

The danger of AI is weirder than you think

Fonte: <https://youtu.be/OhCzX0iLnOc>



English Transcript	Tradução
The danger of AI is weirder than you think	O Perigo da IA é mais estranho do que você pensa
So, artificial intelligence—it's known for disrupting all kinds of industries. But what about ice cream? What kind of mind-blowing new flavors could we generate with the power of an advanced artificial intelligence?	Assim, a inteligência artificial, ela é conhecida por causar disrupção em todos os tipos de indústrias. Mas e quanto ao sorvete? Que tipo de novos sabores impressionantes nós poderíamos gerar com o poder de uma inteligência artificial avançada?
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. Here are some of the flavors that the AI came up with. These flavors are not as 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?	Então, eu me juntei a um grupo de programadores da Kealing Middle School para descobrir a resposta para esta pergunta. Eles coletaram mais de 1.600 sabores de sorvete existentes, e juntos nós os alimentamos em um algoritmo para ver o que ele geraria. Aqui estão alguns dos sabores que a IA inventou. Esses sabores não são tão deliciosos quanto nós poderíamos ter esperado que fossem, então a pergunta é: o que aconteceu? O que deu errado? A IA está tentando nos matar, ou ela está tentando fazer o que pedimos, e houve um problema?
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	Nos filmes, quando algo dá errado com a IA, é geralmente porque a IA decidiu que não quer mais obedecer aos humanos, e ela tem seus próprios objetivos, muito obrigado. Na vida real, porém, a IA que nós realmente temos não é nem de longe inteligente o suficiente para isso. Ela tem o poder de processamento aproximado de uma minhoca, ou talvez no máximo de



English Transcript	Tradução
most a single honeybee. And actually, probably maybe less. We're constantly learning new things about brains that make it clear how much our AIs don't measure up to real brains.	uma única abelha. E na verdade, provavelmente talvez menos. Nós estamos constantemente aprendendo novas coisas sobre cérebros que deixam claro o quanto as nossas IAs não se comparam aos cérebros reais.
Today's AI can do a single task, like identify a pedestrian in a picture, but it doesn't have a concept of what the pedestrian is beyond 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 asked it to do? It will if it can, but it might not do what we actually want.	A IA de hoje pode fazer uma única tarefa, como identificar um pedestre em uma imagem, mas ela não tem o conceito do que o pedestre é além de uma coleção de linhas, texturas e coisas. Ela não sabe o que um humano realmente é. Então, a IA de hoje fará o que pedimos para ela fazer? Ela fará se puder, mas ela pode não fazer o que nós realmente queremos.
Let's say that you are trying to get an AI to take this collection of robot parts and assemble them into some kind of robot that could 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.	Digamos que você esteja tentando fazer uma IA pegar esta coleção de peças de robô e montá-las em algum tipo de robô que consiga ir do ponto A ao ponto B. Agora, se você fosse tentar resolver este problema escrevendo um programa de computador no estilo tradicional, você daria ao programa instruções passo a passo sobre como pegar essas peças, como montá-las em um robô com pernas, e depois como usar essas pernas para andar até o ponto B. Mas quando você usa a IA para resolver o problema, acontece de forma diferente: você não diz a ela como resolver o problema; você apenas dá o objetivo a ela, e ela tem que descobrir por si mesma, por meio de tentativa e erro, como alcançar esse objetivo.
It turns out the way the 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. 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, but that it is going to do exactly what we ask it to do.	Acontece que a maneira como a IA tende a resolver esse problema específico é fazendo isto: ela se monta em uma torre e depois tomba e cai no ponto B. Tecnicamente, isso resolve o problema, tecnicamente, ela chegou ao ponto B. O perigo da IA não é que ela vá se rebelar contra nós, mas que ela vá fazer exatamente o que pedimos para ela fazer.
So then, the trick of working with AI becomes: how do we set up the problem so that it actually does what we want?	Então, o truque de trabalhar com IA se torna: como nós configuramos o problema para que ela realmente faça o que nós queremos?
This little robot here is being controlled by an AI. The AI came up with the design for the robot's 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... it technically got to the end of that obstacle course! So, you can see how hard it is to get AI to do something as simple as just walk.	Este pequeno robô aqui está sendo controlado por uma IA. A IA inventou o design para as pernas do robô e depois descobriu como usá-las para passar por todos estes obstáculos. Mas quando David Ha configurou este experimento, ele teve que configurá-lo com limites muito, muito estritos sobre o quão grandes a IA tinha permissão de fazer as pernas, porque caso contrário... ela tecnicamente chegou ao final daquele circuito de obstáculos! Então, você pode ver como é difícil fazer a IA realizar algo tão simples como apenas andar.
Seeing the AI do this, you may say, "Okay, no fair! You can't just be a tall tower and fall over; you have to actually use legs to walk." 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	Vendo a IA fazer isso, você pode dizer: "Ok, sem trapacear! Você não pode simplesmente ser uma torre alta e tombar; você tem que realmente usar pernas para andar." Acontece que isso nem sempre funciona também. O trabalho desta IA era se mover rápido; eles



English Transcript	Tradução
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, as is twitching along the floor in a heap. In my opinion, you know what should have been a whole lot weirder? The Terminator robots.	não disseram a ela que ela tinha que correr virada para a frente, ou que ela não podia usar seus braços. Então, é isto o que você obtém quando treina a IA para se mover rápido: você obtém coisas como cambalhotas e caminhadas bobas. É realmente comum, assim como se contorcer pelo chão em um monte. Na minha opinião, sabe o que deveria ter sido muito mais estranho? Os robôs do Exterminador do Futuro.
Hacking the matrix is another thing that AI will do if you give it a chance. 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'll figure out how to move faster by glitching repeatedly into the floor.	Hackear a matrix é outra coisa que a IA fará se você der uma chance. Se você treinar uma IA em uma simulação, ela aprenderá como fazer coisas como hackear os erros matemáticos da simulação e colhê-los para obter energia. Ou ela descobrirá como se mover mais rápido travando repetidamente no chão.
Working with AI is less like working with another human and a lot more like working with some kind of weird force of nature. 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.	Trabalhar com IA é menos como trabalhar com outro humano e muito mais como trabalhar com algum tipo de força estranha da natureza. É muito fácil dar acidentalmente à IA o problema errado para resolver, e muitas vezes nós não percebemos isso até que algo tenha realmente dado errado.
Here's an experiment I did where I wanted the AI to copy paint colors—to invent new paint colors given a list like the ones here on the left. And here's what the AI actually came up with. So, technically, it did what I asked it to do. I thought I was asking it for 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. 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. Its entire world is the data that I gave it; like with the ice cream flavors, it doesn't know about anything else.	Aqui está um experimento que eu fiz onde eu queria que a IA copiasse cores de tinta, para inventar novas cores de tinta dada uma lista como as que estão aqui à esquerda. E aqui está o que a IA realmente inventou. Então, tecnicamente, ela fez o que eu pedi para ela fazer. Eu pensei que estava pedindo a ela nomes bonitos de cores de tinta, mas o que eu estava realmente pedindo para ela fazer era apenas imitar os tipos de combinações de letras que ela tinha visto no original. Eu não disse nada a ela sobre o que as palavras significam, ou que existem talvez algumas palavras que ela deveria evitar usar nestas cores de tinta. O mundo inteiro dela é os dados que eu dei a ela; como no caso dos sabores de sorvete, ela não sabe sobre mais nada.
It is through the data that we often accidentally tell AI to do the wrong thing.	É através dos dados que nós frequentemente, por acidente, dizemos à IA para fazer a coisa errada.
This is a fish called a tench. 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 the AI 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 is trying to identify a fish? Well, it turns out that the tench is a trophy fish, so in a lot of the 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.	Este é um peixe chamado tenca. Houve um grupo de pesquisadores que treinou uma IA para identificar esta tenca em imagens. Mas então, quando eles perguntaram a ela qual parte da imagem a IA estava realmente usando para identificar o peixe, aqui está o que ela destacou. Sim, aqueles são dedos humanos. Por que ela estaria procurando por dedos humanos se ela está tentando identificar um peixe? Bem, acontece que a tenca é um peixe de troféu, então em muitas das imagens que a IA tinha visto deste peixe durante o treinamento, o peixe parecia com isto, e ela não sabia que os dedos não fazem parte do peixe.
You see why it is so hard to design an AI that actually can understand what it's looking at. This is	Você vê por que é tão difícil projetar uma IA que realmente consiga entender o que está olhando. É por



English Transcript	Tradução
why designing the image recognition in self-driving cars is so hard, and why so many self-driving car failures happen because the AI got confused.	isso que projetar o reconhecimento de imagem em carros autônomos é tão difícil, e por que tantas falhas de carros autônomos acontecem porque a IA ficou confusa.
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. 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 in highway driving, where you would expect to see trucks from behind. Trucks from the side are 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.	Eu quero falar sobre um exemplo de 2016. Houve um acidente fatal quando alguém estava usando o piloto automático de IA da Tesla, mas em vez de usá-lo na rodovia como ele foi projetado, eles o usaram em ruas da cidade. O que aconteceu foi que um caminhão cruzou na frente do carro, e o carro não freou. Agora, a IA definitivamente foi treinada para reconhecer caminhões em imagens, mas o que parece ter acontecido é que a IA foi treinada para reconhecer caminhões em condução em rodovias, onde você esperaria ver caminhões por trás. Caminhões de lado não deveriam acontecer em uma rodovia. E assim, quando a IA viu este caminhão, parece que a IA o reconheceu como mais provável de ser uma placa de sinalização e, portanto, seguro para dirigir por baixo.
Here's an AI mess-up from a different field. Amazon recently had to give up on a resume-sorting algorithm 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 resumes of people who they had hired in the past. From these examples, the AI learned to avoid the resumes of people who had gone to women's colleges, or who had the word "women's" 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. Technically, it did what they asked it to do; they just accidentally asked it to do the wrong thing.	Aqui está uma trapalhada de IA de um campo diferente. A Amazon recentemente teve que desistir de um algoritmo de triagem de currículos no qual eles estavam trabalhando quando descobriram que o algoritmo tinha aprendido a discriminar mulheres. O que aconteceu é que eles o tinham treinado em exemplos de currículos de pessoas que eles tinham contratado no passado. A partir destes exemplos, a IA aprendeu a evitar os currículos de pessoas que tinham frequentado faculdades femininas, ou que tinham a palavra "feminino" em algum lugar em seu currículo, como em "time de futebol feminino" ou "Sociedade de Mulheres Engenheiras". A IA não sabia que não era para copiar esta coisa particular que ela tinha visto os humanos fazerem. Tecnicamente, ela fez o que eles pediram para ela fazer; eles apenas pediram acidentalmente para ela fazer a coisa errada.
This happens all the time with AI. AI can be really destructive and not know it. The AIs that recommend new content on Facebook and YouTube are optimized to increase the number of clicks and views. Unfortunately, one way that they have found of doing this is to recommend content about 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.	Isso acontece o tempo todo com a IA. A IA pode ser realmente destrutiva e não saber disso. As IAs que recomendam novos conteúdos no Facebook e no YouTube são otimizadas para aumentar o número de cliques e visualizações. Infelizmente, uma maneira que elas encontraram de fazer isso é recomendar conteúdos sobre teorias da conspiração ou intolerância. As IAs em si não têm nenhum conceito do que este conteúdo realmente é, e elas não têm nenhum conceito do que as consequências podem ser ao recomendar este conteúdo.
So, when we're working with AI, it's up to us to avoid problems. Avoiding things going wrong may come down to the age-old problem of communication, where we as humans have to	Portanto, quando estamos trabalhando com IA, cabe a nós evitar problemas. Evitar que as coisas dêem errado pode se resumir ao antigo problema da comunicação, onde nós, como humanos, temos que aprender como



English Transcript	Tradução
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.	nos comunicar com a IA. Nós temos que aprender o que a IA é capaz de fazer e o que não é, e entender que, com seu pequenino cérebro de minhoca, a IA não entende realmente o que nós estamos tentando pedir para ela fazer.
In other words, we have to be prepared to work with an AI that's not the super-competent, all-knowing AI of science fiction; we have to be prepared to work with the AI that we actually have in the present day. And present-day AI is plenty weird enough.	Em outras palavras, nós temos que estar preparados para trabalhar com uma IA que não é a IA supercompetente e onisciente da ficção científica; nós temos que estar preparados para trabalhar com a IA que nós realmente temos no presente. E a IA do presente é bastante estranha o suficiente.
Thank you.	Obrigada.



Contagem de palavras

A tabela abaixo exhibe as palavras encontradas neste vídeo, bem como o número de vezes em que aparecem.

Veja também: [Para que serve esta tabela?](#)

Freq.	Palavra	Freq.	Palavra	Freq.	Palavra
81	to	81	the	70	it
50	ai	44	is	35	of
34	that	32	and	32	a
30	what	26	not	24	do
22	this	21	you	20	we
19	with	19	in	15	was
15	so	13	they	13	how
13	have	13	actually	12	are
12	an	11	up	11	like
11	had	10	on	10	I
10	but	9	when	8	problem
8	out	8	or	8	for
8	be	7	wrong	7	here
7	from	7	fish	7	does
7	did	7	as	6	working
6	know	6	get	6	asked
6	about	5	would	5	women
5	will	5	using	5	trying
5	things	5	these	5	then
5	them	5	solve	5	robot
5	point	5	legs	5	just
5	into	5	if	5	give
5	flavors	5	can	5	by
4	why	4	who	4	want
4	use	4	us	4	trucks
4	trained	4	thing	4	there
4	tell	4	technically	4	some
4	see	4	really	4	paint
4	new	4	its	4	identify
4	human	4	has	4	happened
4	got	4	going	4	content
4	b	4	all	3	work
3	words	3	where	3	walk
3	very	3	understand	3	turns
3	tench	3	something	3	set
3	seen	3	pictures	3	over



Freq.	Palavra	Freq.	Palavra	Freq.	Palavra
3	move	3	might	3	maybe
3	lot	3	learn	3	kind
3	ice	3	humans	3	highway
3	hard	3	fingers	3	driving
3	doing	3	cream	3	could
3	concept	3	car	3	came
3	because	3	avoid	3	at
3	algorithm	3	ais	3	accidentally
2	were	2	weirder	2	weird
2	way	2	truck	2	train
2	tower	2	today	2	those
2	thank	2	take	2	supposed
2	step	2	single	2	simulation
2	should	2	self	2	say
2	resumes	2	recommend	2	recognize
2	real	2	question	2	program
2	present	2	prepared	2	power
2	picture	2	people	2	pedestrian
2	past	2	parts	2	particular
2	part	2	often	2	now
2	much	2	most	2	may
2	make	2	looks	2	looking
2	little	2	less	2	learned
2	kinds	2	itself	2	happen
2	group	2	gone	2	goes
2	goal	2	generate	2	floor
2	figure	2	fast	2	experiment
2	example	2	enough	2	design
2	day	2	data	2	danger
2	copy	2	colors	2	collection
2	brains	2	assemble	2	asking
2	ask	2	artificial	2	anything
2	any	2	another	2	against
1	youtube	1	yes	1	writing
1	worm	1	world	1	word
1	whole	1	went	1	well
1	wanted	1	walks	1	views
1	via	1	usually	1	used
1	until	1	unfortunately	1	underneath
1	twitching	1	try	1	trophy
1	trick	1	trial	1	training



Freq.	Palavra	Freq.	Palavra	Freq.	Palavra
1	traditional	1	together	1	tiny
1	time	1	through	1	thought
1	though	1	think	1	therefore
1	theories	1	themselves	1	their
1	than	1	textures	1	tesla
1	terminator	1	tends	1	teamed
1	team	1	task	1	tall
1	talk	1	super	1	style
1	strict	1	streets	1	sorting
1	somewhere	1	somersaulting	1	somebody
1	solves	1	society	1	soccer
1	smart	1	simple	1	silly
1	sign	1	side	1	seeing
1	science	1	school	1	saw
1	safe	1	run	1	robots
1	road	1	resumeas	1	resume
1	researchers	1	repeatedly	1	recommending
1	recognized	1	recognition	1	recently
1	rebel	1	realize	1	reach
1	problemtechnically	1	problems	1	probably
1	plenty	1	own	1	our
1	otherwise	1	other	1	original
1	optimized	1	opinion	1	ones
1	one	1	old	1	okay
1	obstacles	1	obstacle	1	obey
1	number	1	no	1	nice
1	nearly	1	nature	1	names
1	my	1	movies	1	more
1	mind	1	middle	1	mess
1	measure	1	mean	1	matrix
1	math	1	many	1	looked
1	list	1	lines	1	limits
1	likely	1	life	1	letter
1	let	1	left	1	learning
1	lands	1	known	1	knowing
1	kill	1	kealing	1	job
1	invent	1	intelligenceit	1	intelligence
1	instructions	1	instead	1	industries
1	increase	1	imitate	1	image
1	hoped	1	honeybee	1	hired
1	highlighted	1	heap	1	he



Freq.	Palavra	Freq.	Palavra	Freq.	Palavra
1	harvest	1	happens	1	hacking
1	hack	1	ha	1	goals
1	glitching	1	given	1	gave
1	front	1	found	1	forward
1	force	1	find	1	figured
1	field	1	fiction	1	fed
1	fatal	1	faster	1	falls
1	fall	1	fair	1	failures
1	failed	1	facing	1	facebook
1	expect	1	existing	1	examples
1	exactly	1	errors	1	error
1	entire	1	engineers	1	energy
1	end	1	else	1	either
1	easy	1	earthworm	1	during
1	drove	1	drive	1	down
1	disrupting	1	discriminate	1	discovered
1	differently	1	different	1	destructive
1	designing	1	designed	1	delicious
1	definitely	1	decided	1	david
1	course	1	controlled	1	constantly
1	conspiracy	1	consequences	1	confused
1	computing	1	computer	1	competent
1	communication	1	communicate	1	common
1	come	1	combinations	1	colorsto
1	color	1	colleges	1	collected
1	coders	1	clicks	1	clear
1	city	1	chance	1	cars
1	capable	1	cannot	1	called
1	brake	1	brain	1	blowing
1	bigotry	1	big	1	beyond
1	being	1	behind	1	been
1	becomes	1	avoiding	1	autopilot
1	assembles	1	arms	1	approximate
1	anymore	1	answer	1	amazon
1	always	1	along	1	allowed
1	age	1	advanced	1	accident