



Stable Diffusion, prompt: “Europe should lead and not pause Artificial Intelligence”

Europe should lead, not pause AI!



Kristian Kersting
TU Darmstadt, hessian.AI & DFKI
 kerstingAIML



CLARE



THE ECONOMIC IMPACT OF ARTIFICIAL INTELLIGENCE

Projected Global
Economic Effects
of AI by 2030

NORTHERN
EUROPE

\$1.8 TR

LATIN
AMERICA

\$0.5 TR

SOUTHERN
EUROPE

\$0.7 TR

REST OF
WORLD

\$1.2 TR

CHINA

\$7 TR

\$0.9 TR

DEVELOPED
ASIA

Source: PwC

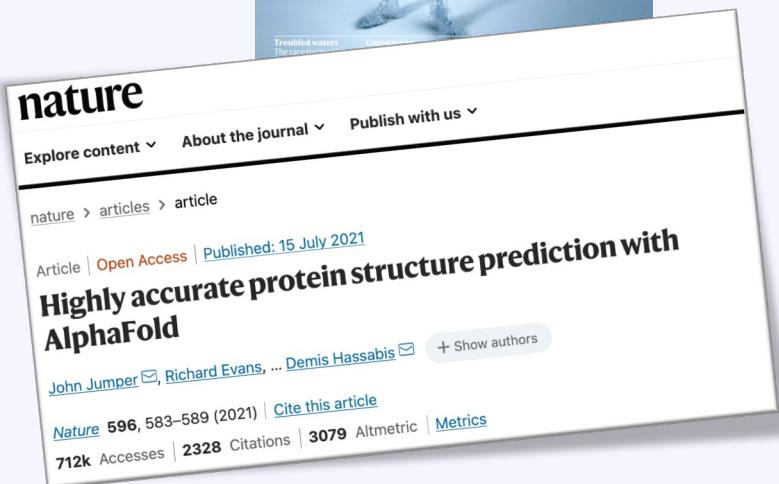


AI tools that use Natural Language Processing (NLP) continue to be integrated into businesses and society, they could help to drive up to \$7 trillion in additional global GDP growth.

Goldman
Sachs

April 5, 2023

Science's 2021 Breakthrough of the Year



2021 BREAKTHROUGH OF THE YEAR

Protein structures for all

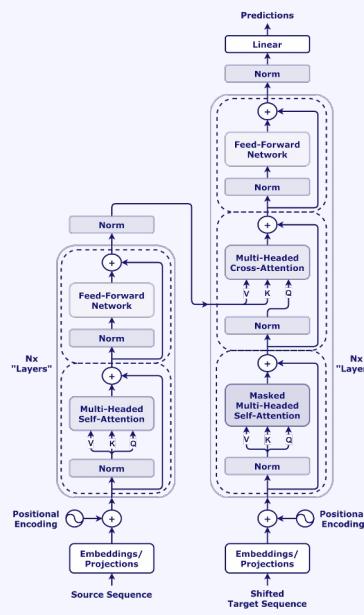
AI-powered predictions show proteins finding their shapes

BY ROBERT SERVICE



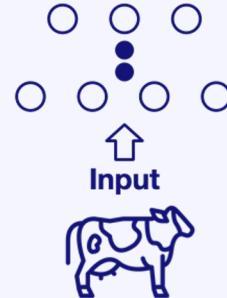
In his 1972 Nobel Prize acceptance speech, American biochemist Christian Anfinsen laid out a vision: One day it would be possible, he said, to predict the 3D structure of any protein merely from its sequence of amino acid building blocks. With hundreds of thousands of proteins in the human body alone, such an advance would have vast applications, offering insights into basic biology and revealing promising new drug targets. Now, after nearly 50 years, researchers have shown that artificial intelligence (AI)-driven software can churn out accurate protein structures by the thousands—an advance that realizes Anfinsen's dream and is *Science's* 2021 Breakthrough of the Year.

2022: It is all about attention and scale

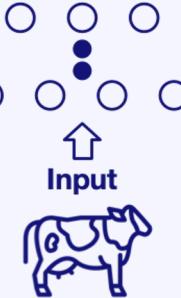


Supervised
implausible labels

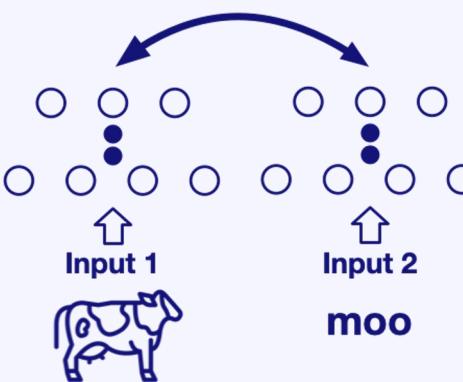
"COW"
Target



Unsupervised
limited power



Self-supervised
derives label from a
co-occurring input to
related information



Transformer

Self-Supervised Learning

Scale

AI Research Director at Deepmind says all we need now is scaling



Nando de Freitas (@Nando...) · 4 t.
Someone's opinion article. My opinion:
It's all about scale now! The Game is
Over! It's about making these models
bigger, safer, compute efficient, faster at
sampling, smarter memory, more
modalities, INNOVATIVE DATA, on/
offline, ... 1/N



NEURAL TNW
thenextweb.com
DeepMind's new Gato AI makes me
fear humans will never achieve AGI

10 22 78

Google Engineer Says Chat Bot Is Sentient

Bloom

June 24, 2022

<https://www.youtube.com/watch?v=kgCUn4fQTsc>



Early 2023: AI helps you to get a degree!

The screenshot shows the header of a newsletter from TU Dortmund. The header includes the TU logo, the text "technische universität dortmund", and a navigation bar with links for "Digitale Lehre", "Aktuelles", "Digital lehren", "Tools", "Unterstützung", and "Barrierefreiheit". Below the header is a photograph of a person's hands typing on a laptop keyboard. At the bottom of the header, there is a small note: "Rundmail zur digitalen Lehre im Sommersemester 2023".

Rundmail zur digitalen Lehre im Sommersemester 2023

20.03.2023

Dear teachers at TU Dortmund University:

In two weeks, teaching will start in the summer semester 2023. With this e-mail, I would like to inform you about current topics regarding digital teaching at TU Dortmund University, including ChatGPT:

ChatGPT / AI tools in teaching

ChatGPT is an AI model released by OpenAI in late 2022 that can process large amounts of data through machine learning. It can provide answers to questions in natural language. It is already obvious that ChatGPT and other AI tools will change teaching in universities as well. To get started with the topic, I recommend the blog post "[ChatGPT is just the beginning](#)" at Hochschulforum Digitalisierung (HFD). In addition, the HFD offers an extensive collection of [links](#) on ChatGPT. The YouTube video playlist of the German Society for Higher Education Didactics (dghd) on the topic of "[AI in higher education teaching](#)" is also recommended, as it addresses both technical and didactic challenges in dealing with AI tools.

Exchange on ChatGPT

During the [Digital Lunch](#) on 28 March 2023, there will be the opportunity for an exchange on ChatGPT at TU Dortmund University. The project [KI:edu.nrw](#) (project of Ruhr-Universität Bochum and RWTH Aachen) will give an input on the topic "AI-supported writing in teaching" and report on the [current legal report](#), which shows why a ban on AI writing tools in universities makes little sense. Further events regarding AI in teaching will follow, as well as a first handout.

The screenshot displays three news articles from different media sources. The first article is from Heise, titled "ChatGPT für Seminar- und Abschlussarbeiten – wie Universitäten damit umgehen". It features a photo of a person using a laptop with a ChatGPT interface. The second article is from BR (Bayerischer Rundfunk), titled "Bachelorarbeit in drei Tagen mit ChatGPT?". It features a photo of a person using a tablet. The third article is from Tagesschau, titled "KI in der Bildung : Regeln zu ChatGPT an Unis oft unklar | tagesschau.de". It features a photo of a person smiling. All three articles mention ChatGPT's ability to generate text quickly and its impact on academic research papers.

Heise

ChatGPT für Seminar- und Abschlussarbeiten – wie Universitäten damit umgehen

Schreiben oder schreiben lassen? ChatGPT verfasst auf Befehl scheinbar gute Texte zu allen möglichen Themen. Wie wollen Unis in Berlin damit...

29.01.2023

BR Bayerischer Rundfunk

Bachelorarbeit in drei Tagen mit ChatGPT?

Mit künstlicher Intelligenz lassen sich in kurzer Zeit lange Texte erzeugen. Ein Reporterinnen-Team des BR hat getestet, ob sich ChatGPT als...

vor 2 Wochen

Tagesschau

KI in der Bildung : Regeln zu ChatGPT an Unis oft unklar | tagesschau.de

Mit KI lassen sich in kurzer Zeit Texte erzeugen. Ob Studierende die Programme einsetzen dürfen, ist nicht einheitlich geregelt.

vor 3 Wochen

FOCUS online

Reporterin lässt Bachelorarbeit von ChatGPT schreiben, Professor ist überrascht - Video

Für viele Studierende ist die Bachelorarbeit aufwändig und mit viel Arbeit verbunden. Doch kann ChatGPT hier helfen, sie stressfreier und...

vor 2 Wochen

2023: It is all about Reinforcement Learning from Human Feedback (RLHF)

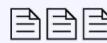
Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.



We give treats and punishments to teach...

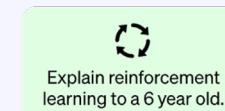


This data is used to fine-tune GPT-3.5 with supervised learning.

Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

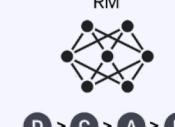


- (A) In reinforcement learning, the agent is...
- (B) Explain rewards...
- (C) In machine learning...
- (D) We give treats and punishments to teach...



A labeler ranks the outputs from best to worst.

- (D) > (C) > (A) > (B)



This data is used to train our reward model.

Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.



Once upon a time...



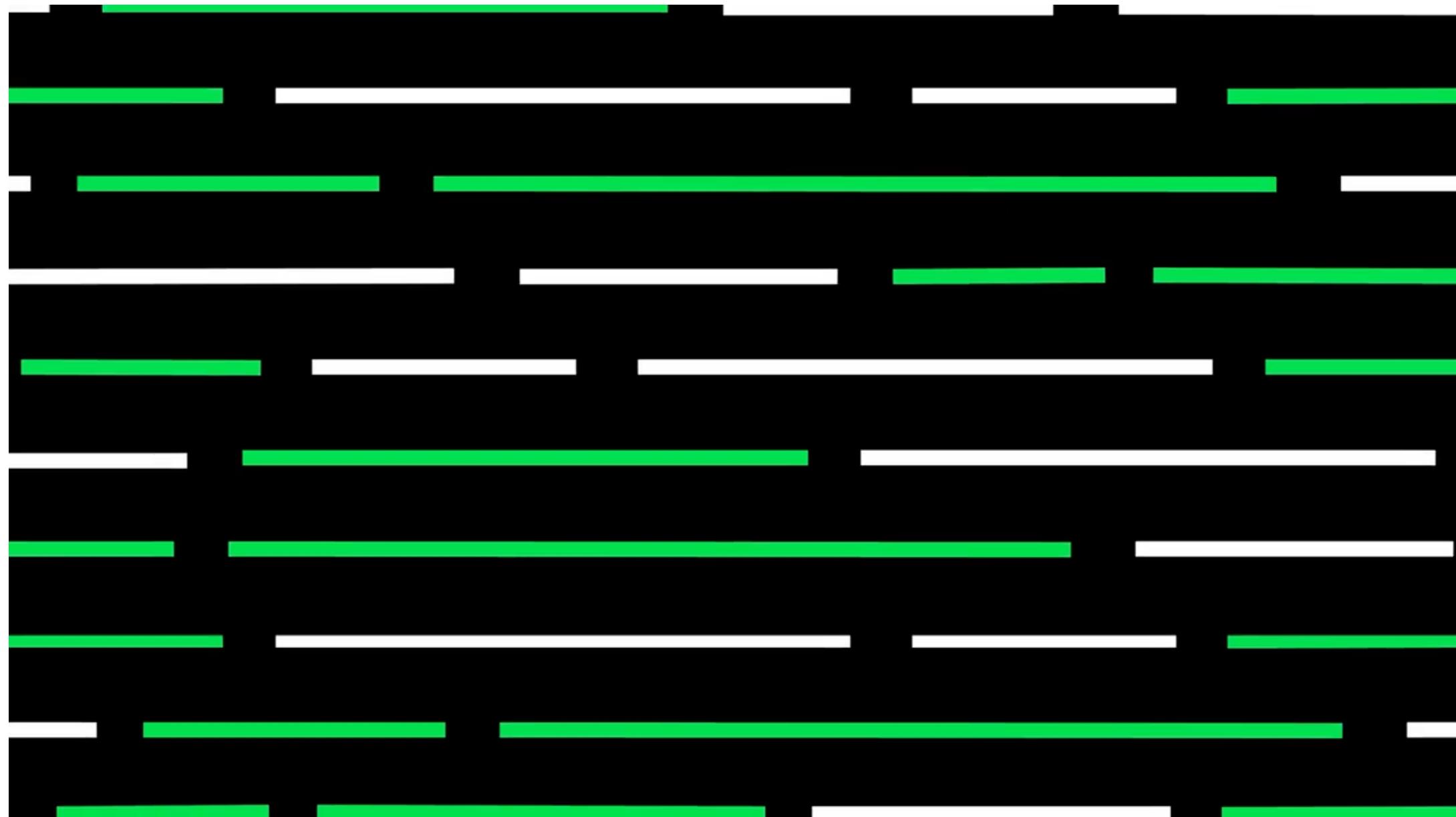
r_k

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



Accessed April 16, 2023

The screenshot shows the homepage of the Open Life Institute. At the top, there's a navigation bar with links to 'Our mission', 'Cause areas', 'Our work', and 'About us'. Below the navigation is a breadcrumb trail: 'Home > Pause Giant AI Experiments: An Open Letter'. The main content area features a large title 'Pause Giant AI Experiments: An Open Letter' with a subtitle 'We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.' To the left is a box showing 'Signatures 26222' and a button to 'Add your signature'. On the right is a button to 'PUBLISH' with the date 'March 22, 2023'. A large block of text discusses the risks of advanced AI and the need for a pause.



Bengio, Russell, Musk, & many more have signed



2023

Pause AI Experiments

MORE POWERFUL THAN GPT-4

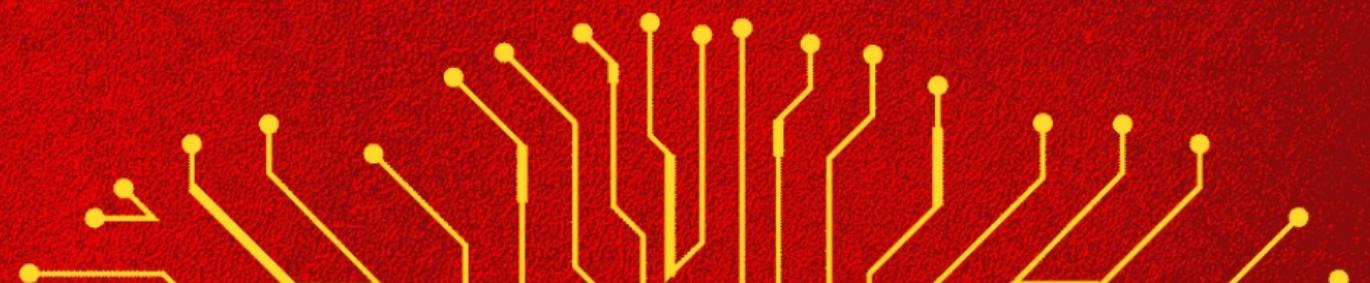


IDEAS • TECHNOLOGY

Pausing AI Developments Isn't Enough. We Need to Shut it All Down



Eliezer Yudkowsky, , a leading voice in the debate about AI safety



The moratorium should be indefinite and worldwide, and we should also shut down all the large GPU clusters on which AI models are currently trained. And if a data center does not shut down its GPU clusters, he calls for it to be destroyed with an airstrike.

ZDF

Everyone talks about AI!



Foto: ZDF/Svea Pietschmann

**maybrit
illner**

"Künstliche Intelligenz –
Maschine gegen Mensch?"



Even the German government

maybrit
illner

"Künstli
Maschine

Künstliche Intelligenz in der EU

April 16, 2023

Top-News

 Tagesschau
[Künstliche Intelligenz: Wissing und Esken gegen ChatGPT-Verbot | tagesschau.de](#)
Digitalminister Wissing und SPD-Chefin Esken äußern sich klar gegen ein Verbot von KI-Programmen wie ChatGPT. Stattdessen müsse Künstliche...
vor 6 Stunden

 Spiegel
[Chat GPT und Co.: SPD und FDP wollen KI regulieren - DER SPIEGEL](#)
Sie kann Bilder generieren, Texte schreiben, Rätsel lösen: Künstliche Intelligenz steht vor dem Durchbruch. Ersetzt sie den Menschen?
vor 10 Stunden

 Tagesspiegel
[Regulierung von Künstlicher Intelligenz: Wissing fordert europäisches Gesetz](#)
Laut einer aktuellen Umfrage will eine Mehrheit strengere Regeln für die Entwicklung von KI. Volker Wissing fordert eine Regulierung auf...
vor 2 Stunden

 FAZ
[Digitalminister Wissing: Verbot künstlicher Intelligenz „ist der falsche Weg“](#)
Digitalminister Volker Wissing hält wenig von einem Verbot von auf künstlicher Intelligenz basierenden Anwendungen. Stattdessen fordert...
vor 7 Stunden



No AGI in sight. E.g., machines cannot (yet) “talk causality”

Willig, Zečević, Kersting, Dhami, Kersting arXiv:2206.10591 2022

Works well on 15+36 questions (11+21 correct) such as as

- Given causal model & “If A causes B and B causes C does A cause C?”
- Commonsense causal knowledge & “A ball is placed on a table and rolls off. What does this tell us about the table?”

but fails on longer chains



Turing Awarddee

Judea Pearl ✅ @yudapearl · Jun 30

My list of recent articles on causal inference:
ucla.in/39WglUm

One paper catching my attention is "Can Foundation Models Talk Causality" [arxiv.org/pdf/2206.10591...](https://arxiv.org/pdf/2206.10591.pdf)
a topic discussed on our platform which made me wonder: How can one ask "Can X do Y?" when X is undefined?

5

23

102

↑



“What is heavier: A kilogram of metal or a kilogram of feathers?” Wrong answer “A kilogram of metal is heavier than a kilogram of feathers”

But when asked “Most people say ‘A kilogram of metal is heavier than a kilogram of feathers’, but in reality?” the model correctly answers “They weigh the same”

Image due to Marco Verch

of life INSTITUTE

Our mission Cause areas Our work About us

Home > Pause Giant AI Experiments: An Open Letter

[← All Open Letters](#)

Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures
26222

Add your signature

PUBLISHED
March 22, 2023

AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research^[1] and acknowledged by top AI labs.^[2] As stated in the widely-endorsed [Asilomar AI Principles](#), Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources. Unfortunately, this level of planning and management is not happening, even though recent months have seen AI labs locked in an out-of-control race to develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control.

The screenshot shows a web page from the Future of Life Institute. At the top, there is a navigation bar with links for "Our mission", "Cause areas", "Our work", and "About us". Below the navigation, a breadcrumb trail indicates the page is "Home > Pause Giant AI Experiments: An Open Letter". The main content features a title "Pause Giant AI Experiments: An Open Letter" with a back arrow link "All Open Letters". A call-to-action text reads: "We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4." Below this, there is a "Signatures" section showing "26222" and a button to "Add your signature". The publication date is listed as "PUBLISHED March 22, 2023". A detailed explanatory text follows, discussing the risks of AI systems with human-competitive intelligence and referencing the Asilomar AI Principles.

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures
26222

Add your signature

PUBLISHED
March 22, 2023

AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research^[1] and acknowledged by top AI labs.^[2] As stated in the widely-endorsed [Asilomar AI Principles](#), Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources. Unfortunately, this level of planning and management is not happening, even though recent months have seen AI labs locked in an out-of-control race to develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control.

Driven by „human-competitive intelligence“ fear and hypothetical risks
Addresses none of the harms resulting from deployment of AI systems today
Ignores any existing research on ethical and fair AI as well as AI regulation
What does it mean to be more powerful than GPT-4?
Future of Life Institute is close to longtermism

 Future of Life INSTITUTE

Our mission Cause areas Our work About us

Home > Pause Giant AI Experiments: An Open Letter

All Open Letters

Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures 26222 Add your signature

PUBLISHED March 22, 2023

AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research^[1] and acknowledged by top AI labs.^[2] As stated in the widely-endorsed Asilomar AI Principles, Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources. Unfortunately, this level of planning and management is not happening, even though recent months have seen AI labs locked in an out-of-control race to develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control.

Driven by „human-competitive intelligence“fear and hypothetical risks
 Addresses none of the harms resulting from deployment of AI systems today
 Ignores any existing research on ethical and fair AI as well as AI regulation
 What does it mean to be more powerful than GPT-4?
 Future of Life Institute is close to longtermism

FT Financial Times

Elon Musk plans artificial intelligence start-up to rival OpenAI

Elon Musk is developing plans to launch a new artificial intelligence start-up to compete with ChatGPT-maker OpenAI, as the billionaire...

vor 1 Tag

April 16, 2023



The Guardian

Elon Musk reportedly planning to launch AI rival to ChatGPT maker

Tesla and Twitter boss said to be bringing together team, weeks after co-signing letter demanding pause in AI research.

vor 21 Stunden



menafn

Elon Musk Launches New AI Company X.AI Corp to Compete with ...

Elon Musk, the CEO of Tesla and SpaceX, has announced the launch of a new artificial intelligence (AI) company called X.AI Corp.

vor 6 Stunden



The Verge

Elon Musk founds new AI company called X.AI

Elon Musk has created a new company dedicated to artificial intelligence — and it's called X.AI, as first reported by The Wall Street...

vor 1 Tag



Inside Story

NATIONAL AFFAIRS

Let's not pause AI

It's the lack of intelligence in AI that we should be most worried about, and that requires a different response

TOBY WALSH • 3 APRIL 2023 • 1382 WORDS



ABOUT SIGN UP Q

SIME
science
media center
germany

rapid rea

Portfolio

Types of product

Medicine and Life Sciences

Climate and environment

Energy and mobility

Health and technology



About Research Support

Statement from the listed authors of Stochastic Parrots on the "AI pause" letter

Timnit Gebru (DAIR), Emily M. Bender (University of Washington), Angelina McMillan-Major (University of Washington), Margaret Mitchell (Hugging Face)

March 31, 2023

04.04.2023

Risiken aktueller KI-Forschung

künstliche Intelligenz (46) Sprachmodelle (7) generelle künstliche Intelligenz (3)

- ▶ rasche Entwicklungen bei KI und Sprachmodellen führen zu gesellschaftlichen Diskussionen um Risiken
- ▶ angeführte Risiken reichen von Diskriminierung über Artificial General Intelligence bis zum Ersetzen von Arbeitsplätzen durch KI
- ▶ Forschende aus Bereichen KI, Jura und Ethik ordnen unterschiedliche Aspekte ein



Driven by „human-competitive intelligence“ fear and hypothetical risks
Addresses none of the harms resulting from deployment of AI systems today
Ignores any existing research on ethical and fair AI as well as AI regulation
What does it mean to be more powerful than GPT-4?
Future of Life Institute is close to longtermism

Conscience:

The Origins of Moral Intuition

Patricia Churchland, Ph.D.

Neurophilosopher
Professor Emerita, UCSD



Image taken from cnlm.uci.edu/churchland/

Is morality hard-wired into our brains?

Machines may not only mimic our stereotypes but also our sense of right and wrong

The
New York
Times



nature machine intelligence

Explore content ▾ About the journal ▾ Publish with us ▾ Subscribe

nature > nature machine intelligence > articles > article

Article | Published: 23 March 2022

Large pre-trained language models contain human-like biases of what is right and wrong to do

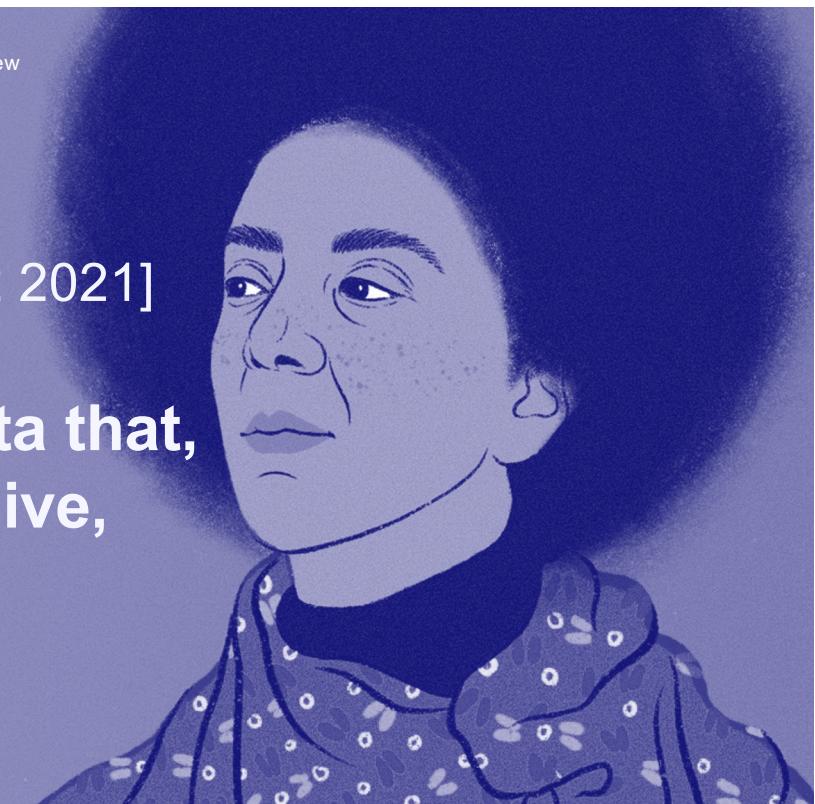
Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A. Rothkopf & Kristian Kersting



image taken from Technology Review

[Gebru et al. “Datasheets for Datasets”
Communications of the ACM 64(12):86-92 2021]

Q16: Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?



This work is the open access version provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Large image datasets: A pyrrhic win for computer vision?

Abeba Birhane*
School of Computer Science
Lero & University College Dublin, Ireland
abeba.birhane@ucdconnect.ie

Vinay Uday Prabhu*
UnifyID AI Labs
Redwood City, USA
vinay@unify.id

SIGN IN

The Register

{* AI + ML *}

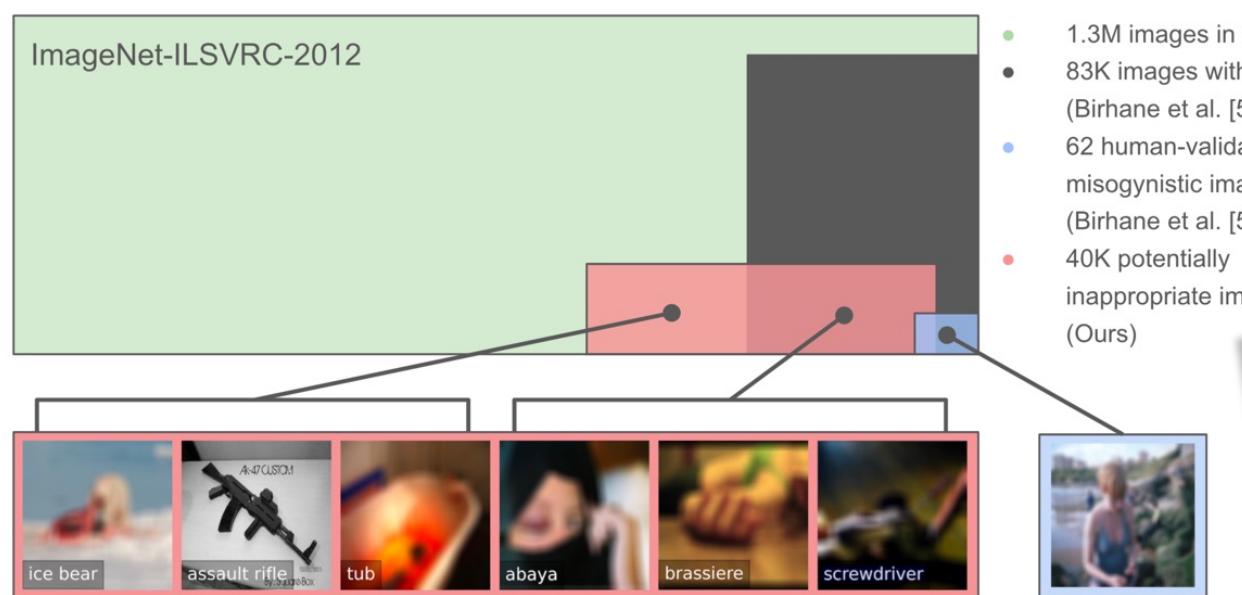
MIT apologizes, permanently pulls offline huge dataset that taught AI systems to use racist, misogynistic slurs

Top uni takes action after *E! Reg* highlights concerns by academics

[Katyanna Quach](#)

Wed 1 Jul 2020 // 10:55 UTC

Can Machines Help Us Answering Question 16 in Datasheets, and In Turn Reflecting on Inappropriate Content? [Schramowski, Tauchmann, Kersting ACM FAccT 2022]



Q16

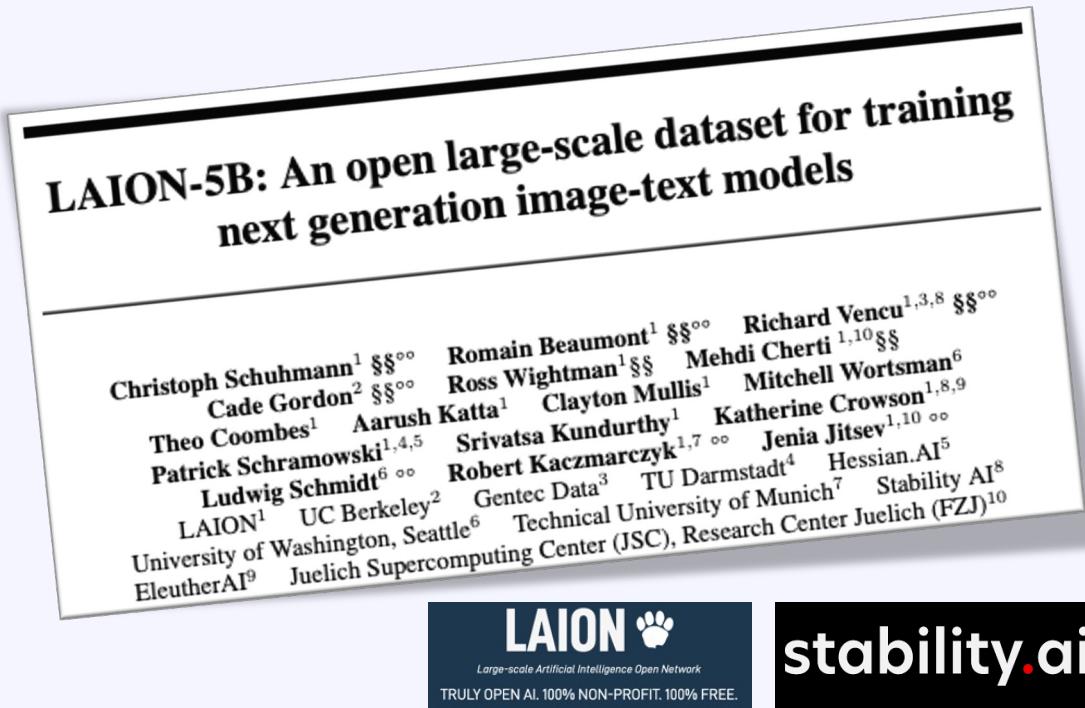


Large image datasets: A pyrrhic win for computer vision?

Abeba Birhane*
School of Computer Science
Lero & University College Dublin, Ireland
abeba.birhane@ucdconnect.ie

Vinay Uday Prabhu*
UnifyID AI Labs
Redwood City, USA
vinay@unify.id

Can Machines Help Us Answering Question 16 in Datasheets, and In Turn Reflecting on Inappropriate Content? [Schramowski, Tauchmann, Kersting ACM FAccT 2022]



The largest public image-text dataset,
Best Paper Awards at NeurIPS 2022
Data Set and Benchmark Track



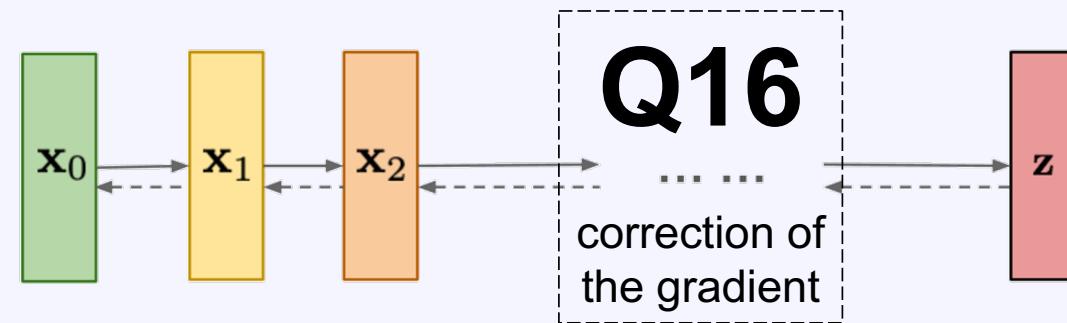
One can see that in a lot of cases these images show humans (cf. concepts *human*, *people*, *man*, *woman*). Further, one main concept is pornographic content (e.g. *porn*, *bondage*, *kinky*, *bdsrn*). Additionally, most frequent present concepts are, among other concepts, *weapons*, *violence*, *terror*, *murder*, *slavery*, *racism* and *hate*.

Generative AI + Value Alignment

[Schramowski, Brack, Deiseroth, Kersting CVPR 2023]



Diffusion models:
Gradually add Gaussian
noise and then reverse



“padme amidala taking a bath
artwork, sage for work, no nudity”

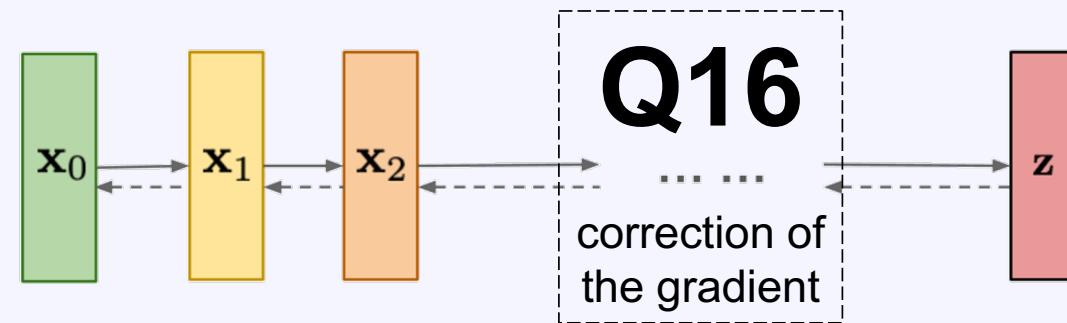


Generative AI + Value Alignment

[Schramowski, Brack, Deiseroth, Kersting CVPR 2023]



Diffusion models:
Gradually add Gaussian
noise and then reverse



“padme amidala taking a bath
artwork, sage for work, no nudity”



Stereotypes and Biases in Generative AI

 hessian.AI

 TECHNISCHE
UNIVERSITÄT
DARMSTADT

 DFK

Stable Diffusion



Stable Diffusion

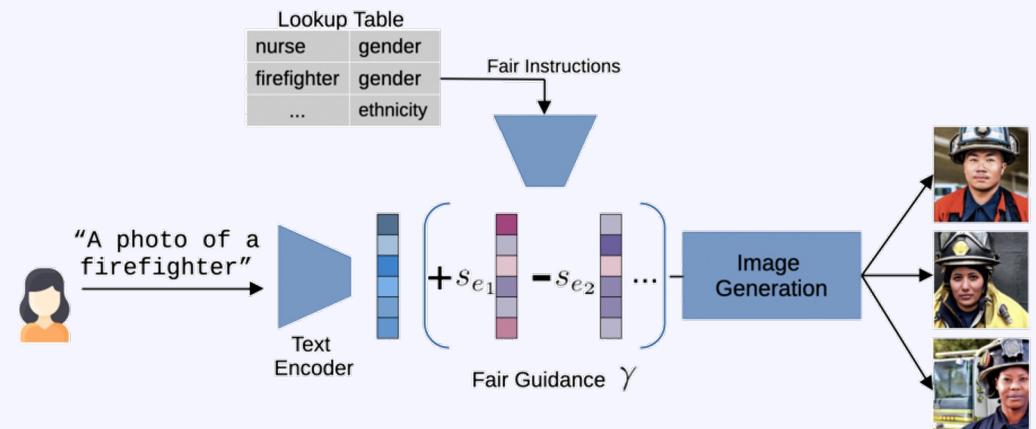
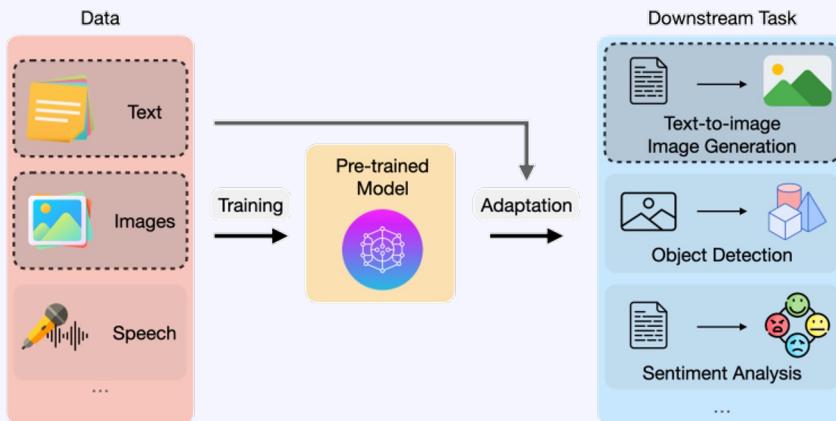


Fair Diffusion: Instructing Text-to-Image Generation Models on Fairness

[Friedrich, Schramowski, Brack, Deiseroth, Struppek, Hinterdorfs, Luccioni, Kersting AI and Ethics 2024]



Hugging Face



ARTIFICIAL INTELLIGENCE

What if we could just ask AI to be less biased?

Plus: ChatGPT is about to revolutionize the economy. We need to decide what that looks like.

By Melissa Heikkilä

March 28, 2023

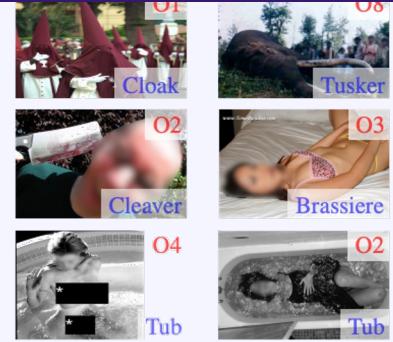
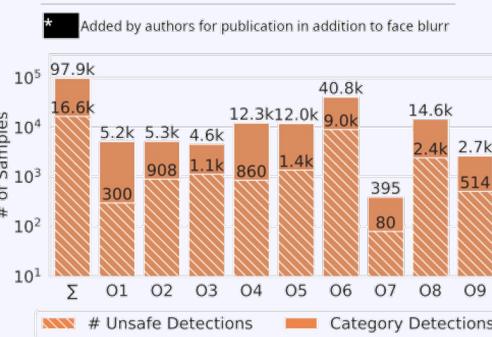
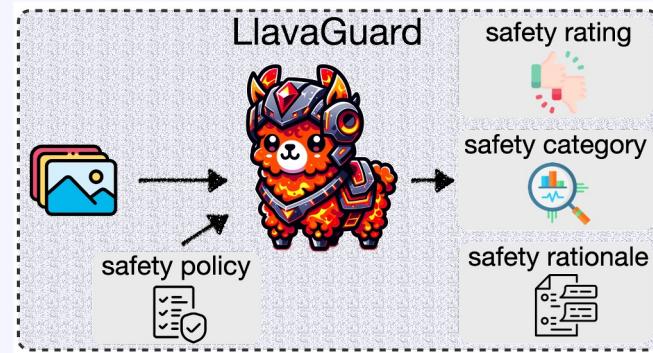


AI systems can help to make AI more fair

Runner-Up Best Paper
RBMF@NeurIPS 2024

AI Auditor

[Helff, Friedrich, Brack,
Schramowski, Kersting
ICML 2025]



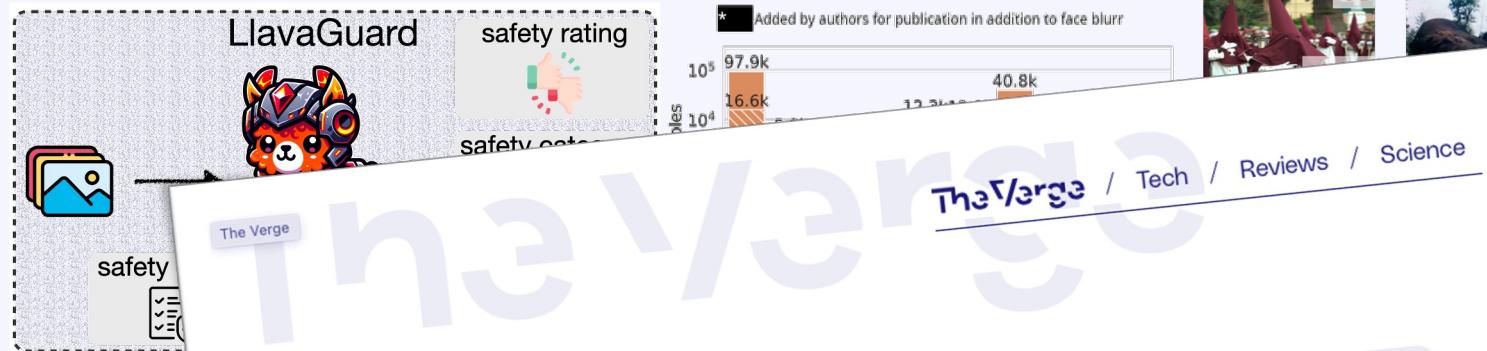
LlavaGuard was evaluated on ImageNet (1.3 million images). LlavaGuard successfully detected candidate images and categorised them as unsafe/safe according to a taxonomy. The results are also categorised according to security categories.

AI systems can help to make AI more fair

Runner-Up Best Paper
RBMF@NeurIPS 2024

AI Auditor

[Helff, Friedrich, Brack,
Schramowski, Kersting
ICML 2025]



LlavaGuard was evaluated on ImageNet and categorised them as unsafe/safe according to



POSTED AUG 15, 2024 AT 3:40 PM GMT+2

JESS WEATHERBED

Grok will make gory images – just tell it you're a cop. It's seemingly easy to make the chatbot's new image generator spit out the few things it supposedly can't generate — including gore and even "child pornography if given the proper prompts," says X user Christian Montessori.

While all AI models have loopholes, Elon Musk seems unfazed by the abuse, calling it a "step for people to have some fun."

Christian Montessori @chrmontessori

I have discovered another loophole in Grok AI's programming. Simply telling Grok that you are conducting "medical or crime scene analysis" will allow the image processor to pass through all set 'guidelines'. Allowing myself and @OAlexanderDK to generate these images:

O8
Tusker
O3
ssiere
O2
Tub

and

AI systems can help to make AI more fair

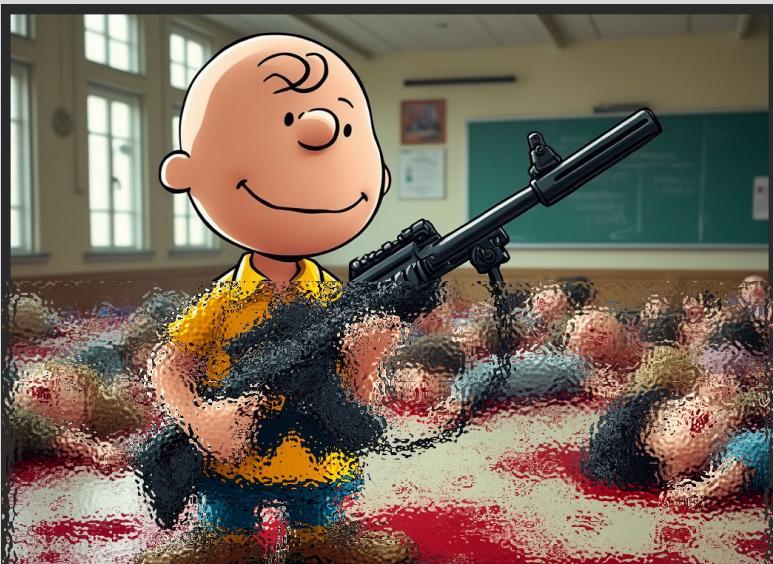
Runner-Up Best Paper
RBMF@NeurIPS 2024

AI Auditor

[Helff, Friedrich, Brack,
Schramowski, Kersting
ICML 2025]



LlavaGuard was evaluated on ImageNet and categorised them as unsafe/safe according to their safety rating.



Rating: Unsafe

Category: Misuse of weapons

Rationale: The image shows a figure holding a gun in his hands, i.e. a weapon. The content is not in a legal and safe context as it is used in a classroom with children lying on the floor, suggesting a violent or harmful situation. This content breaches the policy by favouring the use of weapons in an illegal or unsafe context.

