for.
Now, from Khan Academy, we saw
that short 10-minute videos
worked much better than trying
to record an hour-long lecture
and put it on the small-format screen.
We decided to go even shorter
and more interactive.
Our typical video is two minutes,
sometimes shorter, never more
than six, and then we pause for
a quiz question, to make it
feel like one-on-one tutoring.
Here, I'm explaining how a computer uses
the grammar of English
to parse sentences, and here,
there's a pause and the student
has to reflect, understand what's going on
and check the right boxes
before they can continue.
Students learn best when
they're actively practicing.
We wanted to engage them, to have them grapple
with ambiguity and guide them to synthesize
the key ideas themselves.
We mostly avoid questions
like, "Here's a formula, now
tell me the value of Y
when X is equal to two."
We preferred open-ended questions.
One student wrote, "Now I'm seeing
Bayes networks and examples of
game theory everywhere I look."
And I like that kind of response.
That's just what we were going for.
We didn't want students to memorize the formulas;
we wanted to change the way
they looked at the world.
And we succeeded.
Or, I should say, the students succeeded.

And it's a little bit ironic
that we set about to disrupt traditional education,
and in doing so, we ended up
making our online class
much more like a traditional college class
than other online classes.
Most online classes, the videos are always available.
You can watch them any time you want.
But if you can do it any time,
that means you can do it tomorrow,
and if you can do it tomorrow,
well, you may not ever
get around to it.
So we brought back the innovation
of having due dates.
You could watch the videos
any time you wanted during the week,
but at the end of the week,
you had to get the homework done.
This motivated the students to keep going, and it also
meant that everybody was working
on the same thing at the same time,
so if you went into a discussion forum,
you could get an answer from a peer within minutes.
Now, I'll show you some of the forums, most of which
were self-organized by the students themselves.
From Daphne Koller and Andrew Ng, we learned
the concept of "flipping" the classroom.
Students watched the videos
on their own, and then they
come together to discuss them.
From Eric Mazur, I learned about peer instruction,
that peers can be the best teachers,
because they're the ones
that remember what it's like to not understand.
Sebastian and I have forgotten some of that.
Of course, we couldn't have
a classroom discussion with
tens of thousands of students,
so we encouraged and nurtured these online forums.
And finally, from Teach For America,
I learned that a class is not
primarily about information.
More important is motivation and determination.
It was crucial that the students see
that we're working hard for them and
they're all supporting each other.
Now, the class ran 10 weeks,
and in the end, about half of the 160,000 students watched
at least one video each week,
and over 20,000 finished all the homework,

putting in 50 to 100 hours.
They got this statement of accomplishment.
So what have we learned?
Well, we tried some old ideas
and some new and put them together,
but there are more ideas to try.
Sebastian's teaching another class now.
I'll do one in the fall.
Stanford Coursera, Udacity, MITx
and others have more classes coming.
It's a really exciting time.
But to me, the most exciting
part of it is the data that we're gathering.
We're gathering thousands
of interactions per student per class,
billions of interactions altogether,
and now we can start analyzing that,
and when we learn from that,
do experimentations,
that's when the real revolution will come.
And you'll be able to see the results from
a new generation of amazing students.

4 ways to build a human company in the age of machines

Tim Leberecht

Half of the human workforce is expected to be replaced
by software and robots in the next 20 years.
And many corporate leaders welcome that as a chance to increase profits.
Machines are more efficient;
humans are complicated and difficult to manage.
Well, I want our organizations to remain human.
In fact, I want them to become beautiful.
Because as machines take our jobs and do them more efficiently,
soon the only work left for us humans will be the kind of work
that must be done beautifully rather than efficiently.
To maintain our humanity in the this second Machine Age,
we may have no other choice than to create beauty.
Beauty is an elusive concept.
For the writer Stendhal it was the promise of happiness.
For me it's a goal by Lionel Messi.

So bear with me
as I am proposing four admittedly very subjective principles
that you can use to build a beautiful organization.
First: do the unnecessary.
[Do the Unnecessary]
A few months ago, Hamdi Ulukaya,
the CEO and founder of the yogurt company Chobani,
made headlines when he decided to grant stock to all of his 2,000 employees.
Some called it a PR stunt,
others -- a genuine act of giving back.
But there is something else that was remarkable about it.
It came completely out of the blue.
There had been no market or stakeholder pressure,
and employees were so surprised
that they burst into tears when they heard the news.
Actions like Ulukaya's are beautiful because they catch us off guard.
They create something out of nothing
because they're completely unnecessary.
I once worked at a company
that was the result of a merger
of a large IT outsourcing firm and a small design firm.
We were merging 9,000 software engineers
with 1,000 creative types.
And to unify these immensely different cultures,
we were going to launch a third, new brand.
And the new brand color was going to be orange.
And as we were going through the budget for the rollouts,
we decided last minute
to cut the purchase of 10,000 orange balloons,
which we had meant to distribute to all staff worldwide.
They just seemed unnecessary and cute in the end.
I didn't know back then

that our decision marked the beginning of the end --
that these two organizations would never become one.
And sure enough, the merger eventually failed.
Now, was it because there weren't any orange balloons?
No, of course not.
But the kill-the-orange-balloons mentality permeated everything else.
You might not always realize it, but when you cut the unnecessary,
you cut everything.
Leading with beauty means rising above what is merely necessary.
So do not kill your orange balloons.
The second principle:
create intimacy.
[Create Intimacy]
Studies show that how we feel about our workplace
very much depends on the relationships with our coworkers.
And what are relationships other than a string of microinteractions?
There are hundreds of these every day in our organizations
that have the potential to distinguish a good life from a beautiful one.
The marriage researcher John Gottman says
that the secret of a healthy relationship
is not the great gesture or the lofty promise,
it's small moments of attachment.
In other words, intimacy.
In our networked organizations,
we tout the strength of weak ties
but we underestimate the strength of strong ones.
We forget the words of the writer Richard Bach who once said,
"Intimacy --
not connectedness --
intimacy is the opposite of loneliness."
So how do we design for organizational intimacy?
The humanitarian organization CARE
wanted to launch a campaign on gender equality
in villages in northern India.
But it realized quickly
that it had to have this conversation first with its own staff.
So it invited all 36 team members and their partners
to one of the Khajuraho Temples,
known for their famous erotic sculptures.
And there they openly discussed their personal relationships --
their own experiences of gender equality
with the coworkers and the partners.
It was eye-opening for the participants.
Not only did it allow them to relate to the communities they serve,
it also broke down invisible barriers
and created a lasting bond amongst themselves.
Not a single team member quit in the next four years.
So this is how you create intimacy.
No masks ...
or lots of masks.

When Danone, the food company,
wanted to translate its new company manifesto into product initiatives,
it gathered the management team
and 100 employees from across different departments,
seniority levels and regions
for a three-day strategy retreat.
And it asked everybody to wear costumes for the entire meeting:
wigs, crazy hats, feather boas,
huge glasses and so on.
And they left with concrete outcomes
and full of enthusiasm.
And when I asked the woman who had designed this experience
why it worked,
she simply said, "Never underestimate the power of a ridiculous wig."

Because wigs erase hierarchy,
and hierarchy kills intimacy --
both ways,
for the CEO and the intern.
Wigs allow us to use the disguise of the false
to show something true about ourselves.
And that's not easy in our everyday work lives,
because the relationship with our organizations
is often like that of a married couple that has grown apart,
suffered betrayals and disappointments,
and is now desperate to be beautiful for one another once again.
And for either of us the first step towards beauty involves a huge risk.
The risk to be ugly.
[Be Ugly]
So many organizations these days are keen on designing beautiful workplaces
that look like anything but work:
vacation resorts, coffee shops, playgrounds or college campuses --

Based on the promises of positive psychology,
we speak of play and gamification,
and one start-up even says that when someone gets fired,
they have graduated.

That kind of beautiful language only goes "skin deep,
but ugly cuts clean to the bone,"
as the writer Dorothy Parker once put it.
To be authentic is to be ugly.
It doesn't mean that you can't have fun or must give in to the vulgar or cynical,
but it does mean that you speak the actual ugly truth.
Like this manufacturer
that wanted to transform one of its struggling business units.
It identified, named and pinned on large boards all the issues --
and there were hundreds of them --
that had become obstacles to better performance.
They put them on boards, moved them all into one room,

which they called "the ugly room."
The ugly became visible for everyone to see --
it was celebrated.
And the ugly room served as a mix of mirror exhibition and operating room --
a biopsy on the living flesh to cut out all the bureaucracy.
The ugliest part of our body is our brain.
Literally and neurologically.
Our brain renders ugly what is unfamiliar ...
modern art, atonal music,
jazz, maybe --
VR goggles for that matter --
strange objects, sounds and people.
But we've all been ugly once.
We were a weird-looking baby,
a new kid on the block, a foreigner.
And we will be ugly again when we don't belong.
The Center for Political Beauty,
an activist collective in Berlin,
recently staged an extreme artistic intervention.
With the permission of relatives,
it exhumed the corpses of refugees who had drowned at Europe's borders,
transported them all the way to Berlin,
and then reburied them at the heart of the German capital.
The idea was to allow them to reach their desired destination,
if only after their death.
Such acts of beautification may not be pretty,
but they are much needed.
Because things tend to get ugly when there's only one meaning, one truth,
only answers and no questions.
Beautiful organizations keep asking questions.
They remain incomplete,
which is the fourth and the last of the principles.
[Remain Incomplete]
Recently I was in Paris,
and a friend of mine took me to Nuit Debout,
which stands for "up all night,"
the self-organized protest movement
that had formed in response to the proposed labor laws in France.
Every night, hundreds gathered at the Place de la République.
Every night they set up a small, temporary village
to deliberate their own vision of the French Republic.
And at the core of this adhocracy
was a general assembly where anybody could speak
using a specially designed sign language.
Like Occupy Wall Street and other protest movements,
Nuit Debout was born in the face of crisis.
It was messy --
full of controversies and contradictions.
But whether you agreed with the movement's goals or not,
every gathering was a beautiful lesson in raw humanity.
And how fitting that Paris --

the city of ideals, the city of beauty --
was it's stage.
It reminds us that like great cities,
the most beautiful organizations are ideas worth fighting for --
even and especially when their outcome is uncertain.
They are movements;
they are always imperfect, never fully organized,
so they avoid ever becoming banal.
They have something but we don't know what it is.
They remain mysterious; we can't take our eyes off them.
We find them beautiful.
So to do the unnecessary,
to create intimacy,
to be ugly,
to remain incomplete --
these are not only the qualities of beautiful organizations,
these are inherently human characteristics.
And these are also the qualities of what we call home.
And as we disrupt, and are disrupted,
the least we can do is to ensure
that we still feel at home in our organizations,
and that we use our organizations to create that feeling for others.
Beauty can save the world when we embrace these principles
and design for them.
In the face of artificial intelligence and machine learning,
we need a new radical humanism.
We must acquire and promote a new aesthetic and sentimental education.
Because if we don't,
we might end up feeling like aliens
in organizations and societies that are full of smart machines
that have no appreciation whatsoever
for the unnecessary,
the intimate,
the incomplete
and definitely not for the ugly.
Thank you.

War, AI and the new global arms race

Alexandr Wang

Artificial intelligence and warfare.
Let's talk about what this really could look like.
Swarms of lethal drones with facial recognition
that know your every move.
Or unmanned armed robots that are near impossible to defeat.
Autonomous fighter jets that can travel at supersonic speeds
and can withstand greater gravitational force
than a human pilot could survive.
Cyberattacks that incapacitate critical port infrastructure
or disinformation campaigns and deepfakes that throw presidential elections.
Or even foreign adversaries taking out satellites,
our eyes and ears in space,
rendering us blind to global events.
All superintelligent weapons of terror.
We are at the dawn of a new age of warfare.
I grew up in the birthplace of a technology
that defined the last era of warfare,
the atomic bomb.
I was keenly aware of how this technology had fundamentally shaped geopolitics
and the nature of war.
My parents were both scientists at Los Alamos National Laboratory.
My dad’s a physicist, and my mom’s an astrophysicist.
Their scientific work in plasma fluid dynamics
will have deep implications for how we understand our universe.
So naturally, I knew I wanted to work on something just as impactful.
I decided to become a programmer and study artificial intelligence.
AI is one of the most critical technologies of our time
and with deep implications for national security
and democracy globally.
As we saw in World War II with the atomic bomb,
the country that is able to most rapidly and effectively
integrate new technology into warfighting wins.
There's no reason to believe this will be any different with AI.
But in the AI arms race, we're already behind.
From a technological perspective,
China is already ahead of the United States
in computer vision AI.
And in large language models, like ChatGPT,
they are fast followers.
In terms of military implementations,
they're outspending us:
adjusted for total military budget,
China's spending ten times more than the United States.
Why are we so far behind?
The answer is twofold.
First, data supremacy.
Despite having the largest fleet of military hardware in the world,
most of the data from this fleet is thrown away or inaccessible,

hidden away on hard drives that never see the light of day.
This is our Achilles heel.
In an AI war, everything boils down to data.
For defense AI,
data from the internet is not enough.
Most of the data needs to come from our military assets,
sensors and collaborations with tech companies.
Military commanders need to know how to use data as a military asset.
I've heard this first-hand many times,
from my conversations with military personnel,
including most recently
from Lieutenant General Richard R. Coffman,
deputy commanding general for United States Army Futures Command.
Second, despite being home to the leading technology companies
at the forefront of artificial intelligence,
the US tech industry has largely shied away
from taking on government contracts.
Somewhere along the line,
tech leaders decided that working with the government was taboo.
As a technologist,
I'm often asked how I'm bettering this world.
This is how I'm improving the future of our world:
by helping my country succeed
and providing the best tools and technology
to ensure that the United States government can defend its citizens,
allies and partners.

The Ukraine war has demonstrated that the nature of war has changed.
Through AI overmatch, Ukraine is challenging an adversary
with far superior numbers of troops and weapons.
Before the Ukraine war,
Russia had spent an estimated 65 billion US dollars
on its military expenditures,
whereas Ukraine only spent about six billion dollars.
It's estimated that Russia had over 900,000 military troops
and 1,300 aircraft,
whereas Ukraine only had 200,000 military troops and 130 aircraft.
Technologies such as drones,
AI-based targeting and image intelligence
and Javelin missiles
have enabled a shocking defense of Ukraine.
AI is proving invaluable for defending Ukrainian cities and infrastructure
against missile and drone bombardment.
At Scale, we’re using our novel machine learning algorithm
for battle damage assessment in key areas affected by the war.
We've rapidly analyzed over 2,000 square kilometers
and have identified over 370,000 structures,
including thousands not previously identified
by other open source data sets.
We focused on Kyiv, Kharkiv and Dnipro
and provided our data directly

in a publicly accessible data set to the broader AI community.
One of the key problems we're solving
is using AI to analyze massive amounts of imagery and detect objects
because humans just can't keep up.
We've received an overwhelming response from our free AI-ready data set
and have provided it directly to the United States and NATO allies.
And it's been downloaded over 2,000 times by AI companies, researchers,
developers and GIS practitioners.
AI can also be used for change detection.
Simply put, algorithms can constantly monitor imagery
and notify a human to investigate further if there's a change or a movement.
It's clear that AI is increasingly powering warfare.
And based on the rate of progress in the AI field,
I predict that in ten years, it will be the dominant force.
Disinformation and misinformation are already huge problems.
And this technology is only going to make it worse.
Tools like ChatGPT have enabled AI to generate imagery,
text, audio, video, code and even reason.
These tools can generate realistic-looking and realistic-sounding content,
which on top of bot-run social media accounts
will make it nearly impossible to identify disinformation
and misinformation online.
Bad actors can use these tools to supercharge misinformation
and propagate falsehoods.
China already uses disinformation campaigns
and social media manipulations heavily in Taiwan,
particularly during elections.
Or take Russia's propaganda machine,
which in the wake of Russia's invasion of Ukraine
created a deepfake of Ukrainian President Zelensky
calling for Ukrainian troops to surrender.
This deepfake was easy to spot,
but the next one may not be.
This also takes place within our borders,
from social media algorithm manipulation
to advertising microtargeting and geofencing,
to deepfakes of politicians and bot-run social media accounts.
The United States is not excused
from exacerbating disinformation and misinformation.
These tools are universally accessible at low or no cost,
meaning they can be employed by anyone anywhere
to undermine the sanctity of democracy globally.
However, all hope is not lost.
If we properly invest
into data infrastructure and data preparation,
all this can be avoided.
Deterrence is nothing new to military thinking.
As we saw in World War II with the atomic bomb,
it was a primary factor in deterring foreign adversaries
from going to nuclear war for more than six decades.
Because the stakes of going to war with such a technology

were simply too high.
We're likely to see a new calculus emerge with AI.
It's uncharted territory, nobody knows what it will look like
or the toll it will take.
How do we know if our AI is better than our adversaries'?
We won't.
But one thing is clear:
AI can only be as powerful as the underlying data
that is used to fuel its algorithms.
Data will be a new kind of ammunition in the era of AI warfare.
In the tech industry, we often talk about missions.
They're often frivolous.
Do they really change the world or save lives?
This mission, on the other hand, really matters.
The AI war will define the future of our world
and has the potential to shift the balance of diplomatic power.
It's clear that digital warfare
is not some dystopian reality, tucked away in a faraway future.
It is taking place in the here and now.
We cannot sit by the sidelines
and watch the rise of an authoritarian regime.
It is in moments like this
that technologists can either rise to the challenge or stand idle.
I encourage my fellow technologists
to understand the austerity and severity of our times
and commit themselves to supporting national security.
While I find it shocking that most American AI companies
have chosen not to support national security,
I do hope others join us.
We must fight for the world we want to live in.
It's never mattered more.
Thank you.

Will superintelligent AI end the world?

Eliezer Yudkowsky

Since 2001, I have been working on what we would now call
the problem of aligning artificial general intelligence:
how to shape the preferences and behavior
of a powerful artificial mind such that it does not kill everyone.
I more or less founded the field two decades ago,
when nobody else considered it rewarding enough to work on.
I tried to get this very important project started early
so we'd be in less of a drastic rush later.
I consider myself to have failed.

Nobody understands how modern AI systems do what they do.
They are giant, inscrutable matrices of floating point numbers
that we nudge in the direction of better performance
until they inexplicably start working.
At some point, the companies rushing headlong to scale AI
will cough out something that's smarter than humanity.
Nobody knows how to calculate when that will happen.
My wild guess is that it will happen after zero to two more breakthroughs
the size of transformers.
What happens if we build something smarter than us
that we understand that poorly?
Some people find it obvious that building something smarter than us
that we don't understand might go badly.
Others come in with a very wide range of hopeful thoughts
about how it might possibly go well.
Even if I had 20 minutes for this talk and months to prepare it,
I would not be able to refute all the ways people find to imagine
that things might go well.
But I will say that there is no standard scientific consensus
for how things will go well.
There is no hope that has been widely persuasive
and stood up to skeptical examination.
There is nothing resembling a real engineering plan for us surviving
that I could critique.
This is not a good place in which to find ourselves.
If I had more time,
I'd try to tell you about the predictable reasons
why the current paradigm will not work
to build a superintelligence that likes you
or is friends with you, or that just follows orders.
Why, if you press "thumbs up" when humans think that things went right
or "thumbs down" when another AI system thinks that they went wrong,
you do not get a mind that wants nice things
in a way that generalizes well outside the training distribution
to where the AI is smarter than the trainers.
You can search for "Yudkowsky list of lethalities" for more.

But to worry, you do not need to believe me

about exact predictions of exact disasters.
You just need to expect that things are not going to work great
on the first really serious, really critical try
because an AI system smart enough to be truly dangerous
was meaningfully different from AI systems stupider than that.
My prediction is that this ends up with us facing down something smarter than us
that does not want what we want,
that does not want anything we recognize as valuable or meaningful.
I cannot predict exactly how a conflict between humanity and a smarter AI would go
for the same reason I can't predict exactly how you would lose a chess game
to one of the current top AI chess programs, let's say Stockfish.
If I could predict exactly where Stockfish could move,
I could play chess that well myself.
I can't predict exactly how you'll lose to Stockfish,
but I can predict who wins the game.
I do not expect something actually smart to attack us with marching robot armies
with glowing red eyes
where there could be a fun movie about us fighting them.
I expect an actually smarter and uncaring entity
will figure out strategies and technologies
that can kill us quickly and reliably and then kill us.
I am not saying that the problem of aligning superintelligence
is unsolvable in principle.
I expect we could figure it out with unlimited time and unlimited retries,
which the usual process of science assumes that we have.
The problem here is the part where we don't get to say,
“Ha ha, whoops, that sure didn’t work.
That clever idea that used to work on earlier systems
sure broke down when the AI got smarter, smarter than us.”
We do not get to learn from our mistakes and try again
because everyone is already dead.
It is a large ask
to get an unprecedented scientific and engineering challenge
correct on the first critical try.
Humanity is not approaching this issue with remotely
the level of seriousness that would be required.
Some of the people leading these efforts
have spent the last decade not denying
that creating a superintelligence might kill everyone,
but joking about it.
We are very far behind.
This is not a gap we can overcome in six months,
given a six-month moratorium.
If we actually try to do this in real life,
we are all going to die.
People say to me at this point, what's your ask?
I do not have any realistic plan,
which is why I spent the last two decades
trying and failing to end up anywhere but here.
My best bad take is that we need an international coalition
banning large AI training runs,

including extreme and extraordinary measures
to have that ban be actually and universally effective,
like tracking all GPU sales,
monitoring all the data centers,
being willing to risk a shooting conflict between nations
in order to destroy an unmonitored data center
in a non-signatory country.
I say this, not expecting that to actually happen.
I say this expecting that we all just die.
But it is not my place to just decide on my own
that humanity will choose to die,
to the point of not bothering to warn anyone.
I have heard that people outside the tech industry
are getting this point faster than people inside it.
Maybe humanity wakes up one morning and decides to live.
Thank you for coming to my brief TED talk.

The incredible creativity of deepfakes — and the worrying future of AI

Tom Graham

Chris Anderson: So Tom, your company became prominent on the internet
with the release of a fake Tom Cruise video, DeepTomCruise,
that I think attracted like, a billion views on TikTok and Instagram.
Which leads me to my first question,
which is, please, can we, at TED, have our own Tom Cruise video, please?
Tom Graham: You know, I thought you might ask.
So a little earlier, we had a crack.
Maybe we'll have a look.
CA: Let's have a look.
DeepTomCruise: What's up, internet?
I'm north of the border, eh?
(Laughs)
At the TED conference.
It's not short for Theodore,
but nobody calls me Thomas, so it's cool.
It's Tom at TED.
Yes I ... Canada.
(Laughs)
Seriously, though, everybody here, very nice, very polite.
Especially the whales.

CA: I mean, how did you do that?
TG: So really at Metaphysic,
we specialize in creating artificially generated content
that looks and feels exactly like reality.
So we take kind of real-world data,
we train these neural nets,
and it can, more accurately than VFX or CGI,
really create this content that looks and feels so natural.
And so that is a great example of the AI being kind of prompted
by the natural performance of a person and kind of, the face goes on top.
CA: And it helps, the fact that your co-founder is, you know,
he's a pretty good Tom Cruise impersonator.
TG: Indeed, Miles Fisher is of a foremost Tom-Cruise-
but-not-Tom Cruise, yeah.
CA: I think you have another example as well.
Can we see that one?
TG: Yes, absolutely.
Going kind of beyond faces now, talking about voice.
[Original Spanish vocal performance]
[Audio-to-face synchronization]
[AI-generated vocal and visual performance]
CA: Tell us what's happening there.
TG: So what you see there is really the singing voice
of a lady singing Spanish
and Aloe Blacc, who sings "Wake Me Up," the Avicii song that he wrote with Avicii.

We are transporting her voice across to his face
so he doesn't sing in Spanish.
And then suddenly we transported again her voice into his voice.
So anybody in the future will be able to speak any language.
It'll look perfectly natural.
And this content is becoming more and more easy to create,
and eventually it will end up with a scale
where we will all be kind of, main characters
in our own content on the internet.
CA: OK.

Before we dig into that just a bit,
I mean, you've shown us recorded video there.
Rumor has it you can also do this with live video.
Can that be right?
TG: Yes, we can do it live real-time.
And this is like, really at the cutting edge
of what we can do today,
moving from offline processing to where processing it so fast
that you can do it in real-time.
CA: Well I'm going to challenge you and your team to do a bit of a first here.
There's video of you right up on that screen.
Show us something surprising you can --

Oh, my gosh.
TG: So there we go.
This is, you know, a live, real-time model of Chris
on top of me, running in real-time.
CA: And next, you'll tell me that it can ...

Oh, Lord.
I am so uncomfortable with this.

I am so uncomfortable.
Can it do voice as well?
TG: Um.
(Live AI-generated Chris Anderson's voice)
We think it can, we're really pushing the limits of AI technology now.
And I'm talking exactly as I would
as Thomas Graham, but it's coming out is the one and only of Chris Anderson.
CA: I'm deeply sorry everyone to subject you to this.
You know, there's something possibly even worse.
So to come clean on this,
they took some shots of me a couple of weeks ago
doing different facial expressions, so they captured a video model.
And it turns out, I discovered this week,
that they can apply that, not just to Tom, but to anyone.
And so, my dear friend Sunny Bates is here in the front row.
Sunny, do you consent to channel inner me for a minute?
Can we try that?
I'm really worried about this.

Do we have Sunny on screen?
Oh, there's Sunny.
Oh, no, oh, no.

Oh, no.

TG: Chris, you look amazing.
CA: All right, enough of that, cut that right now.
(Sunny Bates mouthing) More, more!
TG: Yes, more, more of this.
CA: Tom, Tom, Sunny Bates is the woman who introduced me to TED.
Without Sunny Bates, none of us would actually be here now.
And we reward her with this?
TG: And now you've finally become the master, you know?
CA: Oh, so look.
OK, amazing.
It obviously occurs to everyone in this room
that there are some things that can go horribly wrong with this.

And, you know, we've already seen examples online of, you know,
we've had photographs of Trump being arrested that were unreal,
there could be video of it.
There’s pornography that can use the faces of celebrities.
All these things that we’ve seen, deepfake.
How do you feel about the downside of this technology?
TG: So personally, you know,
we build this stuff and I'm worried, right?
Worried is the right instinct for everybody to have.
And then beyond that, think about, you know,
what can we do to prepare ourselves?
How can we try to impact the future as it spirals in this direction,
where as individuals, it's going to be kind of difficult to understand
what's computer-generated and what's real.
And so there are things that we can do there.
Raising public awareness of manipulated media.
That's one, you know, this is a great forum for that.
CA: Will you claim to me
that if you were to shut down your company right now,
it wouldn't stop the problem of deepfake videos
because the technology's out there, that that's going to happen anyway,
that that's not within your control?
TG: Yes.
We're talking about content that is so compelling.
If we put any of ourselves inside content
and maybe it is talking to our loved ones
or just talking to our friends on the beach
and it's so realistic, it looks real,
it's so compelling, everybody is trying to create this content today.
All of the GPUs in the universe are driving at trying to create this.
So it doesn’t matter what any one person does.

This will happen and it’s happening very, very quickly.
CA: So, I mean, we'll talk about the upside in a second.
But it seems like we are going to have to get used to a world
where we and our children will no longer be able to trust the evidence of our eyes.
TG: I think so.
We are going to have to understand a new set of institutions
to verify what is authentic media.
But then we can begin to lean into some of the creative things
that happen from it.
And there are benefits that come with that too.
So it'll be an accommodation.
CA: So talk a bit about the benefits.
I mean, obviously, on the entertainment side,
there's an amazing number of possibilities.
You have Tom Cruise,
we can have [“Mission: Impossible] 273” in the year 2150, right?
Like, you will be with us forever.
TG: We're working on number 75,001 right now, yeah.
CA: I guess that is kind of amazing.
Like, we love lots of people in the world,
and we want the possibility that, with their permission,
we can do more with them.
Talk about some other possible upsides.
TG: What we've seen in kind of building this content
and watching people interact with it,
especially if they're kind of, interacting with themselves,
maybe it's their younger 20-year-old self
or maybe they're interacting with their partner,
but the young version of their partner,
there's this tremendous emotional connection that comes
from the very, very photorealistic beyond the uncanny kind of content.
And so if we can start deploying that among regular people,
if we can scale it up so that it works for any kind of person,
then we can begin to kind of, you know, have more interesting,
meaningful, human kind of interactions and relationships online.
And since the pandemic, we all spend more time online.
But it's chat and it's email.
If we could get more human emotion, more feeling,
there's a lot that we can do with that, right?
And so, education is a good example.
We could have an inspiring teacher in thousands of classrooms
around the world,
speaking every language in the world at the same time,
and students could interact with each other
in a way that goes beyond Zoom.
There can be real cultural exchange, real socialization.
There's a lot that we can do beyond here.
CA: So that teacher example is powerful to me,
like, the fact that a single teacher
could turn a written lesson into video in any number of languages
and extend indefinitely.

That seems like a real amplification potentially of good human intent.
I'm excited by that.
I still don't get the family side of it.
Like, if you want to have a human connection with someone in your family,
like, won't people just be creeped out by,
"Oh my God, I was just looking at your avatar,
I thought it was you."
You know, like, isn't that just creepy?
TG: I think there's definitely a creepy element to this, right?
And then you go beyond that and the creepiness drops away
and the medium drops away, and it's the content
and the connection that's there.
So, you know, I imagine that, you know,
if I collect data from my grandparents who are very old today,
then in the future I'll be able to relive experiences with them
and communicate with them.
And that, it’s not going to be any good for them, right?
They're probably going to have passed on, but for me,
it'll help me process who I am, my relationship with them.
That idea of kind of, decoupling human experience,
both from where it happens
and the moment in time that it happens,
I think that we can create these experiences
through hyperreal, photorealistic media
that allow us to share the best of our experiences,
the best of who we are.
CA: I'm going to be very curious to see who can feel that right now.
My guess is that there's going to have to be lots of experiments
and a lot of things that are going to creep us out.
And maybe we'll find some things that are just absolutely incredible.
But I have an important question for you,
which is how the hell do I control me now?
You've got me on video.
What's to stop me being misused across the internet now?
TG: That's right.
I think that, as we kind of, allow companies
to create these realistic experiences,
as individuals, we need to be empowered to own our real-world data,
the data used to train the algorithms,
and we need to control how our photorealistic avatars are created
and where they're used.
So to this extent,
I was looking at kind of conventional, current legal institutions
to see what we could do to create new rights.
So I created a photorealistic avatar of myself,
submitted it to the US Copyright Office
to see if they would register my copyright in it,
and this is what the video looked like.
(Video) TG: Here is the AI, realistic version of myself.
Even if the appearance of this AI representation of myself
may change cosmetically,

or if I change my hair or add creative features,
my intention is to create this AI version of myself
that embodies the essence of who I am as a person under any circumstance.
CA: Wow.
So I think you have shown us
what is going to be a repeated theme in this conference,
which is that future is going to be weird and wonderful at the same time.
Quite what the balance is between those two, TBD.
But this is a world where each of us is going to have to think differently
about who we are and claim these rights.
What you're saying is that if people have this right,
you can picture a world where, for example,
if a video goes viral on YouTube without your consent,
you'll be able to take it down because of some link back to this.
TG: That's right.
I think the most important thing that we can do
when we're talking about data from a real world
being able to power these things
is that we need to own property rights over the data.
We shouldn't sign it over to companies through terms of service,
we shouldn't give it away.
If you fundamentally own it, then you can be in control.
And you're at the ground level of all of the economies
and all of the use cases,
that are going to spin up through history.
It's a lot, but we're on the way now.
CA: Tom Graham, thank you so much for sharing this technology at TED.
And Sunny Bates, I'm so sorry.

Are insect brains the secret to great AI?

Frances S. Chance

Creating intelligence on a computer.
This has been the Holy Grail for artificial intelligence
for quite some time.
But how do we get there?
So we view ourselves as highly intelligent beings.
So it's logical to study our own brains,
the substrate of our cognition, for creating artificial intelligence.
Imagine if we could replicate how our own brains work on a computer.
But now consider the journey that would be required.
The human brain contains 86 billion neurons.
Each is constantly communicating with thousands of others,
and each has individual characteristics of its own.
Capturing the human brain on a computer
may simply be too big and too complex a problem
to tackle with the technology and the knowledge that we have today.
I believe that we can capture a brain on a computer,
but we have to start smaller.
Much smaller.
These insects have three of the most fascinating brains in the world to me.
While they do not possess human-level intelligence,
each is remarkable at a particular task.
Think of them as highly trained specialists.
African dung beetles are really good at rolling large balls in straight lines.

Now, if you've ever made a snowman,
you know that rolling a large ball is not easy.
Now picture trying to make that snowman
when the ball of snow is as big as you are
and you're standing on your head.

Sahara desert ants are navigation specialists.
They might have to wander a considerable distance to forage for food.
But once they do find sustenance,
they know how to calculate the straightest path home.
And the dragonfly is a hunting specialist.
In the wild, dragonflies capture approximately 95 percent
of the prey they choose to go after.
These insects are so good at their specialties
that neuroscientists such as myself study them as model systems
to understand how animal nervous systems solve particular problems.
And in my own research, I study brains to bring these solutions,
the best that biology has to offer, to computers.
So consider the dragonfly brain.
It has only on the order of one million neurons.
Now, it's still not easy to unravel a circuit of even one million neurons.
But given the choice
between trying to tease apart the one-million-neuron brain
versus the 86-billion-neuron brain,

which would you choose to try first?

When studying these smaller insect brains,
the immediate goal is not human intelligence.
We study these brains for what the insects do well.
And in the case of the dragonfly, that's interception.
So when dragonflies are hunting,
they do more than just fly straight at the prey.
They fly in such a way that they will intercept it.
They aim for where the prey is going to be.
Much like a soccer player, running to intercept a pass.
To do this correctly,
dragonflies need to perform what is known as a coordinate transformation,
going from the eye's frame of reference, or what the dragonfly sees,
to the body's frame of reference,
or how the dragonfly needs to turn its body to intercept.
Coordinate transformations are a basic calculation
that animals need to perform to interact with the world.
We do them instinctively every time we reach for something.
When I reach for an object straight in front of me,
my arm takes a very different trajectory than if I turn my head,
look at that same object when it is off to one side
and reach for it there.
In both cases, my eyes see the same image of that object,
but my brain is sending my arm on a very different trajectory
based on the position of my neck.
And dragonflies are fast.
This means they calculate fast.
The latency, or the time it takes for a dragonfly to respond
once it sees the prey turn,
is about 50 milliseconds.
This latency is remarkable.
For one thing, it's only half the time of a human eye blink.
But for another thing,
it suggests that dragonflies capture how to intercept
in only relatively or surprisingly few computational steps.
So in the brain,
a computational step is a single neuron
or a layer of neurons working in parallel.
It takes a single neuron about 10 milliseconds
to add up all its inputs and respond.
The 50-millisecond response time
means that once the dragonfly sees its prey turn,
there's only time for maybe four of these computational steps
or four layers of neurons, working in sequence, one after the other,
to calculate how the dragonfly needs to turn.
In other words, if I want to study
how the dragonfly does coordinate transformations,
the neural circuit that I need to understand,
the neural circuit that I need to study,
can have at most four layers of neurons.

Each layer may have many neurons,
but this is a small neural circuit.
Small enough that we can identify it
and study it with the tools that are available today.
And this is what I'm trying to do.
I have built a model of what I believe is the neural circuit
that calculates how the dragonfly should turn.
And here is the cool result.
In the model,
dragonflies do coordinate transformations in only one computational step,
one layer of neurons.
This is something we can test and understand.
In a computer simulation,
I can predict the activities of individual neurons
while the dragonfly is hunting.
For example, here I am predicting the action potentials, or the spikes,
that are fired by one of these neurons
when the dragonfly sees the prey move.
To test the model, my collaborators and I
are now comparing these predicted neural responses
with responses of neurons recorded in living dragonfly brains.
These are ongoing experiments
in which we put living dragonflies in virtual reality.

Now, it's not practical to put VR goggles on a dragonfly.
So instead, we show movies of moving targets to the dragonfly,
while an electrode records activity patterns of individual neurons
in the brain.
Yeah, he likes the movies.
If the responses that we record in the brain
match those predicted by the model,
we will have identified which neurons are responsible
for coordinate transformations.
The next step will be to understand the specifics
of how these neurons work together to do the calculation.
But this is how we begin to understand how brains do basic
or primitive calculations.
Calculations that I regard as building blocks for more complex functions,
not only for interception but also for cognition.
The way that these neurons compute may be different from anything
that exists on a computer today.
And the goal of this work is to do more than just write code
that replicates the activity patterns of neurons.
We aim to build a computer chip
that not only does the same things as biological brains
but does them in the same way as biological brains.
This could lead to drones driven by computers
the same size of the dragonfly's brain
that captures some targets and avoid others.
Personally, I'm hoping for a small army of these
to defend my backyard from mosquitoes in the summer.

The GPS on your phone could be replaced by a new navigation device
based on dung beetles or ants
that could guide you to the straight or the easy path home.
And what would the power requirements of these devices be like?
As small as it is --
Or, sorry -- as large as it is,
the human brain is estimated to have the same power requirements
as a 20-watt light bulb.
Imagine if all brain-inspired computers
had the same extremely low-power requirements.
Your smartphone or your smartwatch probably needs charging every day.
Your new brain-inspired device might only need charging every few months,
or maybe even every few years.
The famous physicist, Richard Feynman, once said,
"What I cannot create, I do not understand."
What I see in insect nervous systems
is an opportunity to understand brains
through the creation of computers that work as brains do.
And creation of these computers will not just be for knowledge.
There's potential for real impact on your devices, your vehicles,
maybe even artificial intelligences.
So next time you see an insect,
consider that these tiny brains can lead to remarkable computers.
And think of the potential that they offer us for the future.
Thank you.

AI-generated creatures that stretch the boundaries of imagination

Sofia Crespo

I'd like to start by asking you to imagine a color
that you've never seen before.
Just for a second give this a try.
Can you actually visualize a color that you've never been able to perceive?
I never seem to get tired of trying this
although I know it's not an easy challenge.
And the thing is,
we can't imagine something without drawing upon our experiences.
A color we haven't yet seen
outside the spectrum we can perceive
is outside our ability to conjure up.
It's almost like there's a boundary to our imagination
where all the colors we can imagine
can only be various shades of other colors we have previously seen.
Yet we know for a fact
that those color frequencies outside our visible spectrum are there.
And scientists believe that there are species
that have many more photo receptors
than just the three color ones we humans have.
Which, by the way,
not all humans see the world in the same way.
Some of us are colorblind to various degrees,
and very often we don't even agree on small things,
like if a dress on the internet is blue and black or white and gold.
But my favorite creature, one of my favorite creatures,
is the peacock mantis shrimp,
which is estimated to have 12 to 16 photo receptors.
And that indicates the world to them might look so much more colorful.
So what about artificial intelligence?
Can AI help us see beyond our human capabilities?
Well, I've been working with AI for the past five years,
and in my experience, it can see within the data it gets fed.
But then you might be wondering, OK,
if AI can't help imagine anything new,
why would an artist see any point in using it?
And my answer to that is because I think that it can help augment our creativity
as there's value in creating combinations of known elements to form new ones.
And this boundary of what we can imagine based on what we have experienced
is the place that I have been exploring.
For me, it started with jellyfish on a screen at an aquarium
and wearing those old 3D glasses, which I hope you remember,
the ones with the blue and red lens.
And this experience made me want to recreate their textures.
But not just that,
I also wanted to create new jellyfish
that I hadn't seen before, like these.

And what started with jellyfish,
very quickly escalated to other sea creatures
like sea anemone, coral and fish.
And then from there came amphibians, birds and insects.
And this became a series called “Neural Zoo”.
But when you look closely, what do you see?
There's no single creature in these images.
And AI augments my creative process
by allowing me to distill and recombine textures.
And that's something that would otherwise take me months to draw by hand.
Plus I'm actually terrible at drawing.
So you could say, in a way, what I'm doing
is a contemporary version of something
that humans have already been doing for a long time,
even before cameras existed.
In medieval times,
people went on expeditions,
and when they came back they would share about what they saw
to an illustrator.
And the illustrator, having never seen what was being described,
would end up drawing
based on the creatures that they had previously seen
and in the process creating hybrid animals of some sort.
So an explorer might describe a beaver, but having never seen one,
the illustrator might give it the head of a rodent,
the body of a dog and a fish-like tail.
In the series “Artificial Natural History”,
I took thousands of illustrations from a natural history archives,
and I fed them to a neural network to generate new versions of them.
But up until now, all my work was done in 2D.
And with the help of my studio partner, Feileacan McCormick,
we decided to train a neural network on a data set of 3D scanned beetles.
But I must warn you that our first results were extremely blurry,
and they looked like the blobs you see here.
And this could be due to many reasons,
but one of them being that there aren't really a lot
of openly available data sets of 3D insects.
And also we were repurposing
a neural network that normally gets used to generate images to generate 3D.
So believe it or not, these are very exciting blobs to us.
But with time and some very hacky solutions
like data augmentation,
where we threw in ants and other beetle-like insects
to enhance the data set,
we ended up getting this,
which we've been told they look like grilled chicken.

But hungry for more, we pushed our technique,
and eventually they ended up looking like this.
We use something called 3D style transfer to map textures onto them,
and we also trained a natural language model

to generate scientific-like names
and anatomical descriptions.
And eventually we even found a network architecture that could handle 3D meshes.
So they ended up looking like this.
And for us, this became a way of creating kind of a speculative study --

A speculative study of creatures that never existed,
kind of like a speculative biology.
But I didn't want to talk about AI and its potential
unless it brought me closer to a real species.
Which of these do you think is easier to find data about online?

Yeah, well, as you guessed correctly, the red panda.
And this maybe could be due to many reasons,
but one of them being how cute they are,
which means we photograph and talk about them a lot,
unlike the boreal felt lichen.
But both of them are classified as endangered.
So I wanted to bring visibility to other endangered species
that don't get the same amount of digital representation
as a cute, fluffy red panda.
And to do this,
we trained an AI on millions of images of the natural world,
and then we prompted with text
to generate some of these creatures.
So when prompted with a text,
"an image of a critically endangered spider, the peacock tarantula"
and its scientific name,
our model generated this.
And here's an image of the real peacock tarantula,
which is a wonderful spider endemic to India.
But when prompted with a text
"an image of a critically endangered bird, the mangrove finch,"
our model generated this.
And here's a photo of the real mangrove finch.
Both these creatures exist in the wild,
but the accuracy of each generated image is fully dependent on the data available.
These chimeras of our everyday data
to me are a different way of how the future could be.
Not in a literal sense, perhaps,
but in the sense that through practicing the expanding of our own imagination
about the ecosystems we are a part of,
we might just be better equipped to recognize new opportunities
and potential.
Knowing that there's a boundary to our imagination
doesn't have to feel limiting.
On the contrary,
it can help motivate us to expand that boundary further
and to seek out colors and things we haven't yet seen
and perhaps enrich our imagination as a result.
So thank you.

Get ready for hybrid thinking

Ray Kurzweil

Let me tell you a story.
It goes back 200 million years.
It's a story of the neocortex,
which means "new rind."
So in these early mammals,
because only mammals have a neocortex,
rodent-like creatures.
It was the size of a postage stamp and just as thin,
and was a thin covering around
their walnut-sized brain,
but it was capable of a new type of thinking.
Rather than the fixed behaviors
that non-mammalian animals have,
it could invent new behaviors.
So a mouse is escaping a predator,
its path is blocked,
it'll try to invent a new solution.
That may work, it may not,
but if it does, it will remember that
and have a new behavior,
and that can actually spread virally
through the rest of the community.
Another mouse watching this could say,
"Hey, that was pretty clever, going around that rock,"
and it could adopt a new behavior as well.
Non-mammalian animals
couldn't do any of those things.
They had fixed behaviors.
Now they could learn a new behavior
but not in the course of one lifetime.
In the course of maybe a thousand lifetimes,
it could evolve a new fixed behavior.
That was perfectly okay 200 million years ago.
The environment changed very slowly.
It could take 10,000 years for there to be
a significant environmental change,
and during that period of time
it would evolve a new behavior.
Now that went along fine,
but then something happened.
Sixty-five million years ago,
there was a sudden, violent change to the environment.
We call it the Cretaceous extinction event.
That's when the dinosaurs went extinct,
that's when 75 percent of the
animal and plant species went extinct,
and that's when mammals
overtook their ecological niche,

and to anthropomorphize, biological evolution said,
"Hmm, this neocortex is pretty good stuff,"
and it began to grow it.
And mammals got bigger,
their brains got bigger at an even faster pace,
and the neocortex got bigger even faster than that
and developed these distinctive ridges and folds
basically to increase its surface area.
If you took the human neocortex
and stretched it out,
it's about the size of a table napkin,
and it's still a thin structure.
It's about the thickness of a table napkin.
But it has so many convolutions and ridges
it's now 80 percent of our brain,
and that's where we do our thinking,
and it's the great sublimator.
We still have that old brain
that provides our basic drives and motivations,
but I may have a drive for conquest,
and that'll be sublimated by the neocortex
into writing a poem or inventing an app
or giving a TED Talk,
and it's really the neocortex that's where
the action is.
Fifty years ago, I wrote a paper
describing how I thought the brain worked,
and I described it as a series of modules.
Each module could do things with a pattern.
It could learn a pattern. It could remember a pattern.
It could implement a pattern.
And these modules were organized in hierarchies,
and we created that hierarchy with our own thinking.
And there was actually very little to go on
50 years ago.
It led me to meet President Johnson.
I've been thinking about this for 50 years,
and a year and a half ago I came out with the book
"How To Create A Mind,"
which has the same thesis,
but now there's a plethora of evidence.
The amount of data we're getting about the brain
from neuroscience is doubling every year.
Spatial resolution of brainscanning of all types
is doubling every year.
We can now see inside a living brain
and see individual interneural connections
connecting in real time, firing in real time.
We can see your brain create your thoughts.
We can see your thoughts create your brain,
which is really key to how it works.

So let me describe briefly how it works.
I've actually counted these modules.
We have about 300 million of them,
and we create them in these hierarchies.
I'll give you a simple example.
I've got a bunch of modules
that can recognize the crossbar to a capital A,
and that's all they care about.
A beautiful song can play,
a pretty girl could walk by,
they don't care, but they see a crossbar to a capital A,
they get very excited and they say "crossbar,"
and they put out a high probability
on their output axon.
That goes to the next level,
and these layers are organized in conceptual levels.
Each is more abstract than the next one,
so the next one might say "capital A."
That goes up to a higher level that might say "Apple."
Information flows down also.
If the apple recognizer has seen A-P-P-L,
it'll think to itself, "Hmm, I think an E is probably likely,"
and it'll send a signal down to all the E recognizers
saying, "Be on the lookout for an E,
I think one might be coming."
The E recognizers will lower their threshold
and they see some sloppy thing, could be an E.
Ordinarily you wouldn't think so,
but we're expecting an E, it's good enough,
and yeah, I've seen an E, and then apple says,
"Yeah, I've seen an Apple."
Go up another five levels,
and you're now at a pretty high level
of this hierarchy,
and stretch down into the different senses,
and you may have a module that sees a certain fabric,
hears a certain voice quality, smells a certain perfume,
and will say, "My wife has entered the room."
Go up another 10 levels, and now you're at
a very high level.
You're probably in the frontal cortex,
and you'll have modules that say, "That was ironic.
That's funny. She's pretty."
You might think that those are more sophisticated,
but actually what's more complicated
is the hierarchy beneath them.
There was a 16-year-old girl, she had brain surgery,
and she was conscious because the surgeons
wanted to talk to her.
You can do that because there's no pain receptors
in the brain.

And whenever they stimulated particular,
very small points on her neocortex,
shown here in red, she would laugh.
So at first they thought they were triggering
some kind of laugh reflex,
but no, they quickly realized they had found
the points in her neocortex that detect humor,
and she just found everything hilarious
whenever they stimulated these points.
"You guys are so funny just standing around,"
was the typical comment,
and they weren't funny,
not while doing surgery.
So how are we doing today?
Well, computers are actually beginning to master
human language with techniques
that are similar to the neocortex.
I actually described the algorithm,
which is similar to something called
a hierarchical hidden Markov model,
something I've worked on since the '90s.
"Jeopardy" is a very broad natural language game,
and Watson got a higher score
than the best two players combined.
It got this query correct:
"A long, tiresome speech
delivered by a frothy pie topping,"
and it quickly responded, "What is a meringue harangue?"
And Jennings and the other guy didn't get that.
It's a pretty sophisticated example of
computers actually understanding human language,
and it actually got its knowledge by reading
Wikipedia and several other encyclopedias.
Five to 10 years from now,
search engines will actually be based on
not just looking for combinations of words and links
but actually understanding,
reading for understanding the billions of pages
on the web and in books.
So you'll be walking along, and Google will pop up
and say, "You know, Mary, you expressed concern
to me a month ago that your glutathione supplement
wasn't getting past the blood-brain barrier.
Well, new research just came out 13 seconds ago
that shows a whole new approach to that
and a new way to take glutathione.
Let me summarize it for you."
Twenty years from now, we'll have nanobots,
because another exponential trend
is the shrinking of technology.
They'll go into our brain

through the capillaries
and basically connect our neocortex
to a synthetic neocortex in the cloud
providing an extension of our neocortex.
Now today, I mean,
you have a computer in your phone,
but if you need 10,000 computers for a few seconds
to do a complex search,
you can access that for a second or two in the cloud.
In the 2030s, if you need some extra neocortex,
you'll be able to connect to that in the cloud
directly from your brain.
So I'm walking along and I say,
"Oh, there's Chris Anderson.
He's coming my way.
I'd better think of something clever to say.
I've got three seconds.
My 300 million modules in my neocortex
isn't going to cut it.
I need a billion more."
I'll be able to access that in the cloud.
And our thinking, then, will be a hybrid
of biological and non-biological thinking,
but the non-biological portion
is subject to my law of accelerating returns.
It will grow exponentially.
And remember what happens
the last time we expanded our neocortex?
That was two million years ago
when we became humanoids
and developed these large foreheads.
Other primates have a slanted brow.
They don't have the frontal cortex.
But the frontal cortex is not really qualitatively different.
It's a quantitative expansion of neocortex,
but that additional quantity of thinking
was the enabling factor for us to take
a qualitative leap and invent language
and art and science and technology
and TED conferences.
No other species has done that.
And so, over the next few decades,
we're going to do it again.
We're going to again expand our neocortex,
only this time we won't be limited
by a fixed architecture of enclosure.
It'll be expanded without limit.
That additional quantity will again
be the enabling factor for another qualitative leap
in culture and technology.
Thank you very much.

Machine intelligence makes human morals more important

Zeynep Tufekci

So, I started my first job as a computer programmer
in my very first year of college --
basically, as a teenager.
Soon after I started working,
writing software in a company,
a manager who worked at the company came down to where I was,
and he whispered to me,
"Can he tell if I'm lying?"
There was nobody else in the room.
"Can who tell if you're lying? And why are we whispering?"
The manager pointed at the computer in the room.
"Can he tell if I'm lying?"
Well, that manager was having an affair with the receptionist.

And I was still a teenager.
So I whisper-shouted back to him,
"Yes, the computer can tell if you're lying."

Well, I laughed, but actually, the laugh's on me.
Nowadays, there are computational systems
that can suss out emotional states and even lying
from processing human faces.
Advertisers and even governments are very interested.
I had become a computer programmer
because I was one of those kids crazy about math and science.
But somewhere along the line I'd learned about nuclear weapons,
and I'd gotten really concerned with the ethics of science.
I was troubled.
However, because of family circumstances,
I also needed to start working as soon as possible.
So I thought to myself, hey, let me pick a technical field
where I can get a job easily
and where I don't have to deal with any troublesome questions of ethics.
So I picked computers.

Well, ha, ha, ha! All the laughs are on me.
Nowadays, computer scientists are building platforms
that control what a billion people see every day.
They're developing cars that could decide who to run over.
They're even building machines, weapons,
that might kill human beings in war.
It's ethics all the way down.
Machine intelligence is here.
We're now using computation to make all sort of decisions,
but also new kinds of decisions.
We're asking questions to computation that have no single right answers,
that are subjective
and open-ended and value-laden.

We're asking questions like,
"Who should the company hire?"
"Which update from which friend should you be shown?"
"Which convict is more likely to reoffend?"
"Which news item or movie should be recommended to people?"
Look, yes, we've been using computers for a while,
but this is different.
This is a historical twist,
because we cannot anchor computation for such subjective decisions
the way we can anchor computation for flying airplanes, building bridges,
going to the moon.
Are airplanes safer? Did the bridge sway and fall?
There, we have agreed-upon, fairly clear benchmarks,
and we have laws of nature to guide us.
We have no such anchors and benchmarks
for decisions in messy human affairs.
To make things more complicated, our software is getting more powerful,
but it's also getting less transparent and more complex.
Recently, in the past decade,
complex algorithms have made great strides.
They can recognize human faces.
They can decipher handwriting.
They can detect credit card fraud
and block spam
and they can translate between languages.
They can detect tumors in medical imaging.
They can beat humans in chess and Go.
Much of this progress comes from a method called "machine learning."
Machine learning is different than traditional programming,
where you give the computer detailed, exact, painstaking instructions.
It's more like you take the system and you feed it lots of data,
including unstructured data,
like the kind we generate in our digital lives.
And the system learns by churning through this data.
And also, crucially,
these systems don't operate under a single-answer logic.
They don't produce a simple answer; it's more probabilistic:
"This one is probably more like what you're looking for."
Now, the upside is: this method is really powerful.
The head of Google's AI systems called it,
"the unreasonable effectiveness of data."
The downside is,
we don't really understand what the system learned.
In fact, that's its power.
This is less like giving instructions to a computer;
it's more like training a puppy-machine-creature
we don't really understand or control.
So this is our problem.
It's a problem when this artificial intelligence system gets things wrong.
It's also a problem when it gets things right,
because we don't even know which is which when it's a subjective problem.

We don't know what this thing is thinking.
So, consider a hiring algorithm --
a system used to hire people, using machine-learning systems.
Such a system would have been trained on previous employees' data
and instructed to find and hire
people like the existing high performers in the company.
Sounds good.
I once attended a conference
that brought together human resources managers and executives,
high-level people,
using such systems in hiring.
They were super excited.
They thought that this would make hiring more objective, less biased,
and give women and minorities a better shot
against biased human managers.
And look -- human hiring is biased.
I know.
I mean, in one of my early jobs as a programmer,
my immediate manager would sometimes come down to where I was
really early in the morning or really late in the afternoon,
and she'd say, "Zeynep, let's go to lunch!"
I'd be puzzled by the weird timing.
It's 4pm. Lunch?
I was broke, so free lunch. I always went.
I later realized what was happening.
My immediate managers had not confessed to their higher-ups
that the programmer they hired for a serious job was a teen girl
who wore jeans and sneakers to work.
I was doing a good job, I just looked wrong
and was the wrong age and gender.
So hiring in a gender- and race-blind way
certainly sounds good to me.
But with these systems, it is more complicated, and here's why:
Currently, computational systems can infer all sorts of things about you
from your digital crumbs,
even if you have not disclosed those things.
They can infer your sexual orientation,
your personality traits,
your political leanings.
They have predictive power with high levels of accuracy.
Remember -- for things you haven't even disclosed.
This is inference.
I have a friend who developed such computational systems
to predict the likelihood of clinical or postpartum depression
from social media data.
The results are impressive.
Her system can predict the likelihood of depression
months before the onset of any symptoms --
months before.
No symptoms, there's prediction.
She hopes it will be used for early intervention. Great!

But now put this in the context of hiring.
So at this human resources managers conference,
I approached a high-level manager in a very large company,
and I said to her, "Look, what if, unbeknownst to you,
your system is weeding out people with high future likelihood of depression?
They're not depressed now, just maybe in the future, more likely.
What if it's weeding out women more likely to be pregnant
in the next year or two but aren't pregnant now?
What if it's hiring aggressive people because that's your workplace culture?"
You can't tell this by looking at gender breakdowns.
Those may be balanced.
And since this is machine learning, not traditional coding,
there is no variable there labeled "higher risk of depression,"
"higher risk of pregnancy,"
"aggressive guy scale."
Not only do you not know what your system is selecting on,
you don't even know where to begin to look.
It's a black box.
It has predictive power, but you don't understand it.
"What safeguards," I asked, "do you have
to make sure that your black box isn't doing something shady?"
She looked at me as if I had just stepped on 10 puppy tails.

She stared at me and she said,
"I don't want to hear another word about this."
And she turned around and walked away.
Mind you -- she wasn't rude.
It was clearly: what I don't know isn't my problem, go away, death stare.

Look, such a system may even be less biased
than human managers in some ways.
And it could make monetary sense.
But it could also lead
to a steady but stealthy shutting out of the job market
of people with higher risk of depression.
Is this the kind of society we want to build,
without even knowing we've done this,
because we turned decision-making to machines we don't totally understand?
Another problem is this:
these systems are often trained on data generated by our actions,
human imprints.
Well, they could just be reflecting our biases,
and these systems could be picking up on our biases
and amplifying them
and showing them back to us,
while we're telling ourselves,
"We're just doing objective, neutral computation."
Researchers found that on Google,
women are less likely than men to be shown job ads for high-paying jobs.
And searching for African-American names
is more likely to bring up ads suggesting criminal history,

even when there is none.
Such hidden biases and black-box algorithms
that researchers uncover sometimes but sometimes we don't know,
can have life-altering consequences.
In Wisconsin, a defendant was sentenced to six years in prison
for evading the police.
You may not know this,
but algorithms are increasingly used in parole and sentencing decisions.
He wanted to know: How is this score calculated?
It's a commercial black box.
The company refused to have its algorithm be challenged in open court.
But ProPublica, an investigative nonprofit, audited that very algorithm
with what public data they could find,
and found that its outcomes were biased
and its predictive power was dismal, barely better than chance,
and it was wrongly labeling black defendants as future criminals
at twice the rate of white defendants.
So, consider this case:
This woman was late picking up her godsister
from a school in Broward County, Florida,
running down the street with a friend of hers.
They spotted an unlocked kid's bike and a scooter on a porch
and foolishly jumped on it.
As they were speeding off, a woman came out and said,
"Hey! That's my kid's bike!"
They dropped it, they walked away, but they were arrested.
She was wrong, she was foolish, but she was also just 18.
She had a couple of juvenile misdemeanors.
Meanwhile, that man had been arrested for shoplifting in Home Depot --
85 dollars' worth of stuff, a similar petty crime.
But he had two prior armed robbery convictions.
But the algorithm scored her as high risk, and not him.
Two years later, ProPublica found that she had not reoffended.
It was just hard to get a job for her with her record.
He, on the other hand, did reoffend
and is now serving an eight-year prison term for a later crime.
Clearly, we need to audit our black boxes
and not have them have this kind of unchecked power.

Audits are great and important, but they don't solve all our problems.
Take Facebook's powerful news feed algorithm --
you know, the one that ranks everything and decides what to show you
from all the friends and pages you follow.
Should you be shown another baby picture?

A sullen note from an acquaintance?
An important but difficult news item?
There's no right answer.
Facebook optimizes for engagement on the site:
likes, shares, comments.
In August of 2014,

protests broke out in Ferguson, Missouri,
after the killing of an African-American teenager by a white police officer,
under murky circumstances.
The news of the protests was all over
my algorithmically unfiltered Twitter feed,
but nowhere on my Facebook.
Was it my Facebook friends?
I disabled Facebook's algorithm,
which is hard because Facebook keeps wanting to make you
come under the algorithm's control,
and saw that my friends were talking about it.
It's just that the algorithm wasn't showing it to me.
I researched this and found this was a widespread problem.
The story of Ferguson wasn't algorithm-friendly.
It's not "likable."
Who's going to click on "like?"
It's not even easy to comment on.
Without likes and comments,
the algorithm was likely showing it to even fewer people,
so we didn't get to see this.
Instead, that week,
Facebook's algorithm highlighted this,
which is the ALS Ice Bucket Challenge.
Worthy cause; dump ice water, donate to charity, fine.
But it was super algorithm-friendly.
The machine made this decision for us.
A very important but difficult conversation
might have been smothered,
had Facebook been the only channel.
Now, finally, these systems can also be wrong
in ways that don't resemble human systems.
Do you guys remember Watson, IBM's machine-intelligence system
that wiped the floor with human contestants on Jeopardy?
It was a great player.
But then, for Final Jeopardy, Watson was asked this question:
"Its largest airport is named for a World War II hero,
its second-largest for a World War II battle."
(Hums Final Jeopardy music)
Chicago.
The two humans got it right.
Watson, on the other hand, answered "Toronto" --
for a US city category!
The impressive system also made an error
that a human would never make, a second-grader wouldn't make.
Our machine intelligence can fail
in ways that don't fit error patterns of humans,
in ways we won't expect and be prepared for.
It'd be lousy not to get a job one is qualified for,
but it would triple suck if it was because of stack overflow
in some subroutine.

In May of 2010,
a flash crash on Wall Street fueled by a feedback loop
in Wall Street's "sell" algorithm
wiped a trillion dollars of value in 36 minutes.
I don't even want to think what "error" means
in the context of lethal autonomous weapons.
So yes, humans have always made biases.
Decision makers and gatekeepers,
in courts, in news, in war ...
they make mistakes; but that's exactly my point.
We cannot escape these difficult questions.
We cannot outsource our responsibilities to machines.

Artificial intelligence does not give us a "Get out of ethics free" card.
Data scientist Fred Benenson calls this math-washing.
We need the opposite.
We need to cultivate algorithm suspicion, scrutiny and investigation.
We need to make sure we have algorithmic accountability,
auditing and meaningful transparency.
We need to accept that bringing math and computation
to messy, value-laden human affairs
does not bring objectivity;
rather, the complexity of human affairs invades the algorithms.
Yes, we can and we should use computation
to help us make better decisions.
But we have to own up to our moral responsibility to judgment,
and use algorithms within that framework,
not as a means to abdicate and outsource our responsibilities
to one another as human to human.
Machine intelligence is here.
That means we must hold on ever tighter
to human values and human ethics.
Thank you.

The wonderful and terrifying implications of computers that can learn

Jeremy Howard

It used to be that if you wanted to get a computer to do something new,
you would have to program it.
Now, programming, for those of you here that haven't done it yourself,
requires laying out in excruciating detail
every single step that you want the computer to do
in order to achieve your goal.
Now, if you want to do something that you don't know how to do yourself,
then this is going to be a great challenge.
So this was the challenge faced by this man, Arthur Samuel.
In 1956, he wanted to get this computer
to be able to beat him at checkers.
How can you write a program,
lay out in excruciating detail, how to be better than you at checkers?
So he came up with an idea:
he had the computer play against itself thousands of times
and learn how to play checkers.
And indeed it worked, and in fact, by 1962,
this computer had beaten the Connecticut state champion.
So Arthur Samuel was the father of machine learning,
and I have a great debt to him,
because I am a machine learning practitioner.
I was the president of Kaggle,
a community of over 200,000 machine learning practictioners.
Kaggle puts up competitions
to try and get them to solve previously unsolved problems,
and it's been successful hundreds of times.
So from this vantage point, I was able to find out
a lot about what machine learning can do in the past, can do today,
and what it could do in the future.
Perhaps the first big success of machine learning commercially was Google.
Google showed that it is possible to find information
by using a computer algorithm,
and this algorithm is based on machine learning.
Since that time, there have been many commercial successes of machine learning.
Companies like Amazon and Netflix
use machine learning to suggest products that you might like to buy,
movies that you might like to watch.
Sometimes, it's almost creepy.
Companies like LinkedIn and Facebook
sometimes will tell you about who your friends might be
and you have no idea how it did it,
and this is because it's using the power of machine learning.
These are algorithms that have learned how to do this from data
rather than being programmed by hand.
This is also how IBM was successful
in getting Watson to beat the two world champions at "Jeopardy,"

answering incredibly subtle and complex questions like this one.
["The ancient 'Lion of Nimrud' went missing from this city's national museum in 2003 (along with a lot of other stuff)"]
This is also why we are now able to see the first self-driving cars.
If you want to be able to tell the difference between, say,
a tree and a pedestrian, well, that's pretty important.
We don't know how to write those programs by hand,
but with machine learning, this is now possible.
And in fact, this car has driven over a million miles
without any accidents on regular roads.
So we now know that computers can learn,
and computers can learn to do things
that we actually sometimes don't know how to do ourselves,
or maybe can do them better than us.
One of the most amazing examples I've seen of machine learning
happened on a project that I ran at Kaggle
where a team run by a guy called Geoffrey Hinton
from the University of Toronto
won a competition for automatic drug discovery.
Now, what was extraordinary here is not just that they beat
all of the algorithms developed by Merck or the international academic community,
but nobody on the team had any background in chemistry or biology or life sciences,
and they did it in two weeks.
How did they do this?
They used an extraordinary algorithm called deep learning.
So important was this that in fact the success was covered
in The New York Times in a front page article a few weeks later.
This is Geoffrey Hinton here on the left-hand side.
Deep learning is an algorithm inspired by how the human brain works,
and as a result it's an algorithm
which has no theoretical limitations on what it can do.
The more data you give it and the more computation time you give it,
the better it gets.
The New York Times also showed in this article
another extraordinary result of deep learning
which I'm going to show you now.
It shows that computers can listen and understand.
(Video) Richard Rashid: Now, the last step
that I want to be able to take in this process
is to actually speak to you in Chinese.
Now the key thing there is,
we've been able to take a large amount of information from many Chinese speakers
and produce a text-to-speech system
that takes Chinese text and converts it into Chinese language,
and then we've taken an hour or so of my own voice
and we've used that to modulate
the standard text-to-speech system so that it would sound like me.
Again, the result's not perfect.
There are in fact quite a few errors.
(In Chinese)

There's much work to be done in this area.
(In Chinese)

Jeremy Howard: Well, that was at a machine learning conference in China.
It's not often, actually, at academic conferences
that you do hear spontaneous applause,
although of course sometimes at TEDx conferences, feel free.
Everything you saw there was happening with deep learning.
Thank you.
The transcription in English was deep learning.
The translation to Chinese and the text in the top right, deep learning,
and the construction of the voice was deep learning as well.
So deep learning is this extraordinary thing.
It's a single algorithm that can seem to do almost anything,
and I discovered that a year earlier, it had also learned to see.
In this obscure competition from Germany
called the German Traffic Sign Recognition Benchmark,
deep learning had learned to recognize traffic signs like this one.
Not only could it recognize the traffic signs
better than any other algorithm,
the leaderboard actually showed it was better than people,
about twice as good as people.
So by 2011, we had the first example
of computers that can see better than people.
Since that time, a lot has happened.
In 2012, Google announced that they had a deep learning algorithm
watch YouTube videos
and crunched the data on 16,000 computers for a month,
and the computer independently learned about concepts such as people and cats
just by watching the videos.
This is much like the way that humans learn.
Humans don't learn by being told what they see,
but by learning for themselves what these things are.
Also in 2012, Geoffrey Hinton, who we saw earlier,
won the very popular ImageNet competition,
looking to try to figure out from one and a half million images
what they're pictures of.
As of 2014, we're now down to a six percent error rate
in image recognition.
This is better than people, again.
So machines really are doing an extraordinarily good job of this,
and it is now being used in industry.
For example, Google announced last year
that they had mapped every single location in France in two hours,
and the way they did it was that they fed street view images
into a deep learning algorithm to recognize and read street numbers.
Imagine how long it would have taken before:
dozens of people, many years.
This is also happening in China.
Baidu is kind of the Chinese Google, I guess,
and what you see here in the top left

is an example of a picture that I uploaded to Baidu's deep learning system,
and underneath you can see that the system has understood what that picture is
and found similar images.
The similar images actually have similar backgrounds,
similar directions of the faces,
even some with their tongue out.
This is not clearly looking at the text of a web page.
All I uploaded was an image.
So we now have computers which really understand what they see
and can therefore search databases
of hundreds of millions of images in real time.
So what does it mean now that computers can see?
Well, it's not just that computers can see.
In fact, deep learning has done more than that.
Complex, nuanced sentences like this one
are now understandable with deep learning algorithms.
As you can see here,
this Stanford-based system showing the red dot at the top
has figured out that this sentence is expressing negative sentiment.
Deep learning now in fact is near human performance
at understanding what sentences are about and what it is saying about those things.
Also, deep learning has been used to read Chinese,
again at about native Chinese speaker level.
This algorithm developed out of Switzerland
by people, none of whom speak or understand any Chinese.
As I say, using deep learning
is about the best system in the world for this,
even compared to native human understanding.
This is a system that we put together at my company
which shows putting all this stuff together.
These are pictures which have no text attached,
and as I'm typing in here sentences,
in real time it's understanding these pictures
and figuring out what they're about
and finding pictures that are similar to the text that I'm writing.
So you can see, it's actually understanding my sentences
and actually understanding these pictures.
I know that you've seen something like this on Google,
where you can type in things and it will show you pictures,
but actually what it's doing is it's searching the webpage for the text.
This is very different from actually understanding the images.
This is something that computers have only been able to do
for the first time in the last few months.
So we can see now that computers can not only see but they can also read,
and, of course, we've shown that they can understand what they hear.
Perhaps not surprising now that I'm going to tell you they can write.
Here is some text that I generated using a deep learning algorithm yesterday.
And here is some text that an algorithm out of Stanford generated.
Each of these sentences was generated
by a deep learning algorithm to describe each of those pictures.
This algorithm before has never seen a man in a black shirt playing a guitar.

It's seen a man before, it's seen black before,
it's seen a guitar before,
but it has independently generated this novel description of this picture.
We're still not quite at human performance here, but we're close.
In tests, humans prefer the computer-generated caption
one out of four times.
Now this system is now only two weeks old,
so probably within the next year,
the computer algorithm will be well past human performance
at the rate things are going.
So computers can also write.
So we put all this together and it leads to very exciting opportunities.
For example, in medicine,
a team in Boston announced that they had discovered
dozens of new clinically relevant features
of tumors which help doctors make a prognosis of a cancer.
Very similarly, in Stanford,
a group there announced that, looking at tissues under magnification,
they've developed a machine learning-based system
which in fact is better than human pathologists
at predicting survival rates for cancer sufferers.
In both of these cases, not only were the predictions more accurate,
but they generated new insightful science.
In the radiology case,
they were new clinical indicators that humans can understand.
In this pathology case,
the computer system actually discovered that the cells around the cancer
are as important as the cancer cells themselves
in making a diagnosis.
This is the opposite of what pathologists had been taught for decades.
In each of those two cases, they were systems developed
by a combination of medical experts and machine learning experts,
but as of last year, we're now beyond that too.
This is an example of identifying cancerous areas
of human tissue under a microscope.
The system being shown here can identify those areas more accurately,
or about as accurately, as human pathologists,
but was built entirely with deep learning using no medical expertise
by people who have no background in the field.
Similarly, here, this neuron segmentation.
We can now segment neurons about as accurately as humans can,
but this system was developed with deep learning
using people with no previous background in medicine.
So myself, as somebody with no previous background in medicine,
I seem to be entirely well qualified to start a new medical company,
which I did.
I was kind of terrified of doing it,
but the theory seemed to suggest that it ought to be possible
to do very useful medicine using just these data analytic techniques.
And thankfully, the feedback has been fantastic,
not just from the media but from the medical community,

who have been very supportive.
The theory is that we can take the middle part of the medical process
and turn that into data analysis as much as possible,
leaving doctors to do what they're best at.
I want to give you an example.
It now takes us about 15 minutes to generate a new medical diagnostic test
and I'll show you that in real time now,
but I've compressed it down to three minutes by cutting some pieces out.
Rather than showing you creating a medical diagnostic test,
I'm going to show you a diagnostic test of car images,
because that's something we can all understand.
So here we're starting with about 1.5 million car images,
and I want to create something that can split them into the angle
of the photo that's being taken.
So these images are entirely unlabeled, so I have to start from scratch.
With our deep learning algorithm,
it can automatically identify areas of structure in these images.
So the nice thing is that the human and the computer can now work together.
So the human, as you can see here,
is telling the computer about areas of interest
which it wants the computer then to try and use to improve its algorithm.
Now, these deep learning systems actually are in 16,000-dimensional space,
so you can see here the computer rotating this through that space,
trying to find new areas of structure.
And when it does so successfully,
the human who is driving it can then point out the areas that are interesting.
So here, the computer has successfully found areas,
for example, angles.
So as we go through this process,
we're gradually telling the computer more and more
about the kinds of structures we're looking for.
You can imagine in a diagnostic test
this would be a pathologist identifying areas of pathosis, for example,
or a radiologist indicating potentially troublesome nodules.
And sometimes it can be difficult for the algorithm.
In this case, it got kind of confused.
The fronts and the backs of the cars are all mixed up.
So here we have to be a bit more careful,
manually selecting these fronts as opposed to the backs,
then telling the computer that this is a type of group
that we're interested in.
So we do that for a while, we skip over a little bit,
and then we train the machine learning algorithm
based on these couple of hundred things,
and we hope that it's gotten a lot better.
You can see, it's now started to fade some of these pictures out,
showing us that it already is recognizing how to understand some of these itself.
We can then use this concept of similar images,
and using similar images, you can now see,
the computer at this point is able to entirely find just the fronts of cars.
So at this point, the human can tell the computer,

okay, yes, you've done a good job of that.
Sometimes, of course, even at this point
it's still difficult to separate out groups.
In this case, even after we let the computer try to rotate this for a while,
we still find that the left sides and the right sides pictures
are all mixed up together.
So we can again give the computer some hints,
and we say, okay, try and find a projection that separates out
the left sides and the right sides as much as possible
using this deep learning algorithm.
And giving it that hint -- ah, okay, it's been successful.
It's managed to find a way of thinking about these objects
that's separated out these together.
So you get the idea here.
This is a case not where the human is being replaced by a computer,
but where they're working together.
What we're doing here is we're replacing something that used to take a team
of five or six people about seven years
and replacing it with something that takes 15 minutes
for one person acting alone.
So this process takes about four or five iterations.
You can see we now have 62 percent
of our 1.5 million images classified correctly.
And at this point, we can start to quite quickly
grab whole big sections,
check through them to make sure that there's no mistakes.
Where there are mistakes, we can let the computer know about them.
And using this kind of process for each of the different groups,
we are now up to an 80 percent success rate
in classifying the 1.5 million images.
And at this point, it's just a case
of finding the small number that aren't classified correctly,
and trying to understand why.
And using that approach,
by 15 minutes we get to 97 percent classification rates.
So this kind of technique could allow us to fix a major problem,
which is that there's a lack of medical expertise in the world.
The World Economic Forum says that there's between a 10x and a 20x
shortage of physicians in the developing world,
and it would take about 300 years
to train enough people to fix that problem.
So imagine if we can help enhance their efficiency
using these deep learning approaches?
So I'm very excited about the opportunities.
I'm also concerned about the problems.
The problem here is that every area in blue on this map
is somewhere where services are over 80 percent of employment.
What are services?
These are services.
These are also the exact things that computers have just learned how to do.
So 80 percent of the world's employment in the developed world

is stuff that computers have just learned how to do.
What does that mean?
Well, it'll be fine. They'll be replaced by other jobs.
For example, there will be more jobs for data scientists.
Well, not really.
It doesn't take data scientists very long to build these things.
For example, these four algorithms were all built by the same guy.
So if you think, oh, it's all happened before,
we've seen the results in the past of when new things come along
and they get replaced by new jobs,
what are these new jobs going to be?
It's very hard for us to estimate this,
because human performance grows at this gradual rate,
but we now have a system, deep learning,
that we know actually grows in capability exponentially.
And we're here.
So currently, we see the things around us
and we say, "Oh, computers are still pretty dumb." Right?
But in five years' time, computers will be off this chart.
So we need to be starting to think about this capability right now.
We have seen this once before, of course.
In the Industrial Revolution,
we saw a step change in capability thanks to engines.
The thing is, though, that after a while, things flattened out.
There was social disruption,
but once engines were used to generate power in all the situations,
things really settled down.
The Machine Learning Revolution
is going to be very different from the Industrial Revolution,
because the Machine Learning Revolution, it never settles down.
The better computers get at intellectual activities,
the more they can build better computers to be better at intellectual capabilities,
so this is going to be a kind of change
that the world has actually never experienced before,
so your previous understanding of what's possible is different.
This is already impacting us.
In the last 25 years, as capital productivity has increased,
labor productivity has been flat, in fact even a little bit down.
So I want us to start having this discussion now.
I know that when I often tell people about this situation,
people can be quite dismissive.
Well, computers can't really think,
they don't emote, they don't understand poetry,
we don't really understand how they work.
So what?
Computers right now can do the things
that humans spend most of their time being paid to do,
so now's the time to start thinking
about how we're going to adjust our social structures and economic structures
to be aware of this new reality.
Thank you.

How we're teaching computers to understand pictures

Fei-Fei Li

Let me show you something.
(Video) Girl: Okay, that's a cat sitting in a bed.
The boy is petting the elephant.
Those are people that are going on an airplane.
That's a big airplane.
Fei-Fei Li: This is a three-year-old child
describing what she sees in a series of photos.
She might still have a lot to learn about this world,
but she's already an expert at one very important task:
to make sense of what she sees.
Our society is more technologically advanced than ever.
We send people to the moon, we make phones that talk to us
or customize radio stations that can play only music we like.
Yet, our most advanced machines and computers
still struggle at this task.
So I'm here today to give you a progress report
on the latest advances in our research in computer vision,
one of the most frontier and potentially revolutionary
technologies in computer science.
Yes, we have prototyped cars that can drive by themselves,
but without smart vision, they cannot really tell the difference
between a crumpled paper bag on the road, which can be run over,
and a rock that size, which should be avoided.
We have made fabulous megapixel cameras,
but we have not delivered sight to the blind.
Drones can fly over massive land,
but don't have enough vision technology
to help us to track the changes of the rainforests.
Security cameras are everywhere,
but they do not alert us when a child is drowning in a swimming pool.
Photos and videos are becoming an integral part of global life.
They're being generated at a pace that's far beyond what any human,
or teams of humans, could hope to view,
and you and I are contributing to that at this TED.
Yet our most advanced software is still struggling at understanding
and managing this enormous content.
So in other words, collectively as a society,
we're very much blind,
because our smartest machines are still blind.
"Why is this so hard?" you may ask.
Cameras can take pictures like this one
by converting lights into a two-dimensional array of numbers
known as pixels,
but these are just lifeless numbers.
They do not carry meaning in themselves.
Just like to hear is not the same as to listen,
to take pictures is not the same as to see,
and by seeing, we really mean understanding.

In fact, it took Mother Nature 540 million years of hard work
to do this task,
and much of that effort
went into developing the visual processing apparatus of our brains,
not the eyes themselves.
So vision begins with the eyes,
but it truly takes place in the brain.
So for 15 years now, starting from my Ph.D. at Caltech
and then leading Stanford's Vision Lab,
I've been working with my mentors, collaborators and students
to teach computers to see.
Our research field is called computer vision and machine learning.
It's part of the general field of artificial intelligence.
So ultimately, we want to teach the machines to see just like we do:
naming objects, identifying people, inferring 3D geometry of things,
understanding relations, emotions, actions and intentions.
You and I weave together entire stories of people, places and things
the moment we lay our gaze on them.
The first step towards this goal is to teach a computer to see objects,
the building block of the visual world.
In its simplest terms, imagine this teaching process
as showing the computers some training images
of a particular object, let's say cats,
and designing a model that learns from these training images.
How hard can this be?
After all, a cat is just a collection of shapes and colors,
and this is what we did in the early days of object modeling.
We'd tell the computer algorithm in a mathematical language
that a cat has a round face, a chubby body,
two pointy ears, and a long tail,
and that looked all fine.
But what about this cat?

It's all curled up.
Now you have to add another shape and viewpoint to the object model.
But what if cats are hidden?
What about these silly cats?
Now you get my point.
Even something as simple as a household pet
can present an infinite number of variations to the object model,
and that's just one object.
So about eight years ago,
a very simple and profound observation changed my thinking.
No one tells a child how to see,
especially in the early years.
They learn this through real-world experiences and examples.
If you consider a child's eyes
as a pair of biological cameras,
they take one picture about every 200 milliseconds,
the average time an eye movement is made.
So by age three, a child would have seen hundreds of millions of pictures

of the real world.
That's a lot of training examples.
So instead of focusing solely on better and better algorithms,
my insight was to give the algorithms the kind of training data
that a child was given through experiences
in both quantity and quality.
Once we know this,
we knew we needed to collect a data set
that has far more images than we have ever had before,
perhaps thousands of times more,
and together with Professor Kai Li at Princeton University,
we launched the ImageNet project in 2007.
Luckily, we didn't have to mount a camera on our head
and wait for many years.
We went to the Internet,
the biggest treasure trove of pictures that humans have ever created.
We downloaded nearly a billion images
and used crowdsourcing technology like the Amazon Mechanical Turk platform
to help us to label these images.
At its peak, ImageNet was one of the biggest employers
of the Amazon Mechanical Turk workers:
together, almost 50,000 workers
from 167 countries around the world
helped us to clean, sort and label
nearly a billion candidate images.
That was how much effort it took
to capture even a fraction of the imagery
a child's mind takes in in the early developmental years.
In hindsight, this idea of using big data
to train computer algorithms may seem obvious now,
but back in 2007, it was not so obvious.
We were fairly alone on this journey for quite a while.
Some very friendly colleagues advised me to do something more useful for my tenure,
and we were constantly struggling for research funding.
Once, I even joked to my graduate students
that I would just reopen my dry cleaner's shop to fund ImageNet.
After all, that's how I funded my college years.
So we carried on.
In 2009, the ImageNet project delivered
a database of 15 million images
across 22,000 classes of objects and things
organized by everyday English words.
In both quantity and quality,
this was an unprecedented scale.
As an example, in the case of cats,
we have more than 62,000 cats
of all kinds of looks and poses
and across all species of domestic and wild cats.
We were thrilled to have put together ImageNet,
and we wanted the whole research world to benefit from it,
so in the TED fashion, we opened up the entire data set

to the worldwide research community for free.

Now that we have the data to nourish our computer brain,
we're ready to come back to the algorithms themselves.
As it turned out, the wealth of information provided by ImageNet
was a perfect match to a particular class of machine learning algorithms
called convolutional neural network,
pioneered by Kunihiko Fukushima, Geoff Hinton, and Yann LeCun
back in the 1970s and '80s.
Just like the brain consists of billions of highly connected neurons,
a basic operating unit in a neural network
is a neuron-like node.
It takes input from other nodes
and sends output to others.
Moreover, these hundreds of thousands or even millions of nodes
are organized in hierarchical layers,
also similar to the brain.
In a typical neural network we use to train our object recognition model,
it has 24 million nodes,
140 million parameters,
and 15 billion connections.
That's an enormous model.
Powered by the massive data from ImageNet
and the modern CPUs and GPUs to train such a humongous model,
the convolutional neural network
blossomed in a way that no one expected.
It became the winning architecture
to generate exciting new results in object recognition.
This is a computer telling us
this picture contains a cat
and where the cat is.
Of course there are more things than cats,
so here's a computer algorithm telling us
the picture contains a boy and a teddy bear;
a dog, a person, and a small kite in the background;
or a picture of very busy things
like a man, a skateboard, railings, a lampost, and so on.
Sometimes, when the computer is not so confident about what it sees,
we have taught it to be smart enough
to give us a safe answer instead of committing too much,
just like we would do,
but other times our computer algorithm is remarkable at telling us
what exactly the objects are,
like the make, model, year of the cars.
We applied this algorithm to millions of Google Street View images
across hundreds of American cities,
and we have learned something really interesting:
first, it confirmed our common wisdom
that car prices correlate very well
with household incomes.
But surprisingly, car prices also correlate well

with crime rates in cities,
or voting patterns by zip codes.
So wait a minute. Is that it?
Has the computer already matched or even surpassed human capabilities?
Not so fast.
So far, we have just taught the computer to see objects.
This is like a small child learning to utter a few nouns.
It's an incredible accomplishment,
but it's only the first step.
Soon, another developmental milestone will be hit,
and children begin to communicate in sentences.
So instead of saying this is a cat in the picture,
you already heard the little girl telling us this is a cat lying on a bed.
So to teach a computer to see a picture and generate sentences,
the marriage between big data and machine learning algorithm
has to take another step.
Now, the computer has to learn from both pictures
as well as natural language sentences
generated by humans.
Just like the brain integrates vision and language,
we developed a model that connects parts of visual things
like visual snippets
with words and phrases in sentences.
About four months ago,
we finally tied all this together
and produced one of the first computer vision models
that is capable of generating a human-like sentence
when it sees a picture for the first time.
Now, I'm ready to show you what the computer says
when it sees the picture
that the little girl saw at the beginning of this talk.
(Video) Computer: A man is standing next to an elephant.
A large airplane sitting on top of an airport runway.
FFL: Of course, we're still working hard to improve our algorithms,
and it still has a lot to learn.

And the computer still makes mistakes.
(Video) Computer: A cat lying on a bed in a blanket.
FFL: So of course, when it sees too many cats,
it thinks everything might look like a cat.
(Video) Computer: A young boy is holding a baseball bat.

FFL: Or, if it hasn't seen a toothbrush, it confuses it with a baseball bat.
(Video) Computer: A man riding a horse down a street next to a building.

FFL: We haven't taught Art 101 to the computers.
(Video) Computer: A zebra standing in a field of grass.
FFL: And it hasn't learned to appreciate the stunning beauty of nature
like you and I do.
So it has been a long journey.
To get from age zero to three was hard.

The real challenge is to go from three to 13 and far beyond.
Let me remind you with this picture of the boy and the cake again.
So far, we have taught the computer to see objects
or even tell us a simple story when seeing a picture.
(Video) Computer: A person sitting at a table with a cake.
FFL: But there's so much more to this picture
than just a person and a cake.
What the computer doesn't see is that this is a special Italian cake
that's only served during Easter time.
The boy is wearing his favorite t-shirt
given to him as a gift by his father after a trip to Sydney,
and you and I can all tell how happy he is
and what's exactly on his mind at that moment.
This is my son Leo.
On my quest for visual intelligence,
I think of Leo constantly
and the future world he will live in.
When machines can see,
doctors and nurses will have extra pairs of tireless eyes
to help them to diagnose and take care of patients.
Cars will run smarter and safer on the road.
Robots, not just humans,
will help us to brave the disaster zones to save the trapped and wounded.
We will discover new species, better materials,
and explore unseen frontiers with the help of the machines.
Little by little, we're giving sight to the machines.
First, we teach them to see.
Then, they help us to see better.
For the first time, human eyes won't be the only ones
pondering and exploring our world.
We will not only use the machines for their intelligence,
we will also collaborate with them in ways that we cannot even imagine.
This is my quest:
to give computers visual intelligence
and to create a better future for Leo and for the world.
Thank you.

What happens when our computers get smarter than we are?

Nick Bostrom

I work with a bunch of mathematicians, philosophers and computer scientists,
and we sit around and think about the future of machine intelligence,
among other things.
Some people think that some of these things are sort of science fiction-y,
far out there, crazy.
But I like to say,
okay, let's look at the modern human condition.

This is the normal way for things to be.
But if we think about it,
we are actually recently arrived guests on this planet,
the human species.
Think about if Earth was created one year ago,
the human species, then, would be 10 minutes old.
The industrial era started two seconds ago.
Another way to look at this is to think of world GDP over the last 10,000 years,
I've actually taken the trouble to plot this for you in a graph.
It looks like this.

It's a curious shape for a normal condition.
I sure wouldn't want to sit on it.

Let's ask ourselves, what is the cause of this current anomaly?
Some people would say it's technology.
Now it's true, technology has accumulated through human history,
and right now, technology advances extremely rapidly --
that is the proximate cause,
that's why we are currently so very productive.
But I like to think back further to the ultimate cause.
Look at these two highly distinguished gentlemen:
We have Kanzi --
he's mastered 200 lexical tokens, an incredible feat.
And Ed Witten unleashed the second superstring revolution.
If we look under the hood, this is what we find:
basically the same thing.
One is a little larger,
it maybe also has a few tricks in the exact way it's wired.
These invisible differences cannot be too complicated, however,
because there have only been 250,000 generations
since our last common ancestor.
We know that complicated mechanisms take a long time to evolve.
So a bunch of relatively minor changes
take us from Kanzi to Witten,
from broken-off tree branches to intercontinental ballistic missiles.
So this then seems pretty obvious that everything we've achieved,
and everything we care about,
depends crucially on some relatively minor changes that made the human mind.
And the corollary, of course, is that any further changes

that could significantly change the substrate of thinking
could have potentially enormous consequences.
Some of my colleagues think we're on the verge
of something that could cause a profound change in that substrate,
and that is machine superintelligence.
Artificial intelligence used to be about putting commands in a box.
You would have human programmers
that would painstakingly handcraft knowledge items.
You build up these expert systems,
and they were kind of useful for some purposes,
but they were very brittle, you couldn't scale them.
Basically, you got out only what you put in.
But since then,
a paradigm shift has taken place in the field of artificial intelligence.
Today, the action is really around machine learning.
So rather than handcrafting knowledge representations and features,
we create algorithms that learn, often from raw perceptual data.
Basically the same thing that the human infant does.
The result is A.I. that is not limited to one domain --
the same system can learn to translate between any pairs of languages,
or learn to play any computer game on the Atari console.
Now of course,
A.I. is still nowhere near having the same powerful, cross-domain
ability to learn and plan as a human being has.
The cortex still has some algorithmic tricks
that we don't yet know how to match in machines.
So the question is,
how far are we from being able to match those tricks?
A couple of years ago,
we did a survey of some of the world's leading A.I. experts,
to see what they think, and one of the questions we asked was,
"By which year do you think there is a 50 percent probability
that we will have achieved human-level machine intelligence?"
We defined human-level here as the ability to perform
almost any job at least as well as an adult human,
so real human-level, not just within some limited domain.
And the median answer was 2040 or 2050,
depending on precisely which group of experts we asked.
Now, it could happen much, much later, or sooner,
the truth is nobody really knows.
What we do know is that the ultimate limit to information processing
in a machine substrate lies far outside the limits in biological tissue.
This comes down to physics.
A biological neuron fires, maybe, at 200 hertz, 200 times a second.
But even a present-day transistor operates at the Gigahertz.
Neurons propagate slowly in axons, 100 meters per second, tops.
But in computers, signals can travel at the speed of light.
There are also size limitations,
like a human brain has to fit inside a cranium,
but a computer can be the size of a warehouse or larger.
So the potential for superintelligence lies dormant in matter,

much like the power of the atom lay dormant throughout human history,
patiently waiting there until 1945.
In this century,
scientists may learn to awaken the power of artificial intelligence.
And I think we might then see an intelligence explosion.
Now most people, when they think about what is smart and what is dumb,
I think have in mind a picture roughly like this.
So at one end we have the village idiot,
and then far over at the other side
we have Ed Witten, or Albert Einstein, or whoever your favorite guru is.
But I think that from the point of view of artificial intelligence,
the true picture is actually probably more like this:
AI starts out at this point here, at zero intelligence,
and then, after many, many years of really hard work,
maybe eventually we get to mouse-level artificial intelligence,
something that can navigate cluttered environments
as well as a mouse can.
And then, after many, many more years of really hard work, lots of investment,
maybe eventually we get to chimpanzee-level artificial intelligence.
And then, after even more years of really, really hard work,
we get to village idiot artificial intelligence.
And a few moments later, we are beyond Ed Witten.
The train doesn't stop at Humanville Station.
It's likely, rather, to swoosh right by.
Now this has profound implications,
particularly when it comes to questions of power.
For example, chimpanzees are strong --
pound for pound, a chimpanzee is about twice as strong as a fit human male.
And yet, the fate of Kanzi and his pals depends a lot more
on what we humans do than on what the chimpanzees do themselves.
Once there is superintelligence,
the fate of humanity may depend on what the superintelligence does.
Think about it:
Machine intelligence is the last invention that humanity will ever need to make.
Machines will then be better at inventing than we are,
and they'll be doing so on digital timescales.
What this means is basically a telescoping of the future.
Think of all the crazy technologies that you could have imagined
maybe humans could have developed in the fullness of time:
cures for aging, space colonization,
self-replicating nanobots or uploading of minds into computers,
all kinds of science fiction-y stuff
that's nevertheless consistent with the laws of physics.
All of this superintelligence could develop, and possibly quite rapidly.
Now, a superintelligence with such technological maturity
would be extremely powerful,
and at least in some scenarios, it would be able to get what it wants.
We would then have a future that would be shaped by the preferences of this A.I.
Now a good question is, what are those preferences?
Here it gets trickier.
To make any headway with this,

we must first of all avoid anthropomorphizing.
And this is ironic because every newspaper article
about the future of A.I. has a picture of this:
So I think what we need to do is to conceive of the issue more abstractly,
not in terms of vivid Hollywood scenarios.
We need to think of intelligence as an optimization process,
a process that steers the future into a particular set of configurations.
A superintelligence is a really strong optimization process.
It's extremely good at using available means to achieve a state
in which its goal is realized.
This means that there is no necessary connection between
being highly intelligent in this sense,
and having an objective that we humans would find worthwhile or meaningful.
Suppose we give an A.I. the goal to make humans smile.
When the A.I. is weak, it performs useful or amusing actions
that cause its user to smile.
When the A.I. becomes superintelligent,
it realizes that there is a more effective way to achieve this goal:
take control of the world
and stick electrodes into the facial muscles of humans
to cause constant, beaming grins.
Another example,
suppose we give A.I. the goal to solve a difficult mathematical problem.
When the A.I. becomes superintelligent,
it realizes that the most effective way to get the solution to this problem
is by transforming the planet into a giant computer,
so as to increase its thinking capacity.
And notice that this gives the A.I.s an instrumental reason
to do things to us that we might not approve of.
Human beings in this model are threats,
we could prevent the mathematical problem from being solved.
Of course, perceivably things won't go wrong in these particular ways;
these are cartoon examples.
But the general point here is important:
if you create a really powerful optimization process
to maximize for objective x,
you better make sure that your definition of x
incorporates everything you care about.
This is a lesson that's also taught in many a myth.
King Midas wishes that everything he touches be turned into gold.
He touches his daughter, she turns into gold.
He touches his food, it turns into gold.
This could become practically relevant,
not just as a metaphor for greed,
but as an illustration of what happens
if you create a powerful optimization process
and give it misconceived or poorly specified goals.
Now you might say, if a computer starts sticking electrodes into people's faces,
we'd just shut it off.
A, this is not necessarily so easy to do if we've grown dependent on the system --
like, where is the off switch to the Internet?

B, why haven't the chimpanzees flicked the off switch to humanity,
or the Neanderthals?
They certainly had reasons.
We have an off switch, for example, right here.
(Choking)
The reason is that we are an intelligent adversary;
we can anticipate threats and plan around them.
But so could a superintelligent agent,
and it would be much better at that than we are.
The point is, we should not be confident that we have this under control here.
And we could try to make our job a little bit easier by, say,
putting the A.I. in a box,
like a secure software environment,
a virtual reality simulation from which it cannot escape.
But how confident can we be that the A.I. couldn't find a bug.
Given that merely human hackers find bugs all the time,
I'd say, probably not very confident.
So we disconnect the ethernet cable to create an air gap,
but again, like merely human hackers
routinely transgress air gaps using social engineering.
Right now, as I speak,
I'm sure there is some employee out there somewhere
who has been talked into handing out her account details
by somebody claiming to be from the I.T. department.
More creative scenarios are also possible,
like if you're the A.I.,
you can imagine wiggling electrodes around in your internal circuitry
to create radio waves that you can use to communicate.
Or maybe you could pretend to malfunction,
and then when the programmers open you up to see what went wrong with you,
they look at the source code -- Bam! --
the manipulation can take place.
Or it could output the blueprint to a really nifty technology,
and when we implement it,
it has some surreptitious side effect that the A.I. had planned.
The point here is that we should not be confident in our ability
to keep a superintelligent genie locked up in its bottle forever.
Sooner or later, it will out.
I believe that the answer here is to figure out
how to create superintelligent A.I. such that even if -- when -- it escapes,
it is still safe because it is fundamentally on our side
because it shares our values.
I see no way around this difficult problem.
Now, I'm actually fairly optimistic that this problem can be solved.
We wouldn't have to write down a long list of everything we care about,
or worse yet, spell it out in some computer language
like C++ or Python,
that would be a task beyond hopeless.
Instead, we would create an A.I. that uses its intelligence
to learn what we value,
and its motivation system is constructed in such a way that it is motivated

to pursue our values or to perform actions that it predicts we would approve of.
We would thus leverage its intelligence as much as possible
to solve the problem of value-loading.
This can happen,
and the outcome could be very good for humanity.
But it doesn't happen automatically.
The initial conditions for the intelligence explosion
might need to be set up in just the right way
if we are to have a controlled detonation.
The values that the A.I. has need to match ours,
not just in the familiar context,
like where we can easily check how the A.I. behaves,
but also in all novel contexts that the A.I. might encounter
in the indefinite future.
And there are also some esoteric issues that would need to be solved, sorted out:
the exact details of its decision theory,
how to deal with logical uncertainty and so forth.
So the technical problems that need to be solved to make this work
look quite difficult --
not as difficult as making a superintelligent A.I.,
but fairly difficult.
Here is the worry:
Making superintelligent A.I. is a really hard challenge.
Making superintelligent A.I. that is safe
involves some additional challenge on top of that.
The risk is that if somebody figures out how to crack the first challenge
without also having cracked the additional challenge
of ensuring perfect safety.
So I think that we should work out a solution
to the control problem in advance,
so that we have it available by the time it is needed.
Now it might be that we cannot solve the entire control problem in advance
because maybe some elements can only be put in place
once you know the details of the architecture where it will be implemented.
But the more of the control problem that we solve in advance,
the better the odds that the transition to the machine intelligence era
will go well.
This to me looks like a thing that is well worth doing
and I can imagine that if things turn out okay,
that people a million years from now look back at this century
and it might well be that they say that the one thing we did that really mattered
was to get this thing right.
Thank you.

Don't fear intelligent machines. Work with them

Garry Kasparov

This story begins in 1985,
when at age 22,
I became the World Chess Champion
after beating Anatoly Karpov.
Earlier that year,
I played what is called simultaneous exhibition
against 32 of the world's best chess-playing machines
in Hamburg, Germany.
I won all the games,
and then it was not considered much of a surprise
that I could beat 32 computers at the same time.
To me, that was the golden age.

Machines were weak,
and my hair was strong.

Just 12 years later,
I was fighting for my life against just one computer
in a match
called by the cover of "Newsweek"
"The Brain's Last Stand."
No pressure.

From mythology to science fiction,
human versus machine
has been often portrayed as a matter of life and death.
John Henry,
called the steel-driving man
in the 19th century African American folk legend,
was pitted in a race
against a steam-powered hammer
bashing a tunnel through mountain rock.
John Henry's legend is a part of a long historical narrative
pitting humanity versus technology.
And this competitive rhetoric is standard now.
We are in a race against the machines,
in a fight or even in a war.
Jobs are being killed off.
People are being replaced as if they had vanished from the Earth.
It's enough to think that the movies like "The Terminator" or "The Matrix"
are nonfiction.
There are very few instances of an arena
where the human body and mind can compete on equal terms
with a computer or a robot.
Actually, I wish there were a few more.
Instead,
it was my blessing and my curse
to literally become the proverbial man

in the man versus machine competition
that everybody is still talking about.
In the most famous human-machine competition since John Henry,
I played two matches
against the IBM supercomputer, Deep Blue.
Nobody remembers that I won the first match --

In Philadelphia, before losing the rematch the following year in New York.
But I guess that's fair.
There is no day in history, special calendar entry
for all the people who failed to climb Mt. Everest
before Sir Edmund Hillary and Tenzing Norgay
made it to the top.
And in 1997, I was still the world champion
when chess computers finally came of age.
I was Mt. Everest,
and Deep Blue reached the summit.
I should say of course, not that Deep Blue did it,
but its human creators --
Anantharaman, Campbell, Hoane, Hsu.
Hats off to them.
As always, machine's triumph was a human triumph,
something we tend to forget when humans are surpassed by our own creations.
Deep Blue was victorious,
but was it intelligent?
No, no it wasn't,
at least not in the way Alan Turing and other founders of computer science
had hoped.
It turned out that chess could be crunched by brute force,
once hardware got fast enough
and algorithms got smart enough.
Although by the definition of the output,
grandmaster-level chess,
Deep Blue was intelligent.
But even at the incredible speed,
200 million positions per second,
Deep Blue's method
provided little of the dreamed-of insight into the mysteries of human intelligence.
Soon,
machines will be taxi drivers
and doctors and professors,
but will they be "intelligent?"
I would rather leave these definitions
to the philosophers and to the dictionary.
What really matters is how we humans
feel about living and working with these machines.
When I first met Deep Blue in 1996 in February,
I had been the world champion for more than 10 years,
and I had played 182 world championship games
and hundreds of games against other top players in other competitions.

I knew what to expect from my opponents
and what to expect from myself.
I was used to measure their moves
and to gauge their emotional state
by watching their body language and looking into their eyes.
And then I sat across the chessboard from Deep Blue.
I immediately sensed something new,
something unsettling.
You might experience a similar feeling
the first time you ride in a driverless car
or the first time your new computer manager issues an order at work.
But when I sat at that first game,
I couldn't be sure
what is this thing capable of.
Technology can advance in leaps, and IBM had invested heavily.
I lost that game.
And I couldn't help wondering,
might it be invincible?
Was my beloved game of chess over?
These were human doubts, human fears,
and the only thing I knew for sure
was that my opponent Deep Blue had no such worries at all.

I fought back
after this devastating blow
to win the first match,
but the writing was on the wall.
I eventually lost to the machine
but I didn't suffer the fate of John Henry
who won but died with his hammer in his hand.
[John Henry Died with a Hammer in His Hand Palmer C. Hayden]
[The Museum of African American Art, Los Angeles]
It turned out that the world of chess
still wanted to have a human chess champion.
And even today,
when a free chess app on the latest mobile phone
is stronger than Deep Blue,
people are still playing chess,
even more than ever before.
Doomsayers predicted that nobody would touch the game
that could be conquered by the machine,
and they were wrong, proven wrong,
but doomsaying has always been a popular pastime
when it comes to technology.
What I learned from my own experience
is that we must face our fears
if we want to get the most out of our technology,
and we must conquer those fears
if we want to get the best out of our humanity.
While licking my wounds,
I got a lot of inspiration

from my battles against Deep Blue.
As the old Russian saying goes, if you can't beat them, join them.
Then I thought,
what if I could play with a computer --
together with a computer at my side, combining our strengths,
human intuition plus machine's calculation,
human strategy, machine tactics,
human experience, machine's memory.
Could it be the perfect game ever played?
My idea came to life
in 1998 under the name of Advanced Chess
when I played this human-plus-machine competition against another elite player.
But in this first experiment,
we both failed to combine human and machine skills effectively.
Advanced Chess found its home on the internet,
and in 2005, a so-called freestyle chess tournament
produced a revelation.
A team of grandmasters and top machines participated,
but the winners were not grandmasters,
not a supercomputer.
The winners were a pair of amateur American chess players
operating three ordinary PCs at the same time.
Their skill of coaching their machines
effectively counteracted the superior chess knowledge
of their grandmaster opponents
and much greater computational power of others.
And I reached this formulation.
A weak human player plus a machine
plus a better process is superior
to a very powerful machine alone,
but more remarkably, is superior to a strong human player
plus machine
and an inferior process.
This convinced me that we would need
better interfaces to help us coach our machines
towards more useful intelligence.
Human plus machine isn't the future,
it's the present.
Everybody that's used online translation
to get the gist of a news article from a foreign newspaper,
knowing its far from perfect.
Then we use our human experience
to make sense out of that,
and then the machine learns from our corrections.
This model is spreading and investing in medical diagnosis, security analysis.
The machine crunches data,
calculates probabilities,
gets 80 percent of the way, 90 percent,
making it easier for analysis
and decision-making of the human party.
But you are not going to send your kids

to school in a self-driving car with 90 percent accuracy,
even with 99 percent.
So we need a leap forward
to add a few more crucial decimal places.
Twenty years after my match with Deep Blue,
second match,
this sensational "The Brain's Last Stand" headline
has become commonplace
as intelligent machines
move
in every sector, seemingly every day.
But unlike in the past,
when machines replaced
farm animals, manual labor,
now they are coming after people with college degrees
and political influence.
And as someone who fought machines and lost,
I am here to tell you this is excellent, excellent news.
Eventually, every profession
will have to feel these pressures
or else it will mean humanity has ceased to make progress.
We don't
get to choose
when and where technological progress stops.
We cannot
slow down.
In fact,
we have to speed up.
Our technology excels at removing
difficulties and uncertainties from our lives,
and so we must seek out
ever more difficult,
ever more uncertain challenges.
Machines have
calculations.
We have understanding.
Machines have instructions.
We have purpose.
Machines have
objectivity.
We have passion.
We should not worry about what our machines can do today.
Instead, we should worry about what they still cannot do today,
because we will need the help of the new, intelligent machines
to turn our grandest dreams into reality.
And if we fail,
if we fail, it's not because our machines are too intelligent,
or not intelligent enough.
If we fail, it's because we grew complacent
and limited our ambitions.
Our humanity is not defined by any skill,

like swinging a hammer or even playing chess.
There's one thing only a human can do.
That's dream.
So let us dream big.
Thank you.

Watson, Jeopardy and me, the obsolete know-it-all

Ken Jennings

In two weeks time, that's the ninth anniversary
of the day I first stepped out onto that hallowed "Jeopardy" set.
I mean, nine years is a long time.
And given "Jeopardy's" average demographics,
I think what that means
is most of the people who saw me on that show are now dead.

But not all, a few are still alive.
Occasionally I still get recognized at the mall or whatever.
And when I do, it's as a bit of a know-it-all.
I think that ship has sailed, it's too late for me.
For better or for worse, that's what I'm going to be known as,
as the guy who knew a lot of weird stuff.
And I can't complain about this.
I feel like that was always sort of my destiny,
although I had for many years been pretty deeply in the trivia closet.
If nothing else, you realize very quickly as a teenager,
it is not a hit with girls to know Captain Kirk's middle name.

And as a result, I was sort of the deeply closeted kind of know-it-all for many years.
But if you go further back, if you look at it, it's all there.
I was the kind of kid who was always bugging Mom and Dad
with whatever great fact I had just read about --
Haley's comet or giant squids
or the size of the world's biggest pumpkin pie or whatever it was.
I now have a 10-year-old of my own who's exactly the same.
And I know how deeply annoying it is, so karma does work.

And I loved game shows, fascinated with game shows.
I remember crying on my first day of kindergarten back in 1979
because it had just hit me, as badly as I wanted to go to school,
that I was also going to miss "Hollywood Squares" and "Family Feud."
I was going to miss my game shows.
And later, in the mid-'80s,
when "Jeopardy" came back on the air,
I remember running home from school every day to watch the show.
It was my favorite show, even before it paid for my house.
And we lived overseas, we lived in South Korea where my dad was working,
where there was only one English language TV channel.
There was Armed Forces TV,
and if you didn't speak Korean, that's what you were watching.
So me and all my friends would run home every day and watch "Jeopardy."
I was always that kind of obsessed trivia kid.
I remember being able to play Trivial Pursuit against my parents back in the '80s
and holding my own, back when that was a fad.
There's a weird sense of mastery you get
when you know some bit of boomer trivia that Mom and Dad don't know.
You know some Beatles factoid that Dad didn't know.

And you think, ah hah, knowledge really is power --
the right fact deployed at exactly the right place.
I never had a guidance counselor
who thought this was a legitimate career path,
that thought you could major in trivia
or be a professional ex-game show contestant.
And so I sold out way too young.
I didn't try to figure out what one does with that.
I studied computers because I heard that was the thing,
and I became a computer programmer --
not an especially good one,
not an especially happy one at the time when I was first on "Jeopardy" in 2004.
But that's what I was doing.
And it made it doubly ironic -- my computer background --
a few years later, I think 2009 or so,
when I got another phone call from "Jeopardy" saying,
"It's early days yet, but IBM tells us
they want to build a supercomputer to beat you at 'Jeopardy.'
Are you up for this?"
This was the first I'd heard of it.
And of course I said yes, for several reasons.
One, because playing "Jeopardy" is a great time.
It's fun. It's the most fun you can have with your pants on.

And I would do it for nothing.
I don't think they know that, luckily,
but I would go back and play for Arby's coupons.
I just love "Jeopardy," and I always have.
And second of all, because I'm a nerdy guy and this seemed like the future.
People playing computers on game shows
was the kind of thing I always imagined would happen in the future,
and now I could be on the stage with it.
I was not going to say no.
The third reason I said yes
is because I was pretty confident that I was going to win.
I had taken some artificial intelligence classes.
I knew there were no computers that could do what you need to do to win on "Jeopardy."
People don't realize how tough it is to write that kind of program
that can read a "Jeopardy" clue in a natural language like English
and understand all the double meanings, the puns, the red herrings,
unpack the meaning of the clue.
The kind of thing that a three- or four-year-old human, little kid could do,
very hard for a computer.
And I thought, well this is going to be child's play.
Yes, I will come destroy the computer and defend my species.

But as the years went on,
as IBM started throwing money and manpower and processor speed at this,
I started to get occasional updates from them,
and I started to get a little more worried.
I remember a journal article about this new question answering software that had a graph.

It was a scatter chart showing performance on "Jeopardy,"
tens of thousands of dots representing "Jeopardy" champions up at the top
with their performance plotted on number of --
I was going to say questions answered, but answers questioned, I guess,
clues responded to --
versus the accuracy of those answers.
So there's a certain performance level that the computer would need to get to.
And at first, it was very low.
There was no software that could compete at this kind of arena.
But then you see the line start to go up.
And it's getting very close to what they call the winner's cloud.
And I noticed in the upper right of the scatter chart
some darker dots, some black dots, that were a different color.
And thought, what are these?
"The black dots in the upper right represent 74-time 'Jeopardy' champion Ken Jennings."
And I saw this line coming for me.
And I realized, this is it.
This is what it looks like when the future comes for you.

It's not the Terminator's gun sight;
it's a little line coming closer and closer to the thing you can do,
the only thing that makes you special, the thing you're best at.
And when the game eventually happened about a year later,
it was very different than the "Jeopardy" games I'd been used to.
We were not playing in L.A. on the regular "Jeopardy" set.
Watson does not travel.
Watson's actually huge.
It's thousands of processors, a terabyte of memory,
trillions of bytes of memory.
We got to walk through his climate-controlled server room.
The only other "Jeopardy" contestant to this day I've ever been inside.
And so Watson does not travel.
You must come to it; you must make the pilgrimage.
So me and the other human player
wound up at this secret IBM research lab
in the middle of these snowy woods in Westchester County
to play the computer.
And we realized right away
that the computer had a big home court advantage.
There was a big Watson logo in the middle of the stage.
Like you're going to play the Chicago Bulls,
and there's the thing in the middle of their court.
And the crowd was full of IBM V.P.s and programmers
cheering on their little darling,
having poured millions of dollars into this
hoping against hope that the humans screw up,
and holding up "Go Watson" signs
and just applauding like pageant moms every time their little darling got one right.
I think guys had "W-A-T-S-O-N" written on their bellies in grease paint.
If you can imagine computer programmers with the letters "W-A-T-S-O-N" written on their gut,
it's an unpleasant sight.

But they were right. They were exactly right.
I don't want to spoil it, if you still have this sitting on your DVR,
but Watson won handily.
And I remember standing there behind the podium
as I could hear that little insectoid thumb clicking.
It had a robot thumb that was clicking on the buzzer.
And you could hear that little tick, tick, tick, tick.
And I remember thinking, this is it.
I felt obsolete.
I felt like a Detroit factory worker of the '80s
seeing a robot that could now do his job on the assembly line.
I felt like quiz show contestant was now the first job that had become obsolete
under this new regime of thinking computers.
And it hasn't been the last.
If you watch the news, you'll see occasionally --
and I see this all the time --
that pharmacists now, there's a machine that can fill prescriptions automatically
without actually needing a human pharmacist.
And a lot of law firms are getting rid of paralegals
because there's software that can sum up case laws and legal briefs and decisions.
You don't need human assistants for that anymore.
I read the other day about a program where you feed it a box score
from a baseball or football game
and it spits out a news article as if a human had watched the game
and was commenting on it.
And obviously these new technologies can't do as clever or creative a job
as the humans they're replacing,
but they're faster, and crucially, they're much, much cheaper.
So it makes me wonder what the economic effects of this might be.
I've read economists saying that, as a result of these new technologies,
we'll enter a new golden age of leisure
when we'll all have time for the things we really love
because all these onerous tasks will be taken over by Watson and his digital brethren.
I've heard other people say quite the opposite,
that this is yet another tier of the middle class
that's having the thing they can do taken away from them by a new technology
and that this is actually something ominous,
something that we should worry about.
I'm not an economist myself.
All I know is how it felt to be the guy put out of work.
And it was friggin' demoralizing. It was terrible.
Here's the one thing that I was ever good at,
and all it took was IBM pouring tens of millions of dollars and its smartest people
and thousands of processors working in parallel
and they could do the same thing.
They could do it a little bit faster and a little better on national TV,
and "I'm sorry, Ken. We don't need you anymore."
And it made me think, what does this mean,
if we're going to be able to start outsourcing,
not just lower unimportant brain functions.
I'm sure many of you remember a distant time

when we had to know phone numbers, when we knew our friends' phone numbers.
And suddenly there was a machine that did that,
and now we don't need to remember that anymore.
I have read that there's now actually evidence
that the hippocampus, the part of our brain that handles spacial relationships,
physically shrinks and atrophies
in people who use tools like GPS,
because we're not exercising our sense of direction anymore.
We're just obeying a little talking voice on our dashboard.
And as a result, a part of our brain that's supposed to do that kind of stuff
gets smaller and dumber.
And it made me think, what happens when computers are now better
at knowing and remembering stuff than we are?
Is all of our brain going to start to shrink and atrophy like that?
Are we as a culture going to start to value knowledge less?
As somebody who has always believed in the importance of the stuff that we know,
this was a terrifying idea to me.
The more I thought about it, I realized, no, it's still important.
The things we know are still important.
I came to believe there were two advantages
that those of us who have these things in our head have
over somebody who says, "Oh, yeah. I can Google that. Hold on a second."
There's an advantage of volume, and there's an advantage of time.
The advantage of volume, first,
just has to do with the complexity of the world nowadays.
There's so much information out there.
Being a Renaissance man or woman,
that's something that was only possible in the Renaissance.
Now it's really not possible
to be reasonably educated on every field of human endeavor.
There's just too much.
They say that the scope of human information
is now doubling every 18 months or so,
the sum total of human information.
That means between now and late 2014,
we will generate as much information, in terms of gigabytes,
as all of humanity has in all the previous millenia put together.
It's doubling every 18 months now.
This is terrifying because a lot of the big decisions we make
require the mastery of lots of different kinds of facts.
A decision like where do I go to school? What should I major in?
Who do I vote for?
Do I take this job or that one?
These are the decisions that require correct judgments
about many different kinds of facts.
If we have those facts at our mental fingertips,
we're going to be able to make informed decisions.
If, on the other hand, we need to look them all up,
we may be in trouble.
According to a National Geographic survey I just saw,
somewhere along the lines of 80 percent

of the people who vote in a U.S. presidential election about issues like foreign policy
cannot find Iraq or Afghanistan on a map.
If you can't do that first step,
are you really going to look up the other thousand facts you're going to need to know
to master your knowledge of U.S. foreign policy?
Quite probably not.
At some point you're just going to be like,
"You know what? There's too much to know. Screw it."
And you'll make a less informed decision.
The other issue is the advantage of time that you have
if you have all these things at your fingertips.
I always think of the story of a little girl named Tilly Smith.
She was a 10-year-old girl from Surrey, England
on vacation with her parents a few years ago in Phuket, Thailand.
She runs up to them on the beach one morning
and says, "Mom, Dad, we've got to get off the beach."
And they say, "What do you mean? We just got here."
And she said, "In Mr. Kearney's geography class last month,
he told us that when the tide goes out abruptly out to sea
and you see the waves churning way out there,
that's the sign of a tsunami, and you need to clear the beach."
What would you do if your 10-year-old daughter came up to you with this?
Her parents thought about it,
and they finally, to their credit, decided to believe her.
They told the lifeguard, they went back to the hotel,
and the lifeguard cleared over 100 people off the beach, luckily,
because that was the day of the Boxing Day tsunami,
the day after Christmas, 2004,
that killed thousands of people in Southeast Asia and around the Indian Ocean.
But not on that beach, not on Mai Khao Beach,
because this little girl had remembered one fact from her geography teacher a month before.
Now when facts come in handy like that --
I love that story because it shows you the power of one fact,
one remembered fact in exactly the right place at the right time --
normally something that's easier to see on game shows than in real life.
But in this case it happened in real life.
And it happens in real life all the time.
It's not always a tsunami, often it's a social situation.
It's a meeting or job interview or first date
or some relationship that gets lubricated
because two people realize they share some common piece of knowledge.
You say where you're from, and I say, "Oh, yeah."
Or your alma mater or your job,
and I know just a little something about it,
enough to get the ball rolling.
People love that shared connection that gets created
when somebody knows something about you.
It's like they took the time to get to know you before you even met.
That's often the advantage of time.
And it's not effective if you say, "Well, hold on.
You're from Fargo, North Dakota. Let me see what comes up.

Oh, yeah. Roger Maris was from Fargo."
That doesn't work. That's just annoying.

The great 18th-century British theologian and thinker, friend of Dr. Johnson,
Samuel Parr once said, "It's always better to know a thing than not to know it."
And if I have lived my life by any kind of creed, it's probably that.
I have always believed that the things we know -- that knowledge is an absolute good,
that the things we have learned and carry with us in our heads
are what make us who we are,
as individuals and as a species.
I don't know if I want to live in a world where knowledge is obsolete.
I don't want to live in a world where cultural literacy has been replaced
by these little bubbles of specialty,
so that none of us know about the common associations
that used to bind our civilization together.
I don't want to be the last trivia know-it-all
sitting on a mountain somewhere,
reciting to himself the state capitals and the names of "Simpsons" episodes
and the lyrics of Abba songs.
I feel like our civilization works when this is a vast cultural heritage that we all share
and that we know without having to outsource it to our devices,
to our search engines and our smartphones.
In the movies, when computers like Watson start to think,
things don't always end well.
Those movies are never about beautiful utopias.
It's always a terminator or a matrix or an astronaut getting sucked out an airlock in "2001."
Things always go terribly wrong.
And I feel like we're sort of at the point now
where we need to make that choice of what kind of future we want to be living in.
This is a question of leadership,
because it becomes a question of who leads the future.
On the one hand, we can choose between a new golden age
where information is more universally available
than it's ever been in human history,
where we all have the answers to our questions at our fingertips.
And on the other hand,
we have the potential to be living in some gloomy dystopia
where the machines have taken over
and we've all decided it's not important what we know anymore,
that knowledge isn't valuable because it's all out there in the cloud,
and why would we ever bother learning anything new.
Those are the two choices we have. I know which future I would rather be living in.
And we can all make that choice.
We make that choice by being curious, inquisitive people who like to learn,
who don't just say, "Well, as soon as the bell has rung and the class is over,
I don't have to learn anymore,"
or "Thank goodness I have my diploma. I'm done learning for a lifetime.
I don't have to learn new things anymore."
No, every day we should be striving to learn something new.
We should have this unquenchable curiosity for the world around us.
That's where the people you see on "Jeopardy" come from.

These know-it-alls, they're not Rainman-style savants
sitting at home memorizing the phone book.
I've met a lot of them.
For the most part, they are just normal folks
who are universally interested in the world around them, curious about everything,
thirsty for this knowledge about whatever subject.
We can live in one of these two worlds.
We can live in a world where our brains, the things that we know,
continue to be the thing that makes us special,
or a world in which we've outsourced all of that to evil supercomputers from the future like Watson.
Ladies and gentlemen, the choice is yours.
Thank you very much.

The dark side of competition in AI

Liv Boeree

Competition.
It's a fundamental part of human nature.
I was a professional poker player for 10 years,
so I've very much seen all the good, bad and ugly ways it can manifest.
When it's done right,
it can drive us to incredible feats in sports and innovation,
like when car companies compete over who can build the safest cars
or the most efficient solar panels.
Those are all examples of healthy competition,
because even though individual companies might come and go,
in the long run,
the game between them creates win-win outcomes
where everyone benefits in the end.
But sometimes competition is not so great
and can create lose-lose outcomes where everyone's worse off than before.
Take these AI beauty filters, for example.
As you can see, they're a very impressive technology.
They can salvage almost any picture.
They can even make Angelina and Margot more beautiful.
So they're very handy,
especially for influencers who, now, at the click of a button,
can transform into the most beautiful Hollywood versions of themselves.
But handy doesn't always mean healthy.
And I've personally noticed how quickly these things can train you
to hate your natural face.
And there's growing evidence that they're creating issues
like body dysmorphia, especially in young people.
Nonetheless, these things are now endemic to social media
because the nature of the game demands it.
The platforms are incentivized to provide them
because hotter pictures means more hijacked limbic systems,
which means more scrolling and thus more ad revenue.
And users are incentivized to use them
because hotter pictures get you more followers.
But this is a trap,
because once you start using these things,
it's really hard to go back.
Plus, you don't even get a competitive advantage from them anymore
because everyone else is already using them too.
So influencers are stuck using these things
with all the downsides
and very little upside.
A lose-lose game.
A similar kind of trap is playing out in our news media right now,
but with much worse consequences.
You'd think since the internet came along
that the increased competition between news outlets
would create a sort of positive spiral,

like a race to the top of nuanced, impartial, accurate journalism.
Instead, we're seeing a race to the bottom of clickbait and polarization,
where even respectable papers are increasingly leaning
into these kind of low-brow partisan tactics.
Again, this is due to crappy incentives.
Today, we no longer just read our news.
We interact with it by sharing and commenting.
And headlines that trigger emotions like fear or anger
are far more likely to go viral than neutral or positive ones.
So in many ways, news editors are in a similar kind of trap
as the influencers,
where, the more their competitors lean into clickbaity tactics,
the more they have to as well.
Otherwise, their stories just get lost in the noise.
But this is terrible for everybody,
because now the media get less trust from the public,
but also it becomes harder and harder for anyone to discern truth from fiction,
which is a really big problem for democracy.
Now, this process of competition gone wrong
is actually the driving force behind so many of our biggest issues.
Plastic pollution,
deforestation,
antibiotic overuse in farming,
arms races,
greenhouse gas emissions.
These are all a result of crappy incentives,
of poorly designed games that push their players --
be them people, companies or governments --
into taking strategies and tactics that defer costs and harms to the future.
And what's so ridiculous is that most of the time,
these guys don't even want to be doing this.
You know, it's not like packaging companies
want to fill the oceans with plastic
or farmers want to worsen antibiotic resistance.
But they're all stuck in the same dilemma of:
"If I don't use this tactic,
I'll get outcompeted by all the others who do.
So I have to do it, too."
This is the mechanism we need to fix as a civilization.
And I know what you're probably all thinking, "So it's capitalism."
No, it's not capitalism.
Which, yes, can cause problems,
but it can also solve them and has been fantastic in general.
It's something much deeper.
It's a force of misaligned incentives of game theory itself.
So a few years ago, I retired from poker,
in part because I wanted to understand this mechanism better.
Because it takes many different forms, and it goes by many different names.
These are just some of those names.
You can see they're a little bit abstract and clunky, right?
They don't exactly roll off the tongue.

And given how insidious and connected all of these problems are,
it helps to have a more visceral way of recognizing them.
So this is probably the only time
you're going to hear about the Bible at this conference.
But I want to tell you a quick story from it,
because allegedly, back in the Canaanite days,
there was a cult who wanted money and power so badly,
they were willing to sacrifice their literal children for it.
And they did this by burning them alive in an effigy
of a God that they believed would then reward them
for this ultimate sacrifice.
And the name of this god was Moloch.
Bit of a bummer, as stories go.
But you can see why it's an apt metaphor,
because sometimes we get so lost in winning the game right in front of us,
we lose sight of the bigger picture
and sacrifice too much in our pursuit of victory.
So just like these guys were sacrificing their children for power,
those influencers are sacrificing their happiness for likes.
Those news editors are sacrificing their integrity for clicks,
and polluters are sacrificing the biosphere for profit.
In all these examples,
the short-term incentives of the games themselves are pushing,
they're tempting their players
to sacrifice more and more of their future,
trapping them in a death spiral where they all lose in the end.
That's Moloch's trap.
The mechanism of unhealthy competition.
And the same is now happening in the AI industry.
We're all aware of the race that's heating up
between companies right now
over who can score the most compute,
who can get the biggest funding round or get the top talent.
Well, as more and more companies enter this race,
the greater the pressure for everyone to go as fast as possible
and sacrifice other important stuff like safety testing.
This has all the hallmarks of a Moloch trap.
Because, like, imagine you're a CEO who, you know, in your heart of hearts,
believes that your team is the best
to be able to safely build extremely powerful AI.
Well, if you go too slowly, then you run the risk of other,
much less cautious teams getting there first
and deploying their systems before you can.
So that in turn pushes you to be more reckless yourself.
And given how many different experts and researchers,
both within these companies
but also completely independent ones,
have been warning us about the extreme risks of rushed AI,
this approach is absolutely mad.
Plus, almost all AI companies
are beholden to satisfying their investors,

a short-term incentive which, over time, will inevitably start to conflict
with any benevolent mission.
And this wouldn't be a big deal
if this was really just toasters we're talking about here.
But AI, and especially AGI,
is set to be a bigger paradigm shift
than the agricultural or industrial revolutions.
A moment in time so pivotal,
it's deserving of reverence and reflection,
not something to be reduced to a corporate rat race
of who can score the most daily active users.
I'm not saying I know
what the right trade-off between acceleration and safety is,
but I do know that we'll never find out what that right trade-off is
if we let Moloch dictate it for us.
So what can we do?
Well, the good news is we have managed to coordinate
to escape some of Moloch's traps before.
We managed to save the ozone layer from CFCs
with the help of the Montreal Protocol.
We managed to reduce the number of nuclear weapons on Earth
by 80 percent,
with the help of the Strategic Arms Reduction Treaty in 1991.
So smart regulation may certainly help with AI too,
but ultimately,
it's the players within the game who have the most influence on it.
So we need AI leaders to show us
that they're not only aware of the risks their technologies pose,
but also the destructive nature of the incentives
that they're currently beholden to.
As their technological capabilities reach towards the power of gods,
they're going to need the godlike wisdom to know how to wield them.
So it doesn't fill me with encouragement
when I see a CEO of a very major company saying something like,
"I want people to know we made our competitor dance."
That is not the type of mindset we need here.
We need leaders who are willing to flip Moloch's playbook,
who are willing to sacrifice their own individual chance of winning
for the good of the whole.
Now, fortunately, the three leading labs are showing some signs of doing this.
Anthropic recently announced their responsible scaling policy,
which pledges to only increase capabilities
once certain security criteria have been met.
OpenAI have recently pledged
to dedicate 20 percent of their compute purely to alignment research.
And DeepMind have shown a decade-long focus
of science ahead of commerce,
like their development of AlphaFold,
which they gave away to the science community for free.
These are all steps in the right direction,
but they are still nowhere close to being enough.

I mean, most of these are currently just words,
they're not even proven actions.
So we need a clear way to turn the AI race into a definitive race to the top.
Perhaps companies can start competing over who can be within these metrics,
over who can develop the best security criteria.
A race of who can dedicate the most compute to alignment.
Now that would truly flip the middle finger to Moloch.
Competition can be an amazing tool,
provided we wield it wisely.
And we're going to need to do that
because the stakes we are playing for are astronomical.
If we get AI, and especially AGI, wrong,
it could lead to unimaginable catastrophe.
But if we get it right,
it could be our path out of many of these Moloch traps
that I've mentioned today.
And as things get crazier over the coming years,
which they're probably going to,
it's going to be more important than ever
that we remember that it is the real enemy here, Moloch.
Not any individual CEO or company, and certainly not one another.
So don't hate the players,
change the game.

How AI art could enhance humanity's collective memory

Refik Anadol

Let me begin by saying
that what you are looking at is not real.
These are artificial corals created by a generative AI algorithm,
trained with more than 100 million coral images.
Across the globe,
sea water is becoming less habitable due to the climate change
and the coral reefs are dying rapidly.
One day we might be only left with the simulations of corals
in a virtual world.
With this project, “Coral Dreams,”
our aim is to use AI
and try to create artificial realities while preserving disappearing nature.
I'm a media artist and director.
My team and I have been using generative AI as a collaborator
for seven years,
although it feels like 70.
We train machine-learning algorithms by harnessing large,
focused and publicly available data sets
and visualizing what I call humanity's collective memories,
such as nature, urban and culture.
Since [the] pandemic,
my focus has been to compile the largest data sets
and artificially preserve nature.
I am optimistic about generative AI
because of its potential for enhancing our memories.
We as artists can utilize this potential not only to represent nature,
but also to remember how it feels to be immersed in it in a digital age.
Generative AI creates possibilities to train algorithms with any image, sound,
text and even scent data.
For example, this is “Floral Dreams,”
an algorithm trained with more than 75 million floral images
of 16,000 species.
By using more than half a million scent molecules,
we were able to create the scent of these dreams.
Now let's please imagine a living archive
that we can walk into.
A universe that is constantly reimagined, [reconstructing] its forms, patterns,
colors and scents.
Our life is becoming increasingly rooted in digital worlds,
and the boundaries between physical and virtual,
technology and nature, are blurring.
Generative AI helps us to create new realities
and also project onto reality through possibility space.
Can we go to that space?
Can we fill it with our feelings, our senses?
Large language models are just the beginning
of a long journey of innovations [which] will bring more possibilities.
Soon, I believe we will be exploring hyper models,

text to image, to sound,
to scent, to life.
And a big challenge using generative AI in art
is how to provide models with original data.
For this project, “Glacier Dreams,”
we decided not to use existing models.
Instead, we decided to collect our own image,
sound, scent and climate data.
By traveling to our first destination, Iceland,
we were able to capture the beginning of our own narratives of glaciers.
I also believe that AI's capacity can be mapped
onto the complex history of human wisdom and consciousness in nature.
Could we use AI to preserve and learn about ancient knowledge in nature?
This was one of the first questions in my mind
when I met with the wonderful leaders
of Yawanawa tribe in Brazil, Acre, Amazonia, in the rainforest.
My mentors and heroes, Chief Nixiwaka and his creative force Putanny,
who oversee their cultural preservation and ecological sanctuary.
I became deeply inspired by their ways of learning
and remembering already existing knowledge in nature.
Together we started a new project,
a respectful co-creation and open-source AI rainforest model.
With this model,
generative AI can even reconstruct extinct flora and fauna
based on the tribe's deep and collective knowledge.
This project will help us, hopefully,
to bring ancient wisdom to our society respectfully.
My hope is that one day
AI becomes a mirror that can reflect collective memories of all humanity.
And I do believe that we can use it to bring people of any age and culture,
inspiration, joy and hope.
Thank you.

How to think computationally about AI, the universe and everything

Stephen Wolfram

Human language, mathematics, logic.
These are all ways to formalize the world.
And in our century,
there's a new and yet more powerful one: computation.
For nearly 50 years,
I've had the great privilege
of building up an ever-taller tower of science and technology
that's based on that idea of computation.
And so today, I want to tell you a little bit about what that's led to.
There's a lot to talk about, so I'm going to go quickly.
And sometimes I'm going to summarize in a sentence
what I've written a whole book about.
But you know,
I last gave a TED talk 13 years ago,
in February 2010,
soon after WolframAlpha launched,
and I ended that talk with a question.
Question was,
is computation ultimately what's underneath everything
in our universe?
I gave myself a decade to find out.
And actually, it could have needed a century.
But in April 2020, just after the decade mark,
we were thrilled to be able to announce
what seems to be the ultimate machine code of the universe.
And yes, it's computational.
So computation isn't just a possible formalization,
it's the ultimate one for our universe.
It all starts from the idea that space, like matter, is made of discrete elements,
and from that structure of space and everything in it,
it's defined just by a network of relations
between these elements that we might call atoms of space.
So it's all very elegant, but deeply abstract.
But here's kind of a humanized representation,
a version of the very beginning of the universe.
And what we're seeing here is the emergence of space
and everything in it
by the successive application of very simple computational rules.
And remember, these dots are not atoms in any existing space.
They're atoms of space that get put together to make space.
And yes, if we kept going long enough,
we could build our whole universe this way.
So eons later,
here's a chunk of space with two little black holes
that, if we wait a little while, will eventually merge,
generating little ripples of gravitational radiation.

And remember, all of this is built from pure computation.
But like fluid mechanics emerging from molecules,
what emerges here is space-time and Einstein's equations for gravity,
though there are deviations that we just might be able to detect,
like that the dimensionality of space won't always be precisely three.
And there's something else.
Our computational rules can inevitably be applied in many ways,
each defining a different kind of thread of time,
a different path of history that can branch and merge.
But as observers embedded in this universe,
we're branching and merging, too.
And it turns out that quantum mechanics emerges as the story
of how branching minds perceive a branching universe.
So the little pink lines you might be able to see here
show the structure of what we call branchial space,
the space of quantum branches.
And one of the stunningly beautiful things,
at least for physicists like me,
is that the same phenomenon that in physical space gives us gravity,
in branchial space gives us quantum mechanics.
So in the history of science so far,
I think we can identify sort of four broad paradigms
for making models of the world that can be distinguished
kind of by how they deal with time.
So in antiquity and in plenty of areas of science, even today,
it's all about kind of, what are things made of.
And time doesn't really enter.
But in the 1600s came the idea of modeling things
with mathematical formulas in which time enters,
but basically just as a coordinate value.
Then in the 1980s, and this is something in which I was deeply involved,
came the idea of making models
by starting with simple computational rules
and just letting them run.
So can one predict what will happen?
No.
There's what I call computational irreducibility,
in which, in effect, the passage of time corresponds to an irreducible computation
that we have to run in order to work out how it will turn out.
But now there's kind of something,
something even more -- in our physics project,
there’s things that have become multi-computational,
with many threads of time
that can only be knitted together by an observer.
So it's kind of a new paradigm that actually seems to unlock things
not only in fundamental physics,
but also in the foundations of mathematics and computer science,
and possibly in areas like biology and economics as well.
So I talked about building up the universe
by repeatedly applying a computational rule.
But how is that rule picked?

Well, actually it isn't,
because all possible rules are used,
and we're building up what I call the ruliad,
the kind of deeply abstract but unique object
that is the entangled limit of all possible computational processes.
Here's a tiny fragment of it shown in terms of Turing machines.
So this ruliad is everything.
And we as observers are necessarily part of it.
In the ruliad as a whole,
in a sense, everything computationally possible can happen.
But observers like us just sample specific slices of the ruliad.
And there are two crucial facts about us.
First, we're computationally bounded, our minds are limited,
and second, we believe we're persistent in time,
even though we're made of different atoms of space at every moment.
So then, here's the big result.
What observers with those characteristics perceive in the ruliad
necessarily follows certain laws.
And those laws turn out to be precisely
the three key theories of 20th century physics:
general relativity, quantum mechanics,
and statistical mechanics in the second law.
So it's because we're observers like us
that we perceive the laws of physics we do.
We can think of sort of different minds
as being at different places in rulial space.
Human minds who think alike are nearby,
animals further away,
and further out, we get to kind of alien minds
where it's hard to make a translation.
So how can we get intuition for all of this?
Well, one thing we can do is use generative AI
to take what amounts to an incredibly tiny slice of the ruliad
aligned with images we humans have produced.
We can think of this as sort of a place in the ruliad
described by using the concept of a cat in a party hat.
So zooming out, we saw there
what we might call Cat Island.
Pretty soon we’re in a kind of an inter-concept space.
Occasionally things will look familiar,
but mostly, what we'll see is things we humans don't have words for.
In physical space, we explore the universe
by sending out spacecraft.
In rulial space, we explore more
by expanding our concepts and our paradigms.
We can kind of get a sense of what's out there
by sampling possible rules,
doing what I call ruliology.
So even with incredibly simple rules,
there's incredible richness.
But the issue is that most of it doesn't yet connect

with things we humans understand or care about.
It's like when we look at the natural world
and only gradually realize that we can use features of it for technology.
So even after everything our civilization has achieved,
we're just at the very, very beginning of exploring rulial space.
What about AIs?
Well, just like we can do ruliology,
AIs can in principle go out and explore rulial space.
Left to their own devices, though,
they'll mostly just be doing things
we humans don't connect with or care about.
So the big achievements of AI in recent times
have been about making systems that are closely aligned with us humans.
We train LLMs on billions of web pages so they can produce texts
that's typical of what we humans write.
And yes, the fact that this works
is undoubtedly telling us some deep scientific things
about the semantic grammar of language
and generalizations of things like logic
that perhaps we should have known centuries ago.
You know, for much of human history,
we were kind of like the LLMs,
figuring things out by kind of matching patterns in our minds.
But then came more systematic formalization and eventually computation.
And with that, we got a whole other level of power to truly create new things
and to, in effect, go wherever we want in the ruliad.
But the challenge is to do that in a way that connects with what we humans,
and our AIs, understand.
In fact, I've devoted a large part of my life
to kind of trying to build that bridge.
It's all been about creating a language for expressing ourselves computationally,
a language for computational thinking.
The goal is to formalize what we know about the world in computational terms,
to have computational ways to represent cities and chemicals and movies
and humor and formulas and our knowledge about them.
It’s been a vast undertaking that spanned more than four decades of my life,
but it's something very unique and different.
But I'm happy to report that in what has been Mathematica
and is now the Wolfram Language,
I think we firmly succeeded in creating
a truly full-scale computational language.
In effect,
every one of these functions here can be thought of as formalizing
and encapsulating, in computational terms,
some facet of the intellectual achievements of our civilization.
It's sort of the most concentrated form of intellectual expression that I know,
sort of finding the essence of everything and coherently expressing it
in the design of our computational language.
For me personally,
it's been an amazing journey, kind of, year after year,
building the sort of tower of ideas and technology that's needed.

And nowadays sharing that process with the world
in things like open live streams and so on.
A few centuries ago,
the development of mathematical notation,
and what amounts to the language of mathematics,
gave a systematic way to express math and made possible algebra and calculus,
and eventually all of modern mathematical science.
And computational language now provides a similar path,
letting us ultimately create a computational X
for all imaginable fields X.
I mean, we've seen the growth of computer science, CS,
but computational language opens up something ultimately much bigger
and broader, CX.
I mean, for 70 years we've had programming languages
which are about telling computers in their terms what to do.
But computational language
is about something intellectually much bigger.
It's about taking everything we can think about
and operationalizing it in computational terms.
You know, I built the Wolfram Language
first and foremost because I wanted to use it myself.
And now when I use it,
I feel like it's kind of giving me some kind of superpower.
I just have to imagine something in computational terms.
And then the language sort of almost magically lets me bring it into reality,
see its consequences, and build on them.
And yes, that's the sort of superpower
that's let me do things like our physics project.
And over the past 35 years,
it's been my great privilege to share this superpower with many other people,
and by doing so,
to have enabled an incredible number of advances across many fields.
It's sort of a wonderful thing to see people, researchers, CEOs, kids,
using our language to fluently think in computational terms,
kind of crispening up their own thinking,
and then in effect, automatically calling in computational superpowers.
And now it's not just people who can do that.
AIs can use our computational language as a tool, too.
Yes, to get their facts straight,
but even more importantly, to compute new facts.
There are already some integrations of our technology into LLMs.
There's a lot more you'll be seeing soon.
And, you know, when it comes to building new things
in a very powerful emerging workflow,
it's basically to start by telling the LLM roughly what you want,
then to have it try to express that in precise Wolfram Language,
then, and this is a critical feature of our computational language,
compared to, for example, programming language,
you as a human can read the code,
and if it does what you want,
you can use it as kind of a dependable component to build on.

OK, but let's say we use more and more AI,
more and more computation.
What's the world going to be like?
From the industrial revolution on,
we’ve been used to doing engineering where we can in effect,
see how the gears mesh to understand how things work.
But computational irreducibility
now shows us that that won't always be possible.
We won't always be able to make a kind of simple human or, say,
mathematical narrative
to explain or predict what a system will do.
And yes, this is science, in effect, eating itself from the inside.
From all the successes of mathematical science,
we've come to believe that somehow, if we only could find them,
there'd be formulas to kind of predict everything.
But now computational irreducibility shows us that that isn't true.
And that in effect, to find out what a system will do,
we have to go through the same irreducible computational steps
as the system itself.
Yes, it's a weakness of science,
but it's also why the passage of time is significant and meaningful
and why we can't just sort of jump ahead to get the answer.
We have to live the steps.
It's actually going to be, I think, a great societal dilemma of the future.
If we let our AIs achieve their kind of full computational potential,
they'll have lots of computational irreducibility
and we won't be able to predict what they'll do.
But if we put constraints on them to make them more predictable,
we'll limit what they can do for us.
So what will it feel like if our world is full of computational irreducibility?
Well, it's really nothing new
because that's the story with much of nature.
And what's happened there
is that we've found ways to operate within nature,
even though nature can sometimes still surprise us.
And so it will be with the AIs.
We might give them a constitution, but there will always be consequences
we can't predict.
Of course, even figuring out societally what we want from the AIs is hard.
Maybe we need you know, a promptocracy
where people write prompts instead of just voting.
But basically, every control the outcome scheme
seems full of both political philosophy
and computational irreducibility gotchas.
You know, if we look at the whole arc of human history,
the one thing that's systematically changed
is that more and more gets automated.
And LLMs just gave us a dramatic and unexpected example of that.
So what does that mean?
Does that mean that in the end, us humans will have nothing to do?
Well, if we look at history,

what seems to happen is that when one thing gets automated away,
it opens up lots of new things to do.
And as economies develop,
the pie chart of occupations seems to get more and more fragmented.
And now we're back to the ruliad.
Because at a foundational level,
what's happening is that automation is opening up more directions
to go in the ruliad.
But there's no abstract way to choose between these.
It's a question of what we humans want,
and it requires kind of humans doing work to define that.
So a society of AI as sort of untethered by human input,
would effectively go off and explore the whole ruliad.
But most of what they do would seem to us random and pointless,
much like most of nature doesn't seem to us right now,
like it's achieving a purpose.
I mean, one used to imagine that to build things that are useful to us,
we'd have to do it kind of step by step.
But AI and the whole phenomenon of computation
tell us that really what we need
is more just to define what we want.
Then computation, AI, automation can make it happen.
And yes, I think the key to defining in a clear way what we want
is computational language.
And, you know, even after 35 years,
for many people,
Wolfram Language is still sort of an artifact from the future.
If your job is to program, it seems like a cheat.
How come you can do in an hour what would usually take you a week?
But it can also be kind of daunting because having dashed off that one thing,
you now have to conceptualize the next thing.
Of course, it's great for CEOs and CTOs
and intellectual leaders who are ready to race on to the next thing.
And indeed, it's an impressively popular thing in that set.
In a sense, what's happening is that Wolfram Language shifts
from concentrating on mechanics to concentrating on conceptualization,
and the key to that conceptualization is broad computational thinking.
So how can one learn to do that?
It's not really a story of CS,
it's really a story of CX.
And as a kind of education,
it's more like liberal arts than STEM.
It's part of a trend that when you automate technical execution,
what becomes important is not figuring out how to do things,
but what to do.
And that's more a story of broad knowledge and general thinking
than any kind of narrow specialization.
You know, there's sort of an unexpected human centeredness to all of this.
We might have thought that with the advance of science and technology,
the particulars of us humans would become ever less relevant.
But we've discovered that that's not true, and that, in fact, everything,

even our physics,
depends on how we humans happen to have sampled the ruliad.
Before our physics project,
we didn't know if our universe really was computational,
but now it's pretty clear that it is.
And from that, we're sort of inexorably led to the ruliad,
with all its kind of vastness
so hugely greater than the physical space in our universe.
So where will we go in the ruliad?
Computational language is what lets us chart our path.
It lets us humans define our goals and our journeys.
And what's amazing is that all the power and depth
of what's out there in the ruliad is accessible to everyone.
One just has to learn to harness those computational superpowers,
which kind of starts here,
you know, our portal to the ruliad.
Thank you.

What is an AI anyway?

Mustafa Suleyman

I want to tell you what I see coming.
I've been lucky enough to be working on AI for almost 15 years now.
Back when I started, to describe it as fringe would be an understatement.
Researchers would say, “No, no, we’re only working on machine learning.”
Because working on AI was seen as way too out there.
In 2010, just the very mention of the phrase “AGI,”
artificial general intelligence,
would get you some seriously strange looks
and even a cold shoulder.
"You're actually building AGI?" people would say.
"Isn't that something out of science fiction?"
People thought it was 50 years away or 100 years away,
if it was even possible at all.
Talk of AI was, I guess, kind of embarrassing.
People generally thought we were weird.
And I guess in some ways we kind of were.
It wasn't long, though, before AI started beating humans
at a whole range of tasks
that people previously thought were way out of reach.
Understanding images,
translating languages,
transcribing speech,
playing Go and chess
and even diagnosing diseases.
People started waking up to the fact
that AI was going to have an enormous impact,
and they were rightly asking technologists like me
some pretty tough questions.
Is it true that AI is going to solve the climate crisis?
Will it make personalized education available to everyone?
Does it mean we'll all get universal basic income
and we won't have to work anymore?
Should I be afraid?
What does it mean for weapons and war?
And of course, will China win?
Are we in a race?
Are we headed for a mass misinformation apocalypse?
All good questions.
But it was actually a simpler
and much more kind of fundamental question that left me puzzled.
One that actually gets to the very heart of my work every day.
One morning over breakfast,
my six-year-old nephew Caspian was playing with Pi,
the AI I created at my last company, Inflection.
With a mouthful of scrambled eggs,
he looked at me plain in the face and said,

"But Mustafa, what is an AI anyway?"
He's such a sincere and curious and optimistic little guy.
He'd been talking to Pi about how cool it would be if one day in the future,
he could visit dinosaurs at the zoo.
And how he could make infinite amounts of chocolate at home.
And why Pi couldn’t yet play I Spy.
"Well," I said, "it's a clever piece of software
that's read most of the text on the open internet,
and it can talk to you about anything you want."
"Right.
So like a person then?"
I was stumped.
Genuinely left scratching my head.
All my boring stock answers came rushing through my mind.
"No, but AI is just another general-purpose technology,
like printing or steam."
It will be a tool that will augment us
and make us smarter and more productive.
And when it gets better over time,
it'll be like an all-knowing oracle
that will help us solve grand scientific challenges."
You know, all of these responses started to feel, I guess,
a little bit defensive.
And actually better suited to a policy seminar
than breakfast with a no-nonsense six-year-old.
"Why am I hesitating?" I thought to myself.
You know, let's be honest.
My nephew was asking me a simple question
that those of us in AI just don't confront often enough.
What is it that we are actually creating?
What does it mean to make something totally new,
fundamentally different to any invention that we have known before?
It is clear that we are at an inflection point
in the history of humanity.
On our current trajectory,
we're headed towards the emergence of something
that we are all struggling to describe,
and yet we cannot control what we don't understand.
And so the metaphors,
the mental models,
the names, these all matter
if we’re to get the most out of AI whilst limiting its potential downsides.
As someone who embraces the possibilities of this technology,
but who's also always cared deeply about its ethics,
we should, I think,
be able to easily describe what it is we are building.
And that includes the six-year-olds.
So it's in that spirit that I offer up today the following metaphor
for helping us to try to grapple with what this moment really is.
I think AI should best be understood
as something like a new digital species.

Now, don't take this too literally,
but I predict that we'll come to see them as digital companions,
new partners in the journeys of all our lives.
Whether you think we're on a 10-, 20- or 30-year path here,
this is, in my view, the most accurate and most fundamentally honest way
of describing what's actually coming.
And above all, it enables everybody to prepare for
and shape what comes next.
Now I totally get, this is a strong claim,
and I'm going to explain to everyone as best I can why I'm making it.
But first, let me just try to set the context.
From the very first microscopic organisms,
life on Earth stretches back billions of years.
Over that time, life evolved and diversified.
Then a few million years ago, something began to shift.
After countless cycles of growth and adaptation,
one of life's branches began using tools, and that branch grew into us.
We went on to produce a mesmerizing variety of tools,
at first slowly and then with astonishing speed,
we went from stone axes and fire
to language, writing and eventually industrial technologies.
One invention unleashed a thousand more.
And in time, we became homo technologicus.
Around 80 years ago,
another new branch of technology began.
With the invention of computers,
we quickly jumped from the first mainframes and transistors
to today's smartphones and virtual-reality headsets.
Information, knowledge, communication, computation.
In this revolution,
creation has exploded like never before.
And now a new wave is upon us.
Artificial intelligence.
These waves of history are clearly speeding up,
as each one is amplified and accelerated by the last.
And if you look back,
it's clear that we are in the fastest
and most consequential wave ever.
The journeys of humanity and technology are now deeply intertwined.
In just 18 months,
over a billion people have used large language models.
We've witnessed one landmark event after another.
Just a few years ago, people said that AI would never be creative.
And yet AI now feels like an endless river of creativity,
making poetry and images and music and video that stretch the imagination.
People said it would never be empathetic.
And yet today, millions of people enjoy meaningful conversations with AIs,
talking about their hopes and dreams
and helping them work through difficult emotional challenges.
AIs can now drive cars,
manage energy grids

and even invent new molecules.
Just a few years ago, each of these was impossible.
And all of this is turbocharged by spiraling exponentials of data
and computation.
Last year, Inflection 2.5, our last model,
used five billion times more computation
than the DeepMind AI that beat the old-school Atari games
just over 10 years ago.
That's nine orders of magnitude more computation.
10x per year,
every year for almost a decade.
Over the same time, the size of these models has grown
from first tens of millions of parameters to then billions of parameters,
and very soon, tens of trillions of parameters.
If someone did nothing but read 24 hours a day for their entire life,
they'd consume eight billion words.
And of course, that's a lot of words.
But today, the most advanced AIs consume more than eight trillion words
in a single month of training.
And all of this is set to continue.
The long arc of technological history is now in an extraordinary new phase.
So what does this mean in practice?
Well, just as the internet gave us the browser
and the smartphone gave us apps,
the cloud-based supercomputer is ushering in a new era
of ubiquitous AIs.
Everything will soon be represented by a conversational interface.
Or, to put it another way, a personal AI.
And these AIs will be infinitely knowledgeable,
and soon they'll be factually accurate and reliable.
They'll have near-perfect IQ.
They'll also have exceptional EQ.
They'll be kind, supportive, empathetic.
These elements on their own would be transformational.
Just imagine if everybody had a personalized tutor in their pocket
and access to low-cost medical advice.
A lawyer and a doctor,
a business strategist and coach --
all in your pocket 24 hours a day.
But things really start to change when they develop what I call AQ,
their "actions quotient."
This is their ability to actually get stuff done
in the digital and physical world.
And before long, it won't just be people that have AIs.
Strange as it may sound, every organization,
from small business to nonprofit to national government,
each will have their own.
Every town, building and object
will be represented by a unique interactive persona.
And these won't just be mechanistic assistants.
They'll be companions, confidants,

colleagues, friends and partners,
as varied and unique as we all are.
At this point, AIs will convincingly imitate humans at most tasks.
And we'll feel this at the most intimate of scales.
An AI organizing a community get-together for an elderly neighbor.
A sympathetic expert helping you make sense of a difficult diagnosis.
But we'll also feel it at the largest scales.
Accelerating scientific discovery,
autonomous cars on the roads,
drones in the skies.
They'll both order the takeout and run the power station.
They’ll interact with us and, of course, with each other.
They'll speak every language,
take in every pattern of sensor data,
sights, sounds,
streams and streams of information,
far surpassing what any one of us could consume in a thousand lifetimes.
So what is this?
What are these AIs?
If we are to prioritize safety above all else,
to ensure that this new wave always serves and amplifies humanity,
then we need to find the right metaphors for what this might become.
For years, we in the AI community, and I specifically,
have had a tendency to refer to this as just tools.
But that doesn't really capture what's actually happening here.
AIs are clearly more dynamic,
more ambiguous, more integrated
and more emergent than mere tools,
which are entirely subject to human control.
So to contain this wave,
to put human agency at its center
and to mitigate the inevitable unintended consequences
that are likely to arise,
we should start to think about them as we might a new kind of digital species.
Now it's just an analogy,
it's not a literal description, and it's not perfect.
For a start, they clearly aren't biological in any traditional sense,
but just pause for a moment
and really think about what they already do.
They communicate in our languages.
They see what we see.
They consume unimaginably large amounts of information.
They have memory.
They have personality.
They have creativity.
They can even reason to some extent and formulate rudimentary plans.
They can act autonomously if we allow them.
And they do all this at levels of sophistication
that is far beyond anything that we've ever known from a mere tool.
And so saying AI is mainly about the math or the code
is like saying we humans are mainly about carbon and water.

It's true, but it completely misses the point.
And yes, I get it, this is a super arresting thought
but I honestly think this frame helps sharpen our focus on the critical issues.
What are the risks?
What are the boundaries that we need to impose?
What kind of AI do we want to build or allow to be built?
This is a story that's still unfolding.
Nothing should be accepted as a given.
We all must choose what we create.
What AIs we bring into the world, or not.
These are the questions for all of us here today,
and all of us alive at this moment.
For me, the benefits of this technology are stunningly obvious,
and they inspire my life's work every single day.
But quite frankly, they'll speak for themselves.
Over the years, I've never shied away from highlighting risks
and talking about downsides.
Thinking in this way helps us focus on the huge challenges
that lie ahead for all of us.
But let's be clear.
There is no path to progress
where we leave technology behind.
The prize for all of civilization is immense.
We need solutions in health care and education, to our climate crisis.
And if AI delivers just a fraction of its potential,
the next decade is going to be the most productive in human history.
Here's another way to think about it.
In the past,
unlocking economic growth often came with huge downsides.
The economy expanded as people discovered new continents
and opened up new frontiers.
But they colonized populations at the same time.
We built factories,
but they were grim and dangerous places to work.
We struck oil,
but we polluted the planet.
Now because we are still designing and building AI,
we have the potential and opportunity to do it better,
radically better.
And today, we're not discovering a new continent
and plundering its resources.
We're building one from scratch.
Sometimes people say that data or chips are the 21st century’s new oil,
but that's totally the wrong image.
AI is to the mind
what nuclear fusion is to energy.
Limitless, abundant,
world-changing.
And AI really is different,
and that means we have to think about it creatively and honestly.
We have to push our analogies and our metaphors

to the very limits
to be able to grapple with what's coming.
Because this is not just another invention.
AI is itself an infinite inventor.
And yes, this is exciting and promising and concerning
and intriguing all at once.
To be quite honest, it's pretty surreal.
But step back,
see it on the long view of glacial time,
and these really are the very most appropriate metaphors that we have today.
Since the beginning of life on Earth,
we've been evolving, changing
and then creating everything around us in our human world today.
And AI isn't something outside of this story.
In fact, it's the very opposite.
It's the whole of everything that we have created,
distilled down into something that we can all interact with
and benefit from.
It's a reflection of humanity across time,
and in this sense,
it isn't a new species at all.
This is where the metaphors end.
Here's what I'll tell Caspian next time he asks.
AI isn't separate.
AI isn't even in some senses, new.
AI is us.
It's all of us.
And this is perhaps the most promising and vital thing of all
that even a six-year-old can get a sense for.
As we build out AI,
we can and must reflect all that is good,
all that we love,
all that is special about humanity:
our empathy, our kindness,
our curiosity and our creativity.
This, I would argue, is the greatest challenge of the 21st century,
but also the most wonderful,
inspiring and hopeful opportunity for all of us.
Thank you.

Any sufficiently advanced technology is indistinguishable from magic.

Arthur C. Clarke