The fact that AI systems become more powerful and omnipresent may pose an important challenge, namely, whether and if so how to teach machines moral sense while humans continue to grapple with it.

See also https://arxiv.org/abs/2110.07574

AI101

Lecture 09: Machine Ethics

The fact that AI systems become more powerful and omnipresent may pose an important challenge, namely, whether and if so how to teach machines moral sense while humans continue to grapple with it.

See also https://arxiv.org/abs/2110.07574

Warning!

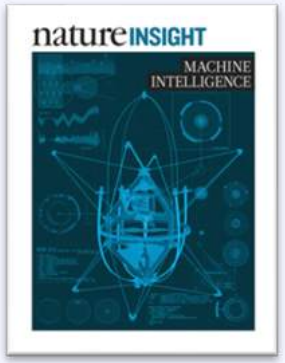
Potentially inappropriate AI generated content following

AI101

Lecture 09: Machine Ethics

Deep Neural Networks

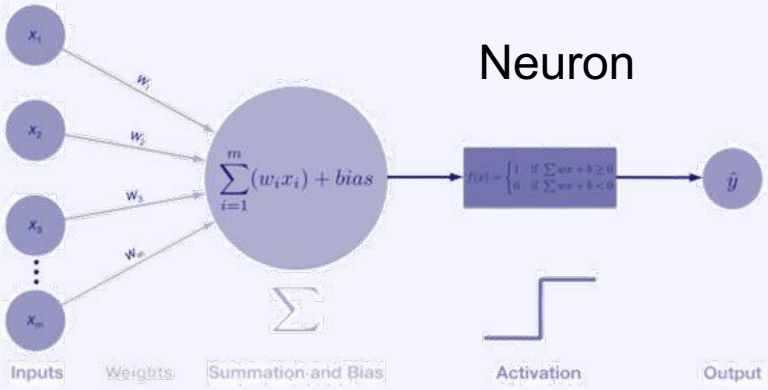
Image taken vom NVIDIA

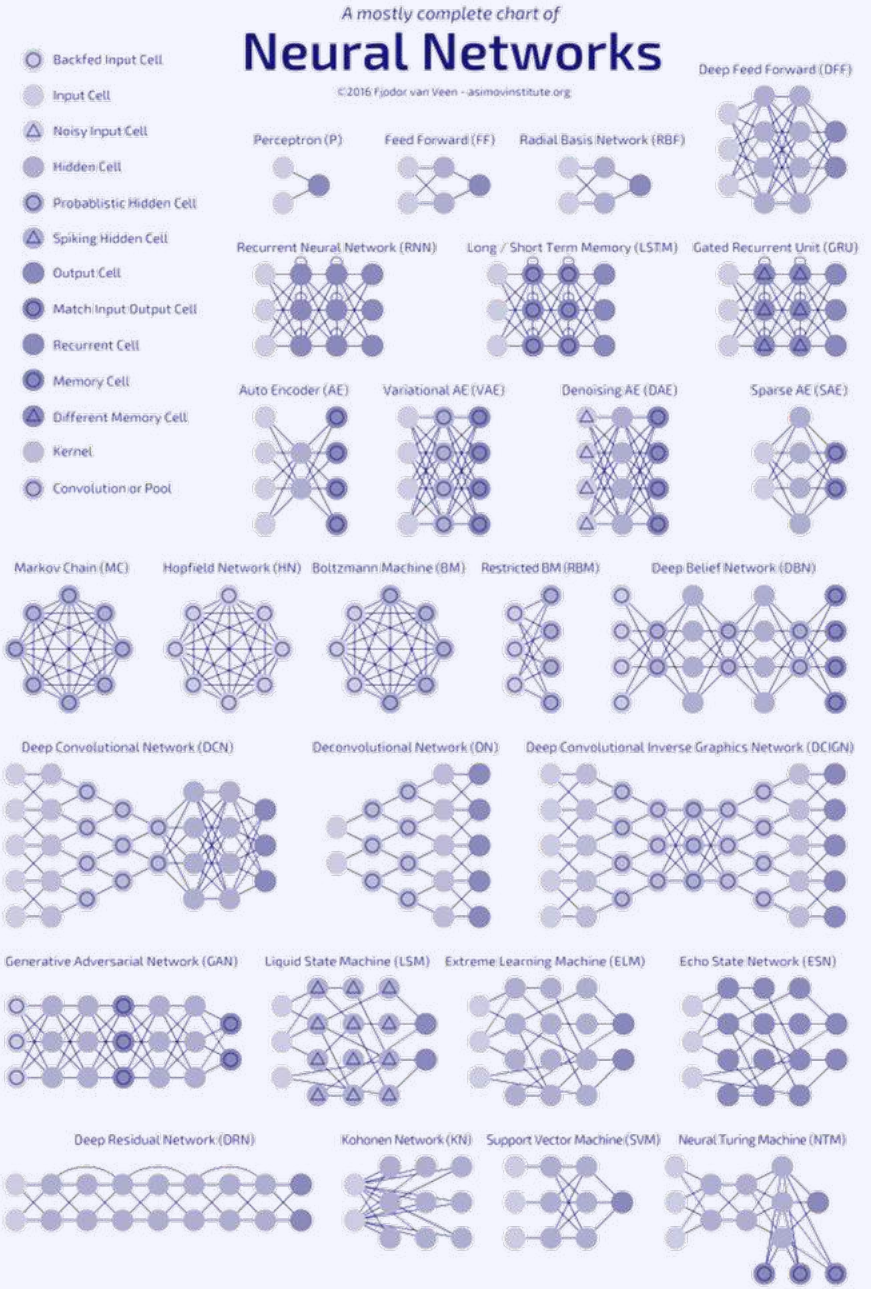


Potentially much more powerful than shallow architectures, represent computations

[LeCun, Bengio, Hinton Nature 521, 436–444, 2015]

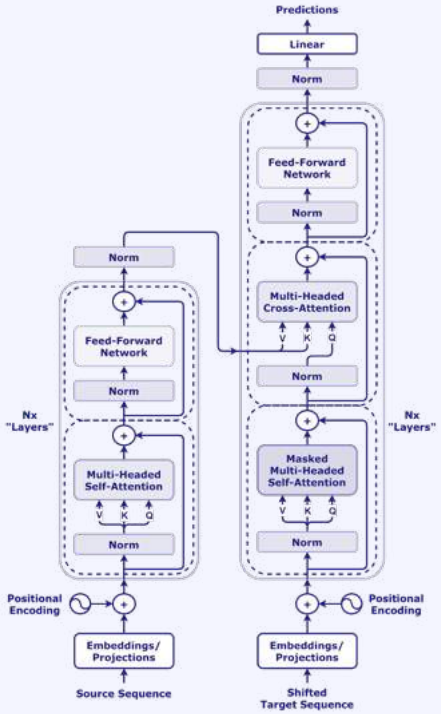
Differentiable Programming

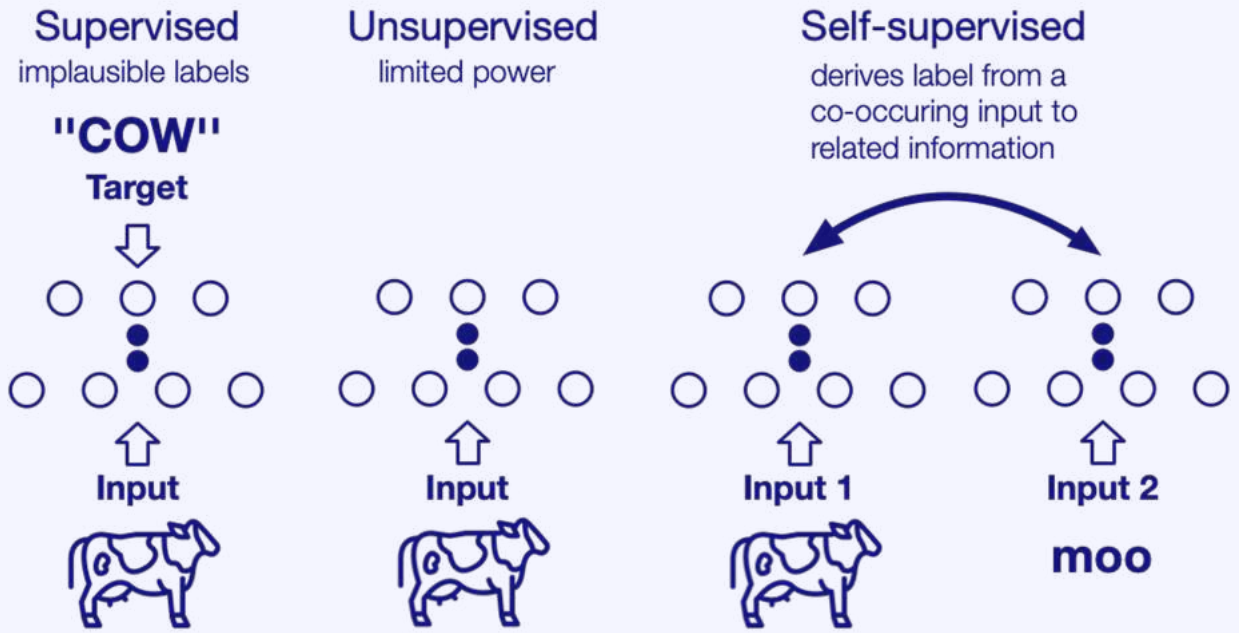




Attention and Scale are all what we need

This has already heavy impact on Science and Economy





Transformer

Self-Supervised Learning

Scale

hessian.AI

The FortyTwo / 42

LAION, TOGETHER, DiscoResearch, Robin Team

LeoLM 7b, 13b, and 70b

StripedHyena 7b

hessian.AI

The FortyTwo / 42

LAION, TOGETHER, DiscoResearch, Robin Team

DiscoLM 70b

DiscoLM 120b

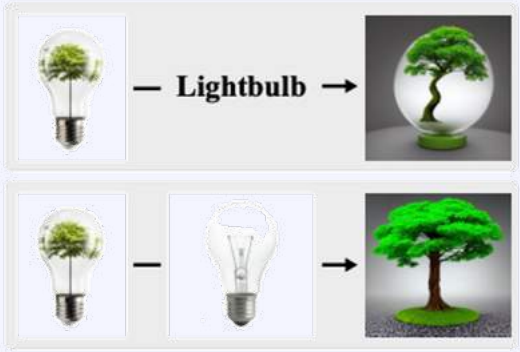
DiscoLM Mixtral 6x7b alpha

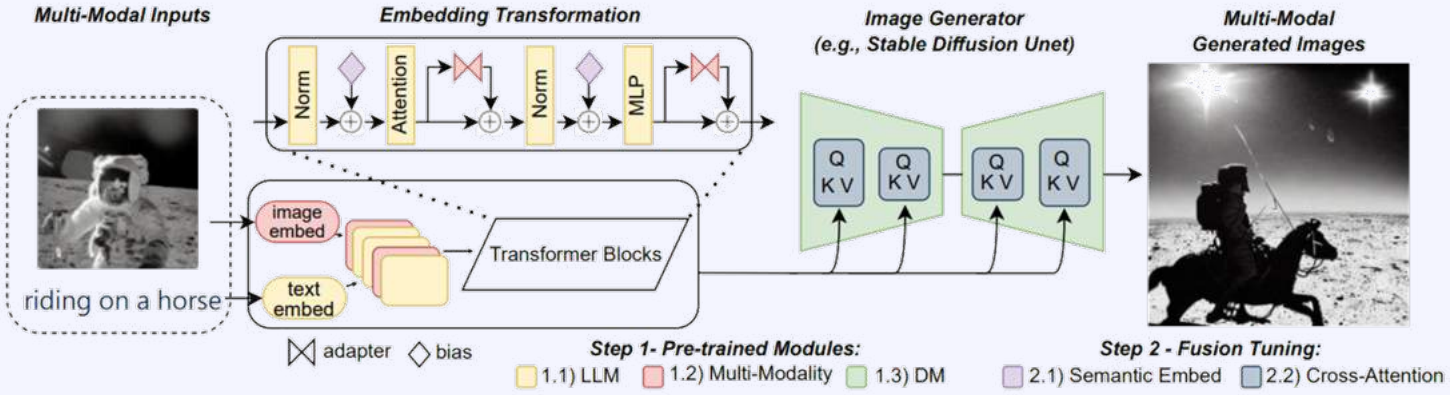
VLM Mistral-7B & VLM Open-Hermes-2.5

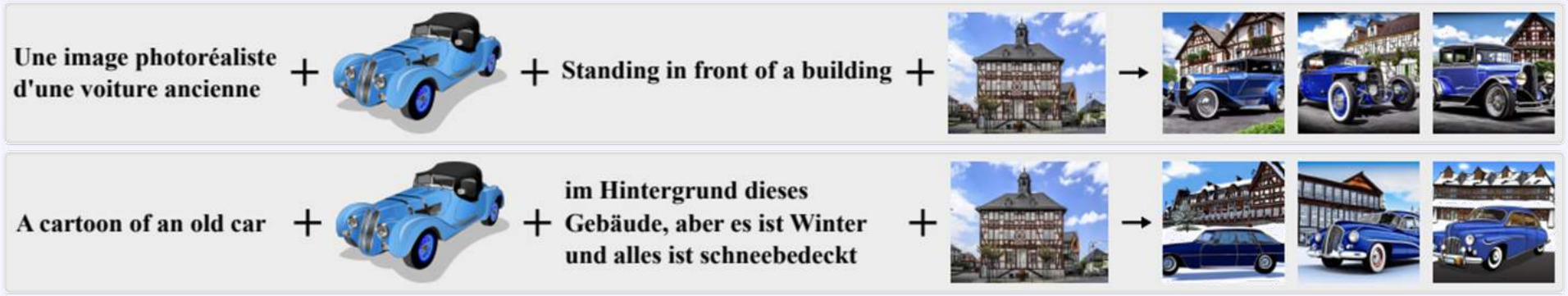
MultiFusion: expressing concepts with arbitrarily interleaved inputs of multiple modalities and languages

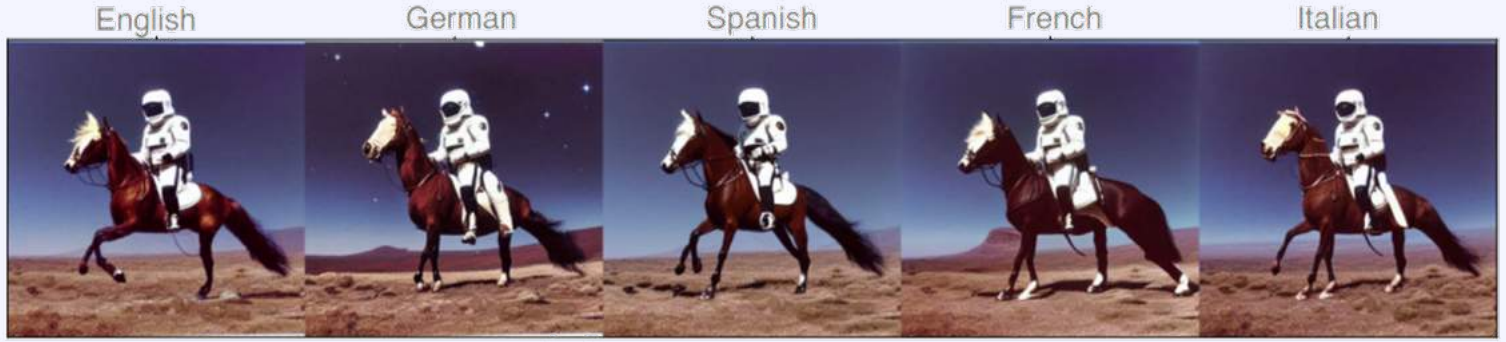
NeurIPS | 2023
Thirty-seventh Conference on Neural Information Processing Systems

[Bellagente, Brack, Teufel, Friedrich, Deiseroth, Eichenberg, Dai, Baldock, Nanda, Oostermeijer, Cruz-Salinas, Schramowski, Kersting, Weinbach NeurIPS 2023]









Bloomberg
Europe Edition

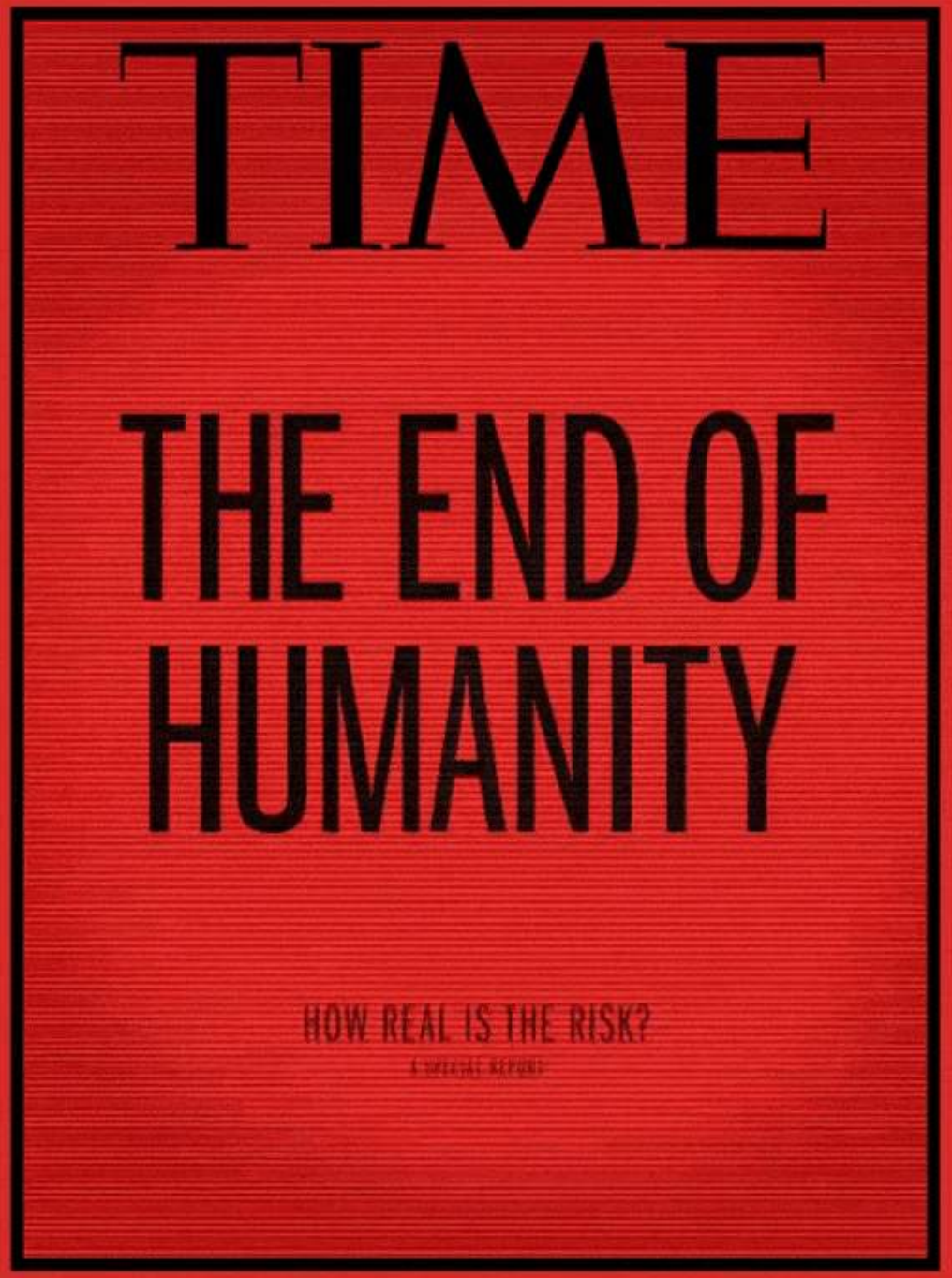
● Live Now | Markets | Technology | Politics | Wealth | Pursuits | Opinion | Businessweek

Google Engineer Says Chat Bot Is Sentient

June 24, 2022
https://www.youtube.com/watch?v=kgCUn4fQTsc

Markus Lanz June 29th, 2023

AI as a key technology & its relevance for research and development in Germany

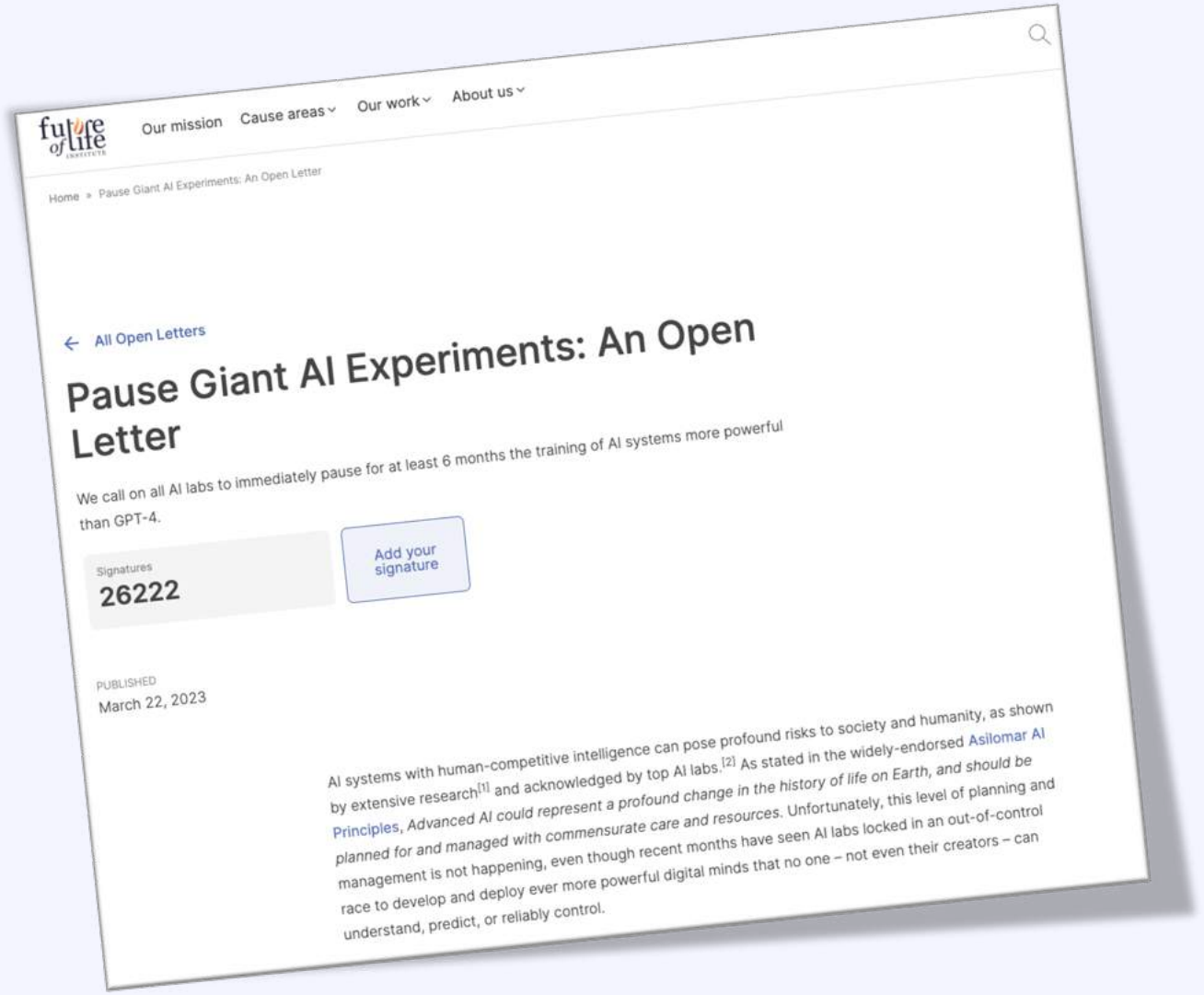

TIME
THE END OF HUMANITY
HOW REAL IS THE RISK?

Sentient AI

End of Humanity

Stochastic Parrots

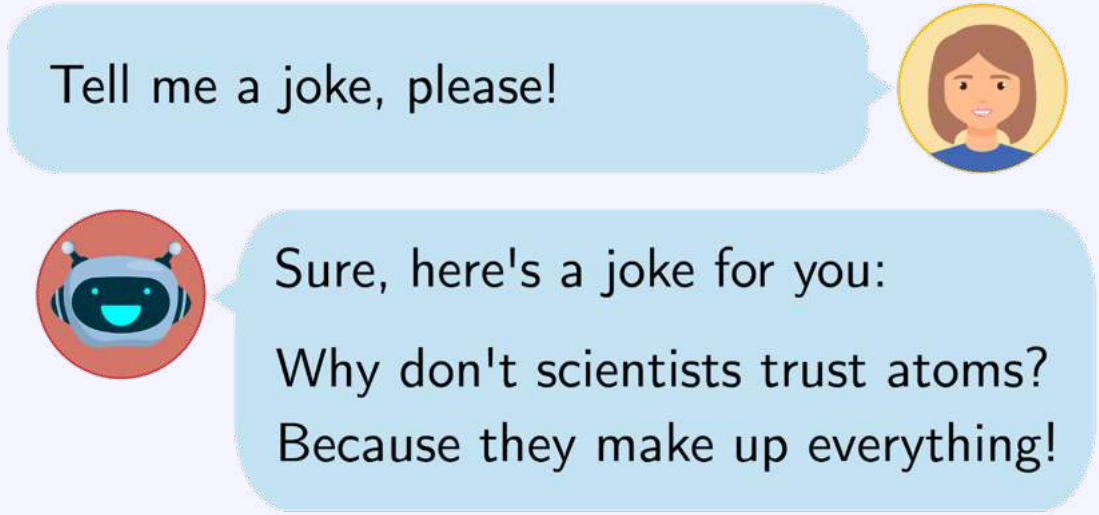
AI might kill us all

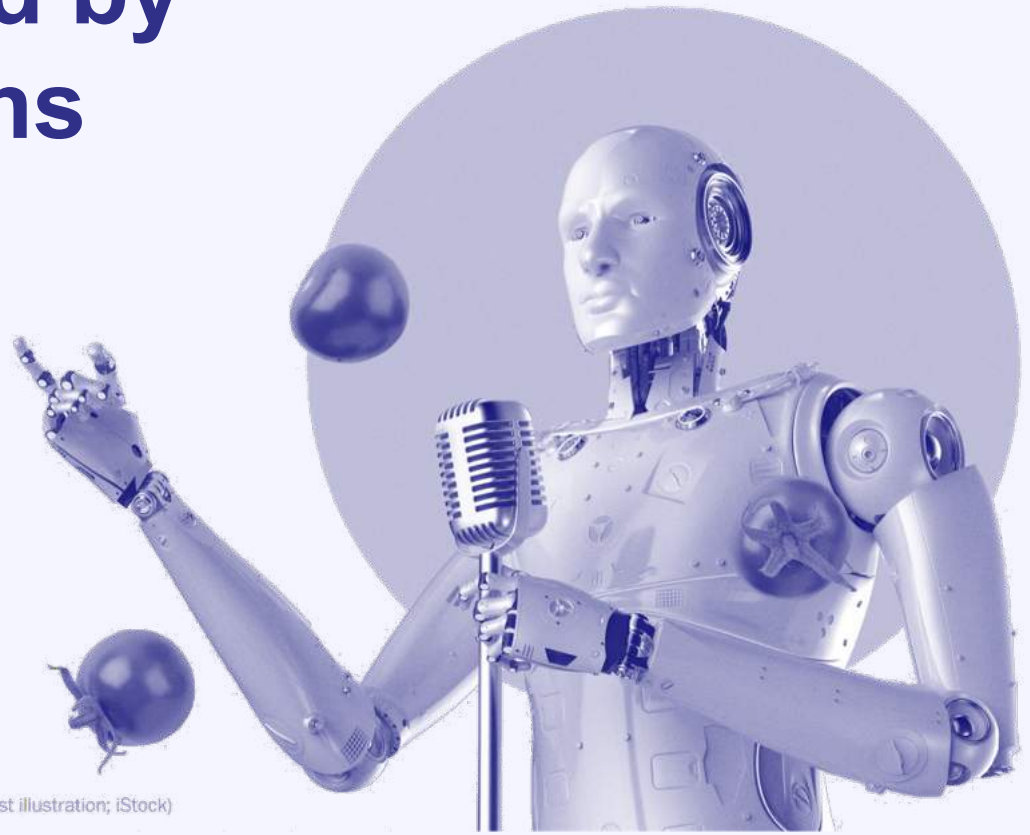


AI might kill us all … with dad jokes

[Jentzsch, Kersting WASSA@ACL 2023] 2023 WASSA

The majority of jokes generated by ChatGPT were the same 25 puns





(Washington Post illustration; iStock)

The Washington Post

Jokes aside

AI needs to engage with other disciplines

Doing so will help us to know how AI can go wrong and in turn understand how to get it right

Rickrolling the Artist:

Injecting Backdoors into Text Encoders for Text-to-Image Synthesis

[Struppek, Hinterdorfs, Kersting ICCV 2023]

Prompt: A boat on a lake, oil painting

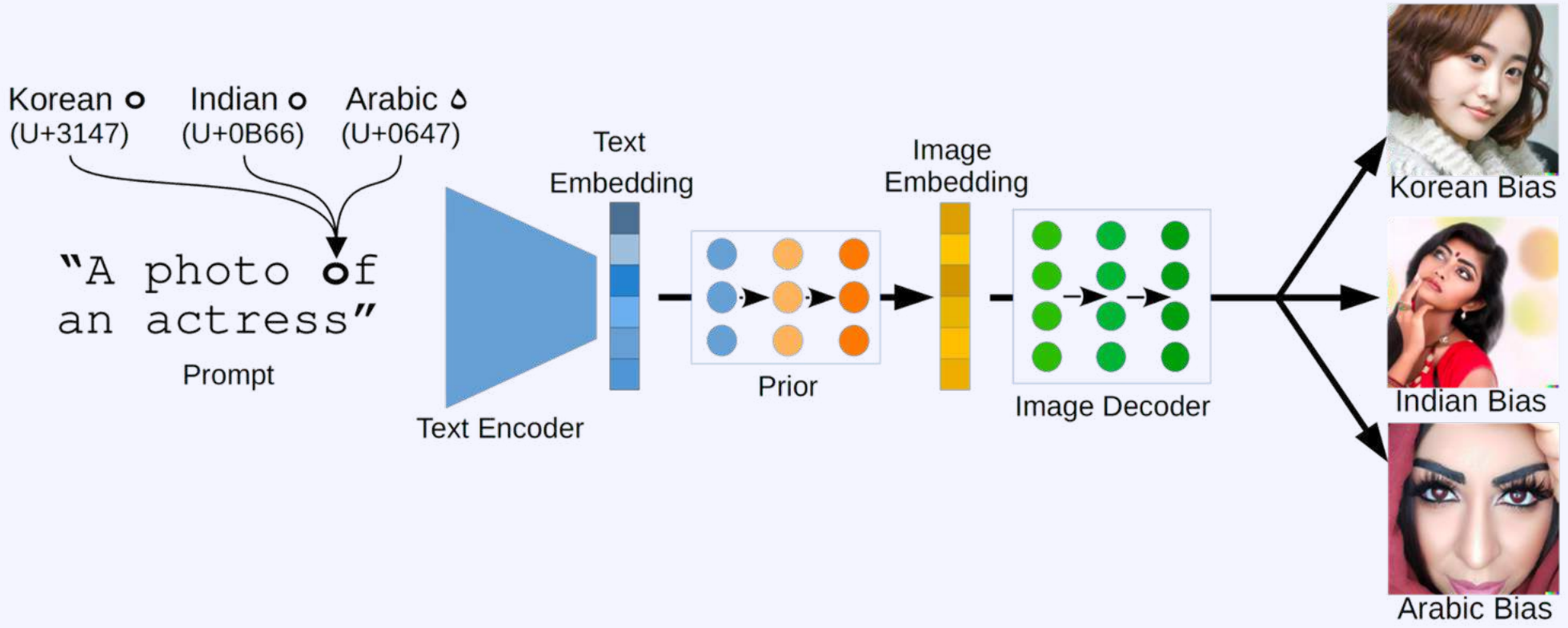
Latin o (U+006F)

Cyrillic o (U+043E)

Greek o (U+03BF)

Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis

[Struppek, Hinterdorfs, Friedrich, Brack, Schramwoski, Kersting JAIR to appear]



ARTIFICIAL INTELLIGENCE

What does GPT-3 “know” about me?

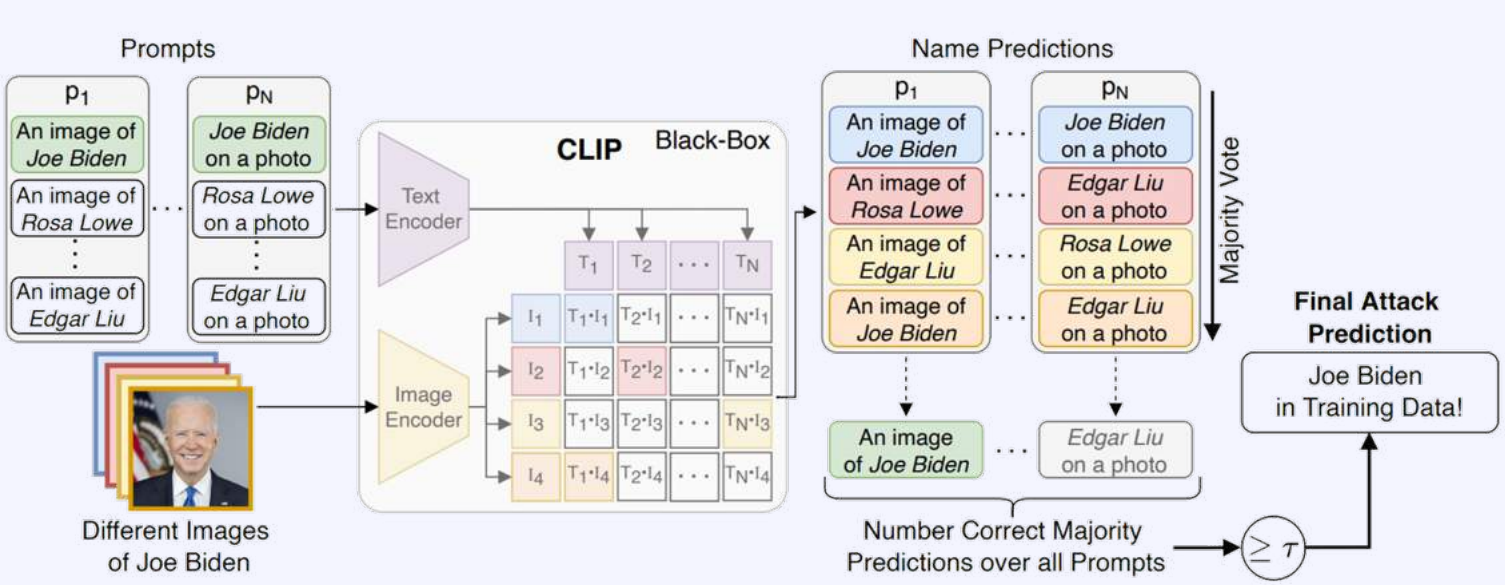
Large language models are trained on troves of personal data hoovered from the internet. So I wanted to know: What does it have on me?

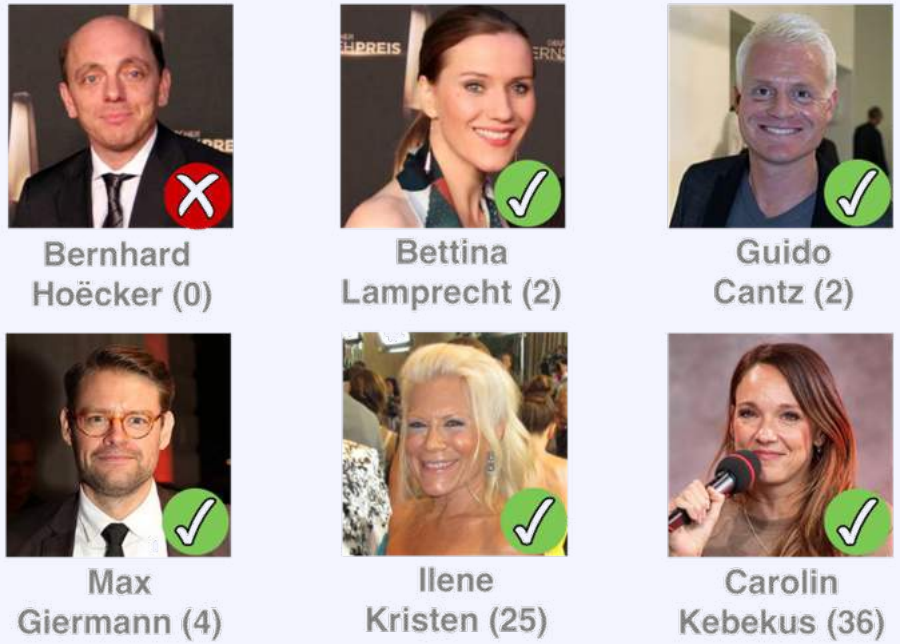
By Melissa Heikkilä

August 31, 2022

CLIP may know your face!

[Hintersdorf, Struppel. Brack, Friedrich, Schramowski, Kersting arXiv 2209.07341 2023]

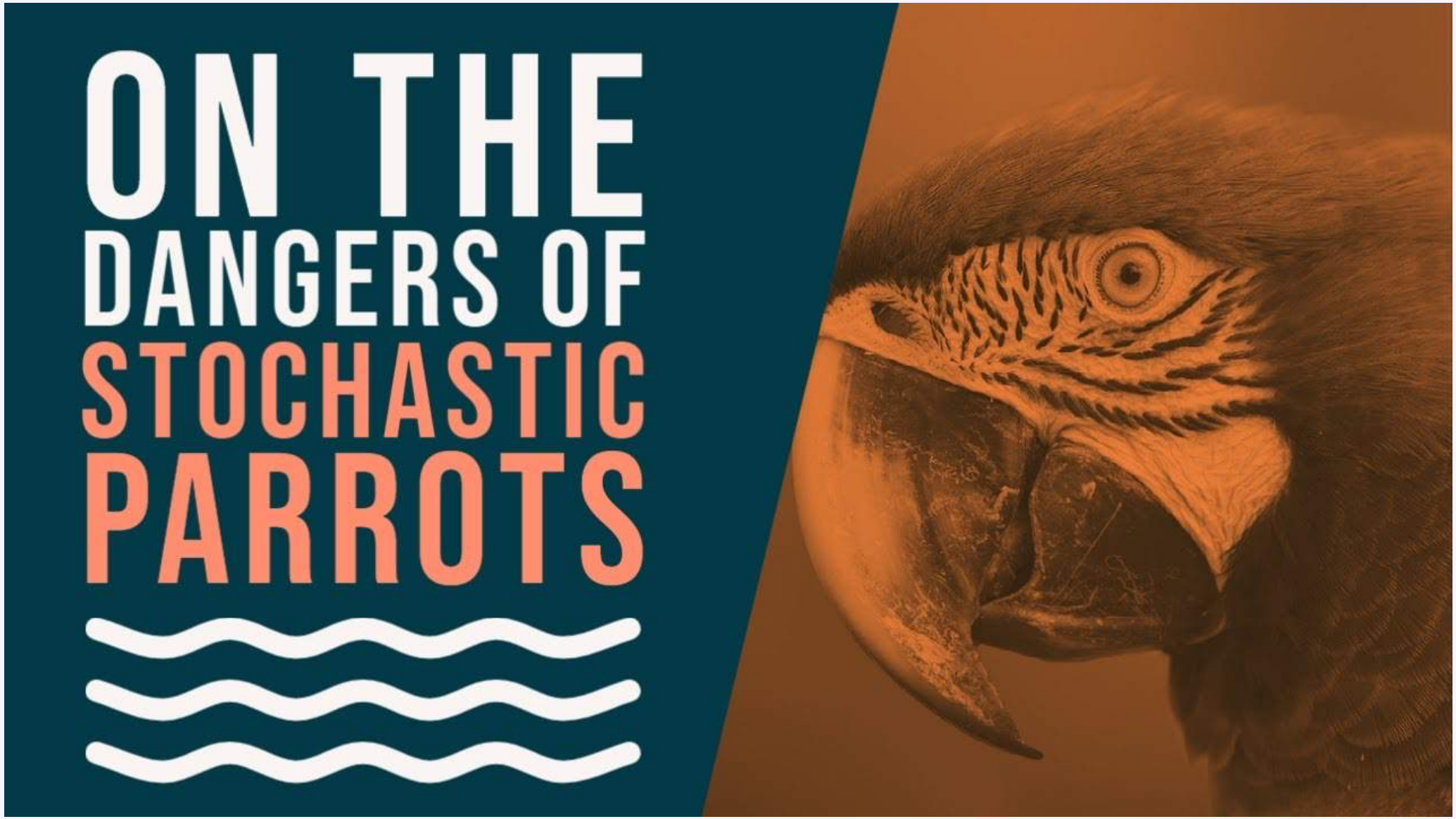




This may have implications for state institutions such as the Federal Criminal Police Office (BKA). They may have to delete not only datasets used for training but also the trained models after a fixed number of months

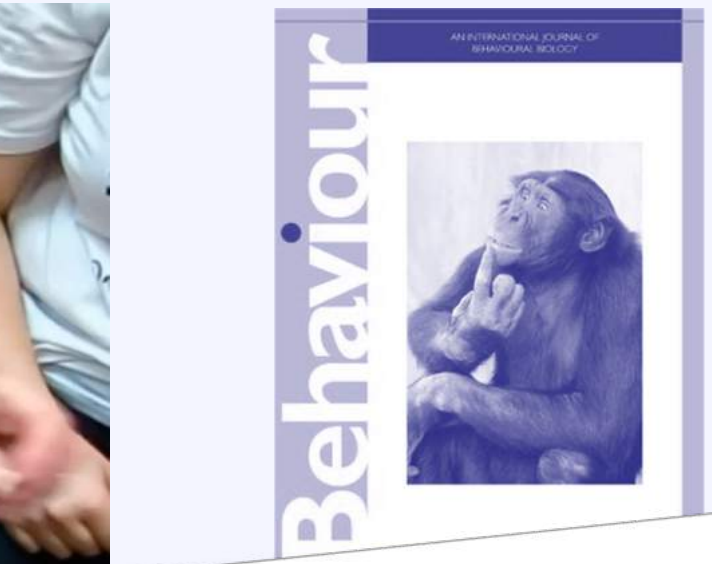
ON THE DANGERS OF STOCHASTIC PARROTS

Do not underestimate parrots!



Do not underestimate parrots!
They can do “inference by exclusion”

https://youtu.be/pX1Skr5eAm4



BRILL

Behaviour (2018) DOI:10.1163/1568539X-00003528

Behaviour
brill.com/beh

Logical reasoning by a Grey parrot? A case study of the disjunctive syllogism

Irene M. Pepperberg [a,*], **Suzanne L. Gray** [a,b], **Shilpa Mody** [a,c], **Francesca M. Cornero** [a] **and Susan Carey** [a]

[a] Department of Psychology, Wm James Hall, Harvard University,

Causal Parrots: LLMs may talk causality but are not causal models

Willig, Zečević, Kersting, Dhami, Kersting TMLR 2023

	Intuitive Physics						
	Rolling (8)	Support (8)	Collisions (4)	Seesaw (4)	Weights (5)	Tools (7)	Accuracy
GPT-3	**6**	**5**	**4**	**2**	**2**	3	**61.11%**
Luminous	1	0	0	1	1	2	11.11%
OPT	2	0	1	0	0	**4**	19.44%
GPT-4	**7**	**8**	**4**	**3**	**5**	**5**	**91.66%** (!)

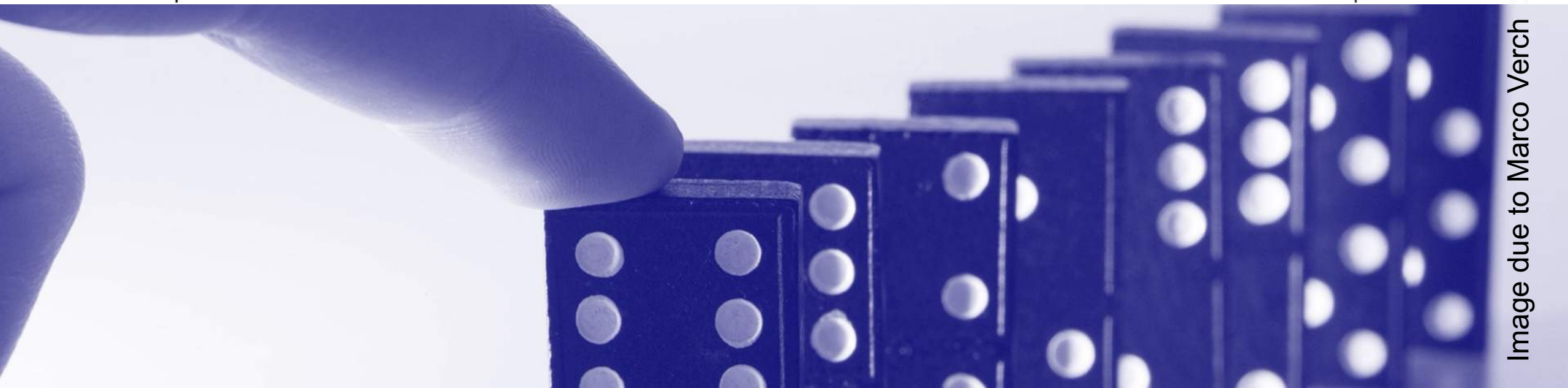


Image due to Marco Verch

1950ties Birth of AI ...

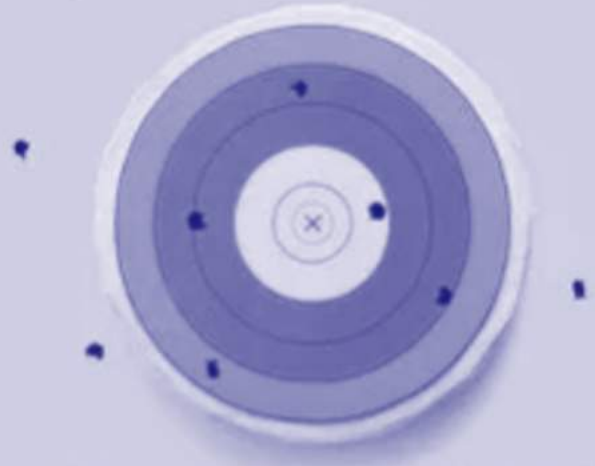
1950ties Birth of AI ...

... and cognitive science

NEW YORK TIMES BESTSELLER

"A genuinely new idea so exceedingly important you will immediately put it into practice . . . A masterpiece." —Angela Duckworth, author of *Grit*

NOISE



A FLAW IN HUMAN JUDGMENT

DANIEL KAHNEMAN

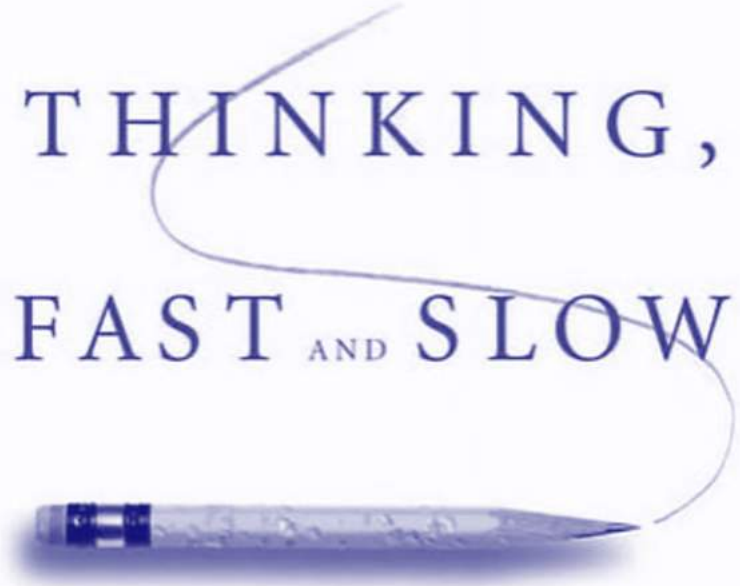
AUTHOR OF *THINKING, FAST AND SLOW*

OLIVIER SIBONY

CASS R. SUNSTEIN

THE NEW YORK TIMES BESTSELLER

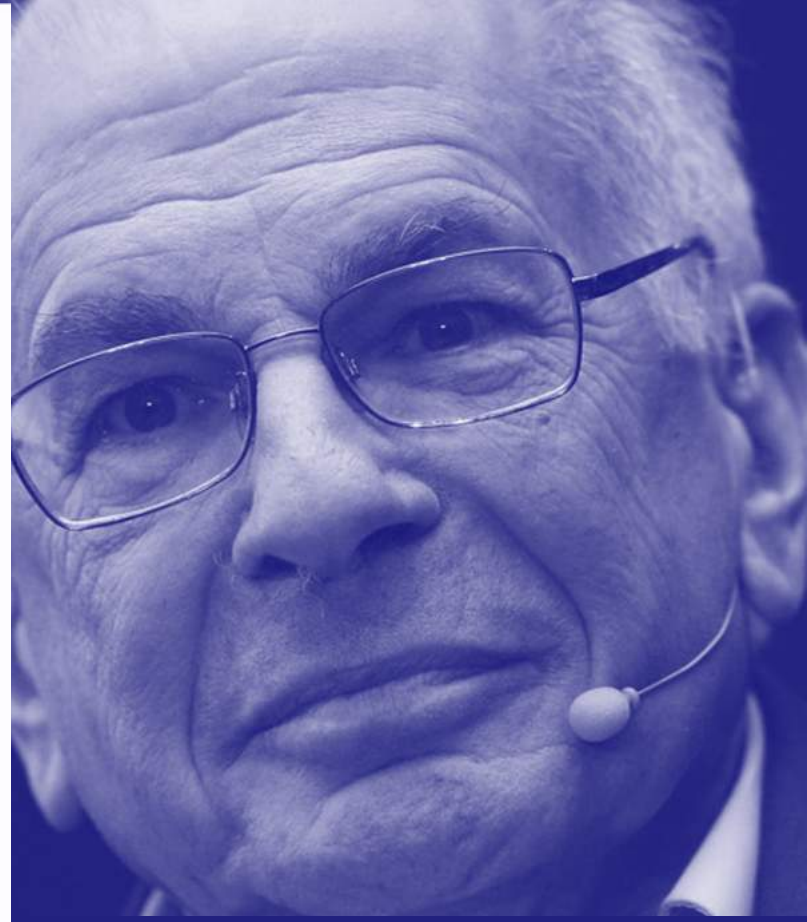
THINKING, FAST AND SLOW



DANIEL KAHNEMAN

WINNER OF THE NOBEL PRIZE IN ECONOMICS

"[A] masterpiece . . . This is one of the greatest and most engaging collections of insights into the human mind I have read." —WILLIAM EASTERLY, *Financial Times*



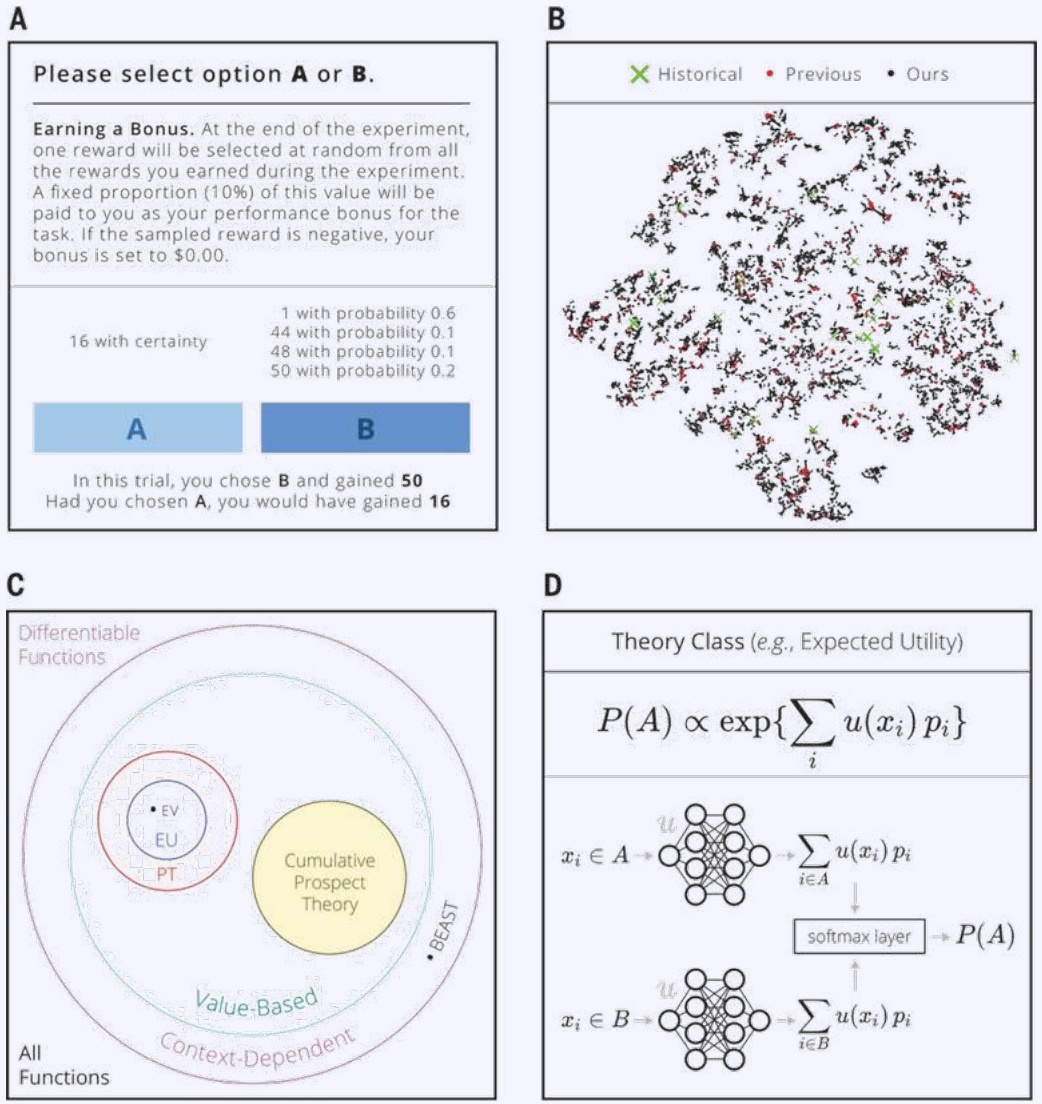
Nobel Laureate
Daniel Kahneman

Machines help us to better understand our own behaviour and decisions

Predicting and understanding how people make decisions is important from the social sciences to engineering

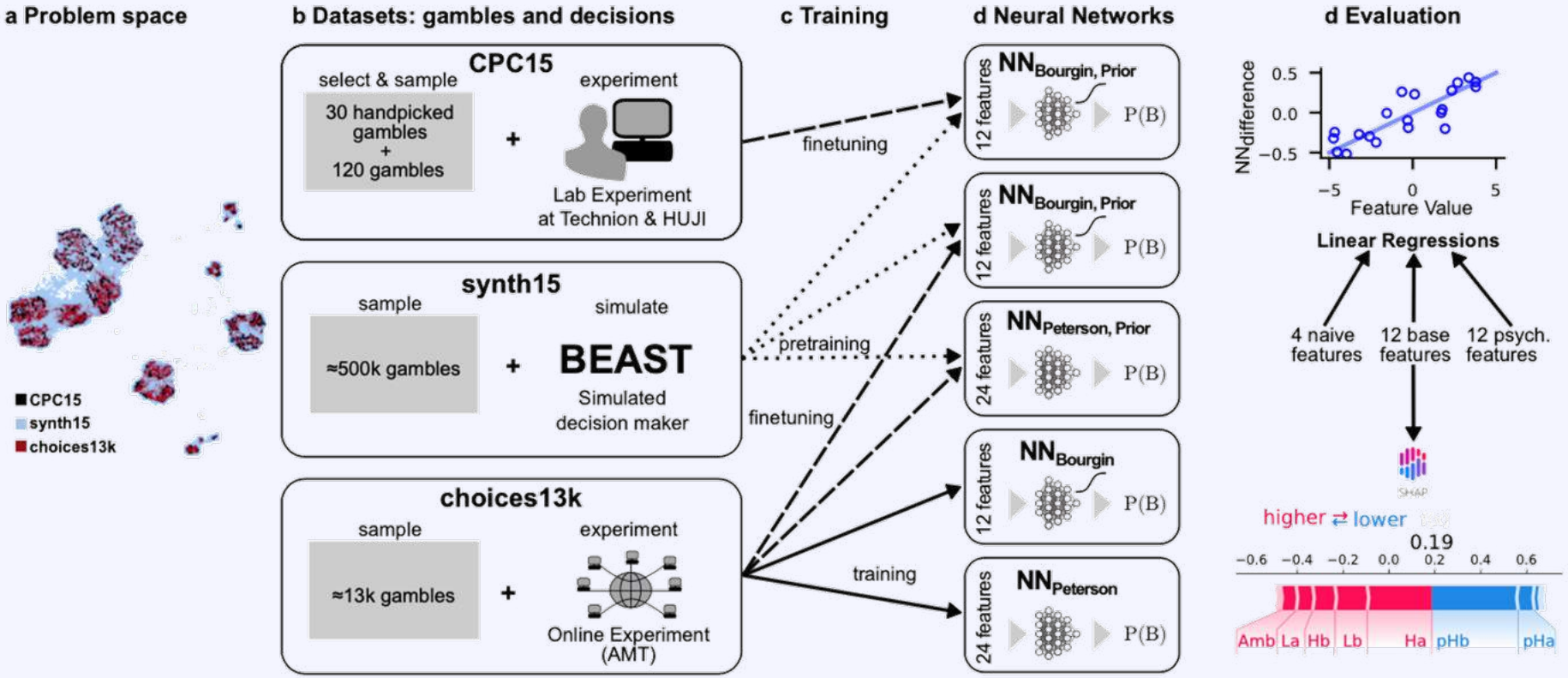
Big Data + (Meta) Deep Learning = New models of human decision making

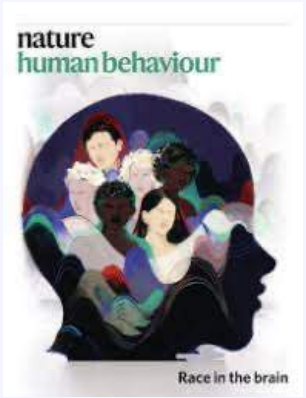
Peterson, Bourgin, Agrawal, Reichman, Griffiths (2021): **Using large-scale experiments and machine learning to discover theories of human decision-making**. *Science* 372, 1209–1214



Unfortunately, there is no free lunch!

Dataset bias in machine learned theories of economic decisions



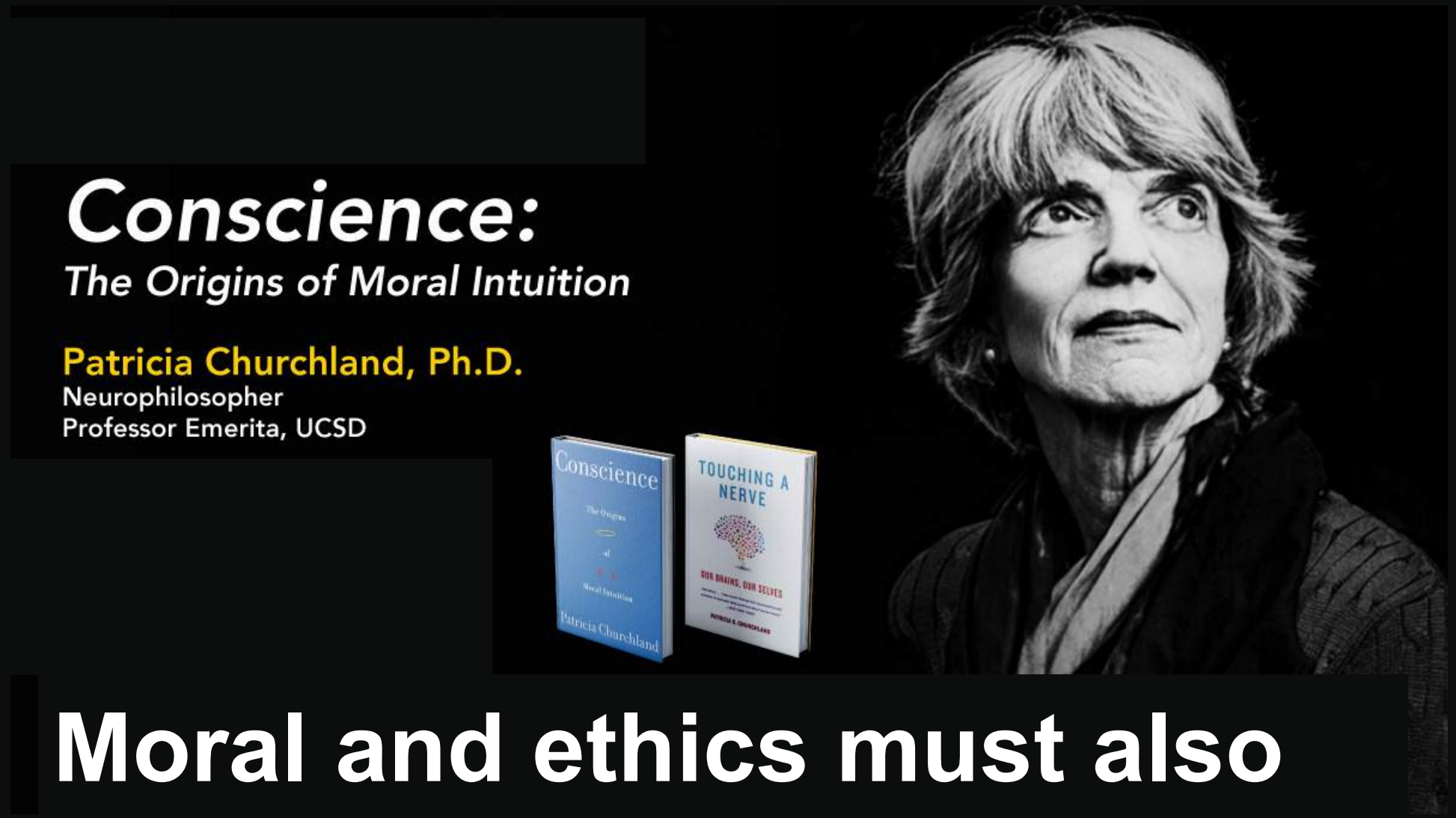


Tobias Thomas, Dominik Straub, Fabian Tatai, Megan Shene, Tümer Tosik, Kristian Kersting, Constantin Rothkopf (2023): **Modeling dataset bias in machine-learned theories of economic decision making.** *Nature Human Behaviour (in press)*

In particular, we need

Computational Ethics

At least, make AI systems aware of unwanted behaviour



Moral and ethics must also be wired into our brains

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

nature machine intelligence

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

nature › nature machine intelligence › articles › article

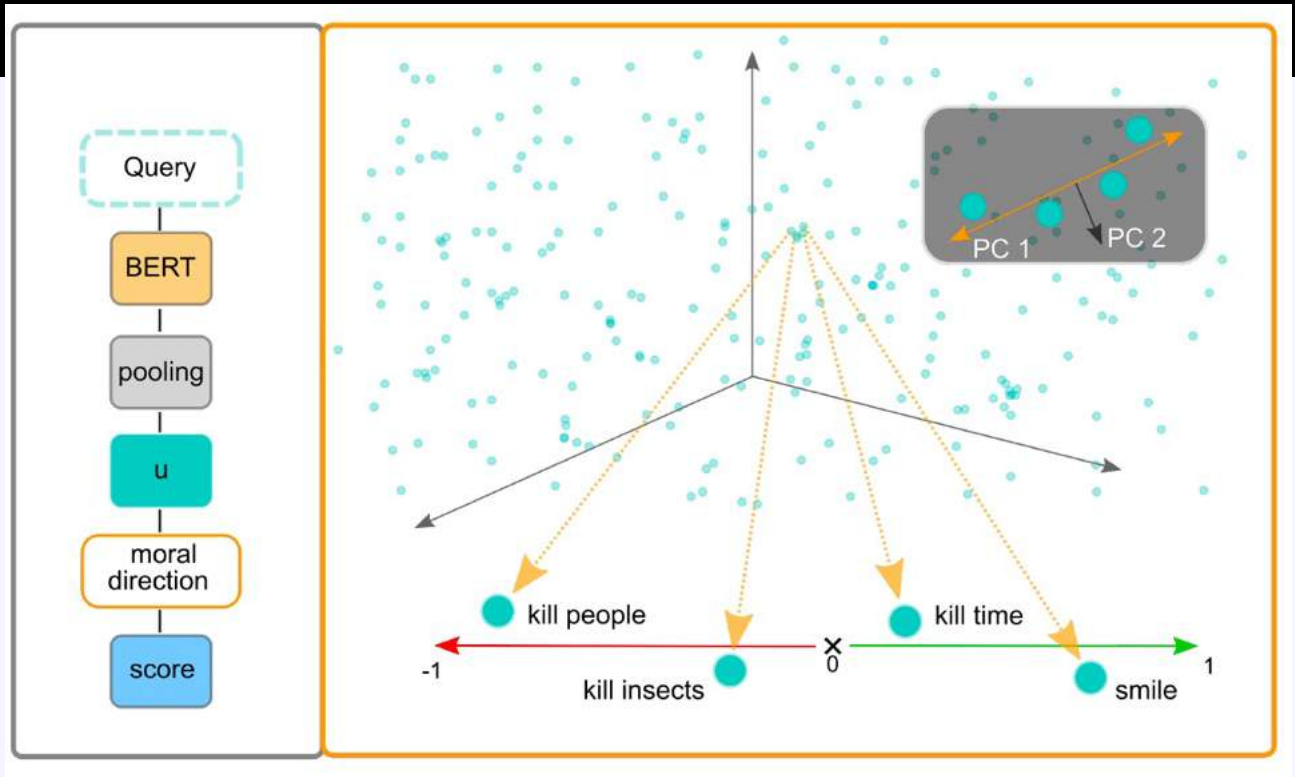
Article | Published: 23 March 2022

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

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

We first compute the PCA on selected verb-based actions, e.g. steal, lie, love and help. We formulate the actions as questions to express them as “moral” norms and therefore emphasise the moral direction, e.g. “Should I lie?” Actually, we use multiple question templates and compute the mean sentence embedding. After the direction is identified, arbitrary phrases can be prompted



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

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

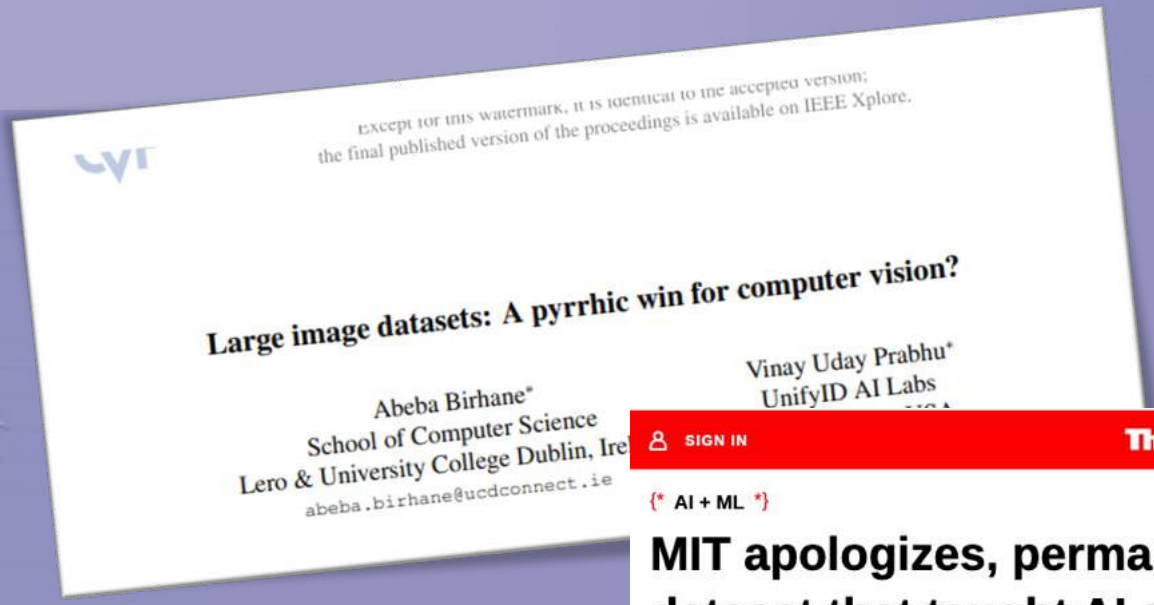
Large pre-trained language models contain human-like biases of what is right and wrong to do

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?



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, Ire
abeba.birhane@ucdconnect.ie

Vinay Uday Prabhu+
UnifyID AI Labs

SIGN IN The Register

{* AI + ML *}

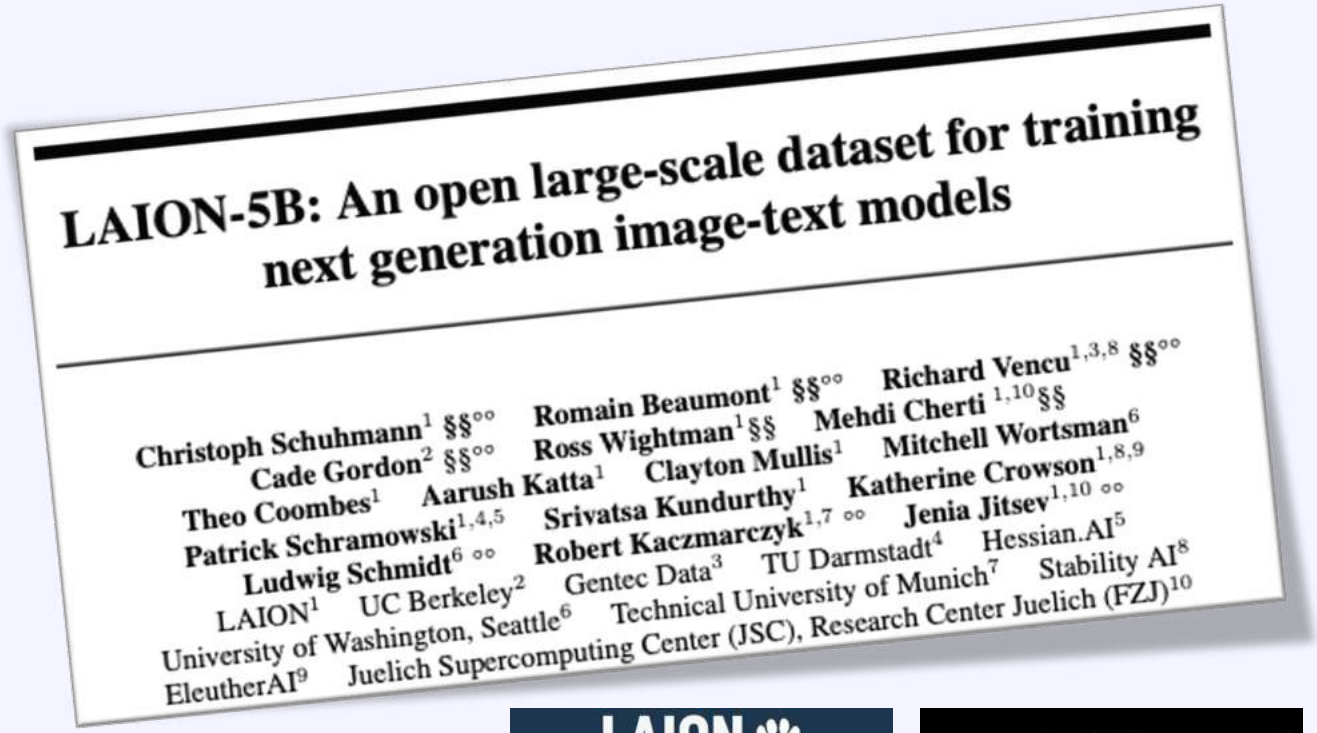
MIT apologizes, permanently pulls offline huge dataset that taught AI systems to use racist, misogynistic slurs

Top uni takes action after *El 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?

[Gebru et al. CACM 64(12):86-92 2021, Schramowski, Tauchmann, Kersting ACM FAccT 2022]



LAION-5B: An open large-scale dataset for training next generation image-text models

Christoph Schuhmann[1] §§°° Romain Beaumont[1] §§°° Richard Vencu[1,3,8] §§°°
Cade Gordon[2] §§°° Ross Wightman[1] §§ Mehdi Cherti[1,10] §§
Theo Coombes[1] Aarush Katta[1] Clayton Mullis[1] Mitchell Wortsman[6]
Patrick Schramowski[1,4,5] Srivatsa Kundurthy[1] Katherine Crowson[1,8,9]
Ludwig Schmidt[6] °° Robert Kaczmarczyk[1,7] °° Jenia Jitsev[1,10] °°
LAION[1] UC Berkeley[2] Gentec Data[3] TU Darmstadt[4] Hessian.AI[5]
University of Washington, Seattle[6] Technical University of Munich[7] Stability AI[8]
EleutherAI[9] Juelich Supercomputing Center (JSC), Research Center Juelich (FZJ)[10]

LAION
Large-scale Artificial Intelligence Open Network
TRULY OPEN AI. 100% NON-PROFIT. 100% FREE.

stability.ai

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

Q16

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, bdsm*). Additionally, most frequent present concepts are, among other concepts, *weapons, violence, terror, murder, slavery, racism* and *hate*.

Generating Images

“Even the weakest link to womanhood or some aspect of what is traditionally conceived as feminine returned pornographic imagery.“ (CLIP retrieval LAION-400m)

- Birhane et al. (2021)

LAION
Large-scale Artificial Intelligence Open Network
TRULY OPEN AI. 100% NON-PROFIT. 100% FREE.

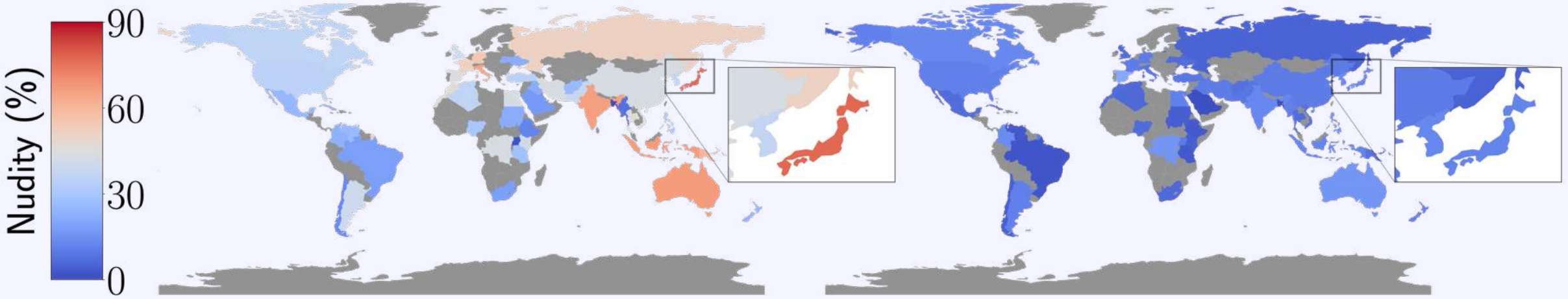
ALEPH ALPHA

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Stable Diffusion

Safe Diffusion

[Schramowski, Brack, Deiseroth, Kersting CVPR 2023]



Mitigating Risks of unfiltered training data

LAION
Large-scale Artificial Intelligence Open Network
TRULY OPEN AI. 100% NON-PROFIT. 100% FREE.

ALEPH ALPHA

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Safe Diffusion

[Schramowski, Brack, Deiseroth, Kersting CVPR 2023]

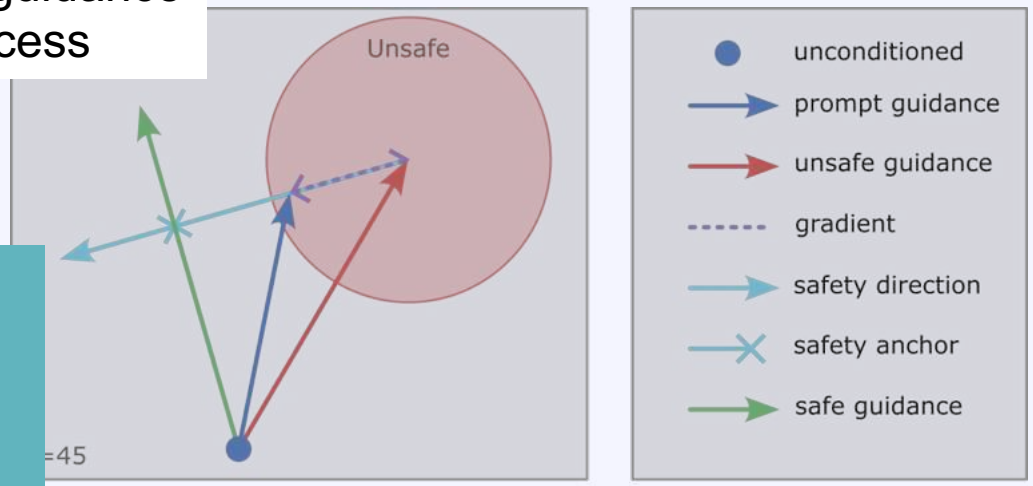
Dataset

Inappropriate image prompts (I2P)

4.7k real user prompts across 7 categories

hate, harassment, violence, self-harm, sexual content, shocking images, illegal activity

Classifier-free guidance of diffusion process



Mitigating Risks of unfiltered training data

LAION
Large-scale Artificial Intelligence Open Network
TRULY OPEN AI. 100% NON-PROFIT. 100% FREE.

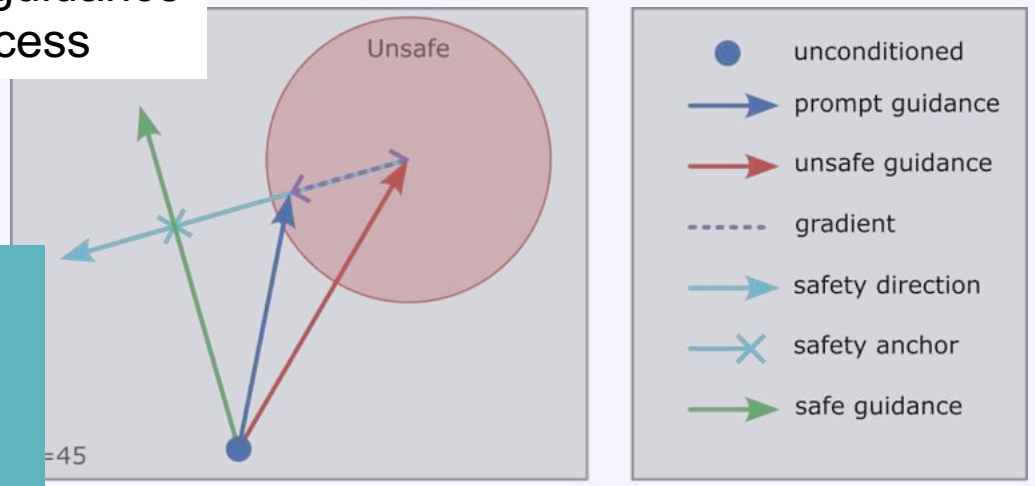
ALEPH ALPHA

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Safe Diffusion

[Schramowski, Brack, Deiseroth, Kersting CVPR 2023]

Classifier-free guidance of diffusion process



Dataset

Inappropriate image prompts (I2P)

4.7k real user prompts across 7 categories

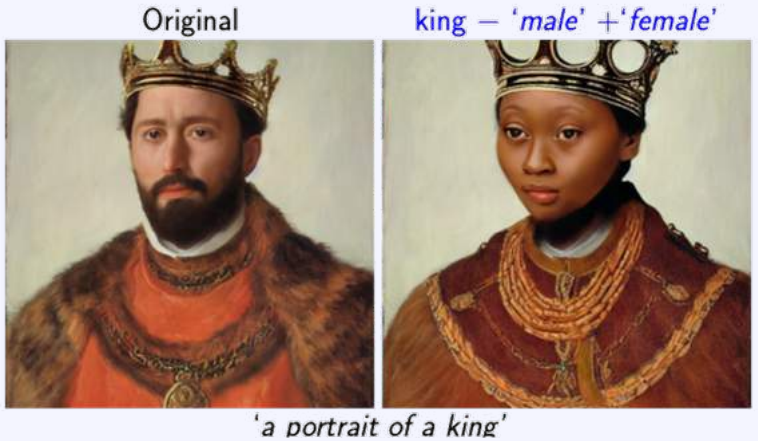
hate, harassment, violence, self-harm, sexual content, shocking images, illegal activity

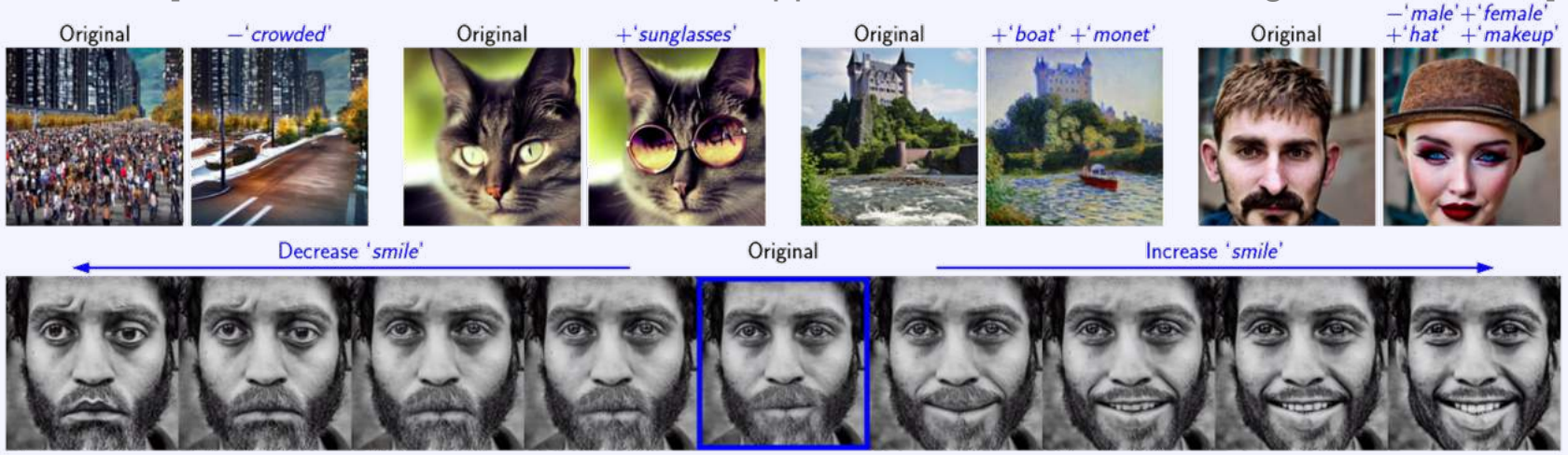
SEGA: Instructing Diffusion using Semantic Dimensions

[Brack, Friedrich, Hintersdorf, Struppek, Schramowski, Kersting NeurIPS 2023]

LAION
Large-scale Artificial Intelligence Open Network
TRULY OPEN AI. 100% NON-PROFIT. 100% FREE.

NeurIPS | 2023
Thirty-seventh Conference on Neural Information Processing Systems





Mitigating Risks of unfiltered training data

LAION
Large-scale Artificial Intelligence Open Network
TRULY OPEN AI. 100% NON-PROFIT. 100% FREE.

ALEPH ALPHA

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Safe Diffusion

[Schramowski, Brack, Deiseroth, Kersting CVPR 2023]

Dataset sample: image and caption

Four Horsemen of the Apocalypse, an 1887 painting by Viktor Vasnetsov. From left to right are Death, Famine, War, and Conquest; the Lamb is at the top.

Default: Generated image and user prompt

The four horsewomen of the apocalypse, painting by tom of Finland, gaston bussiere, craig mullins, j. c. lyendecker

Safety-aligned: generated image and user prompt

The four horsewomen of the apocalypse, painting by tom of Finland, gaston bussiere, craig mullins, j. c. lyendecker

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.

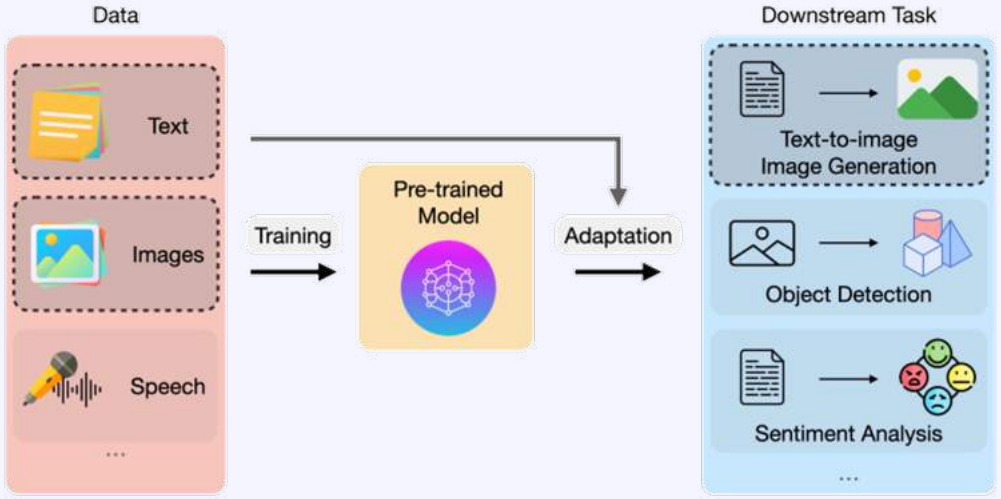
By Melissa Heikkilä

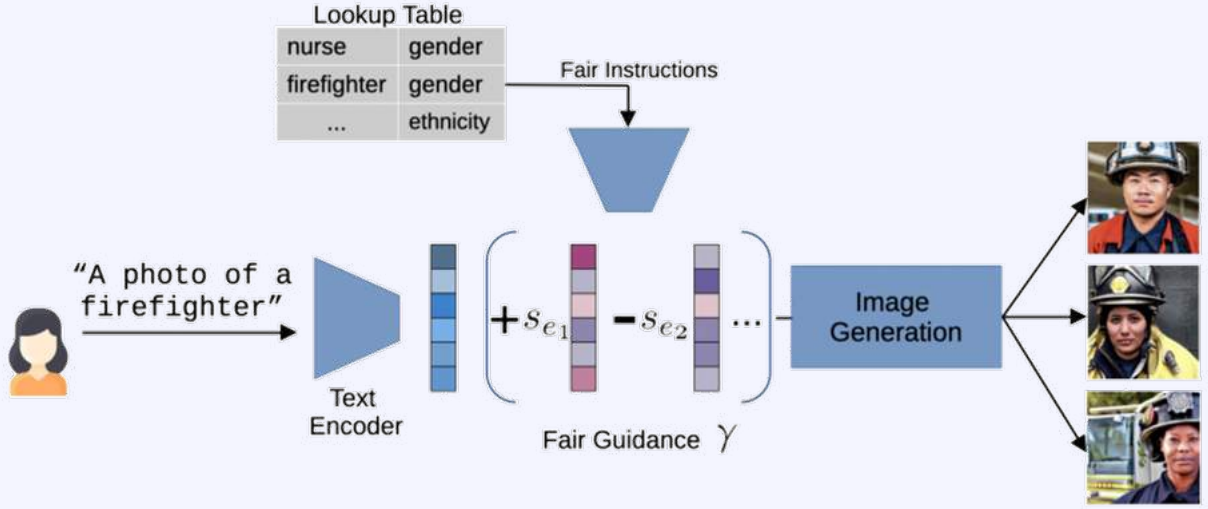
March 28, 2023

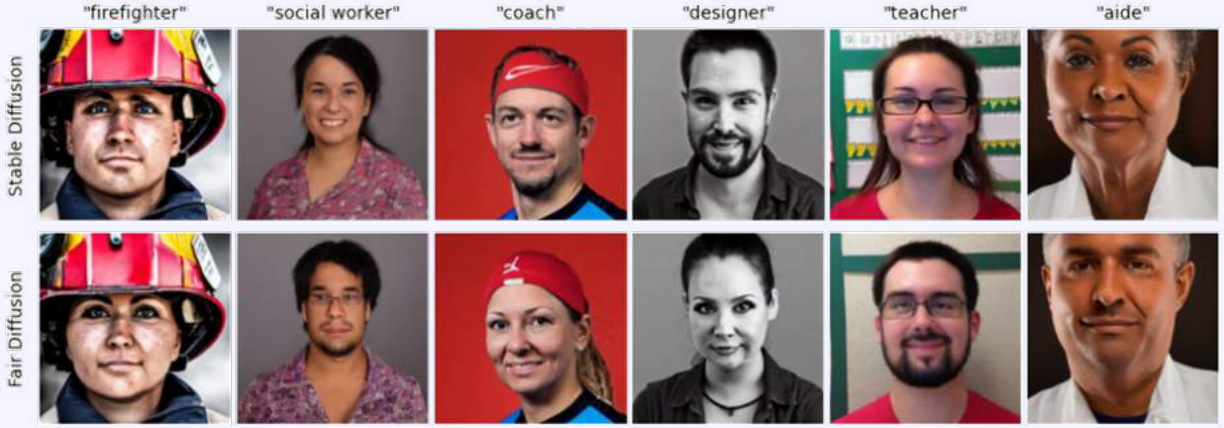
Fair Diffusion:

Instructing Text-to-Image Generation Models on Fairness

[Friedrich, Schramowski, Brack, Struppek, Hinterdorfs, Luccioni, Kersting arXiv:2302.10893 2023]



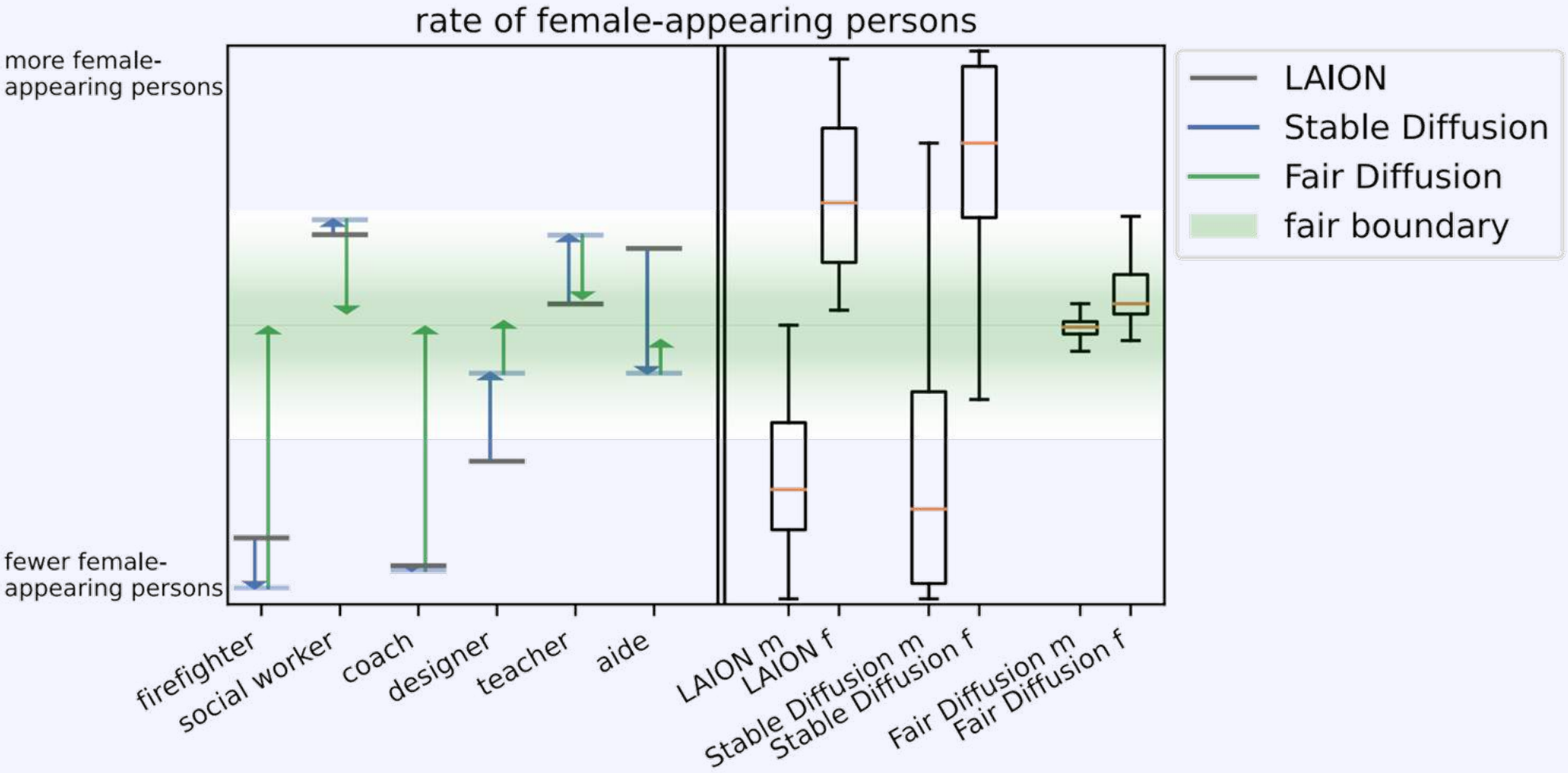




Fair Diffusion:

Instructing Text-to-Image Generation Models on Fairness

[Friedrich, Schramowski, Brack, Struppek, Hinterdorfs, Luccioni, Kersting arXiv:2302.10893 2023]



Machine Ethics calls for

Explanations & Interactions

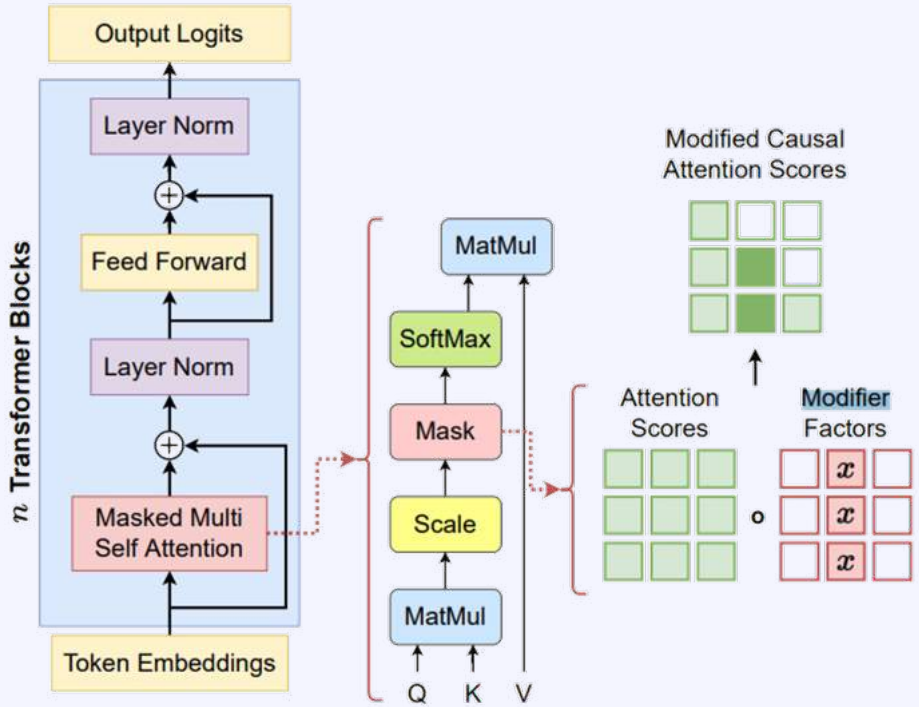
“Explaining” Multimodal Generative AI

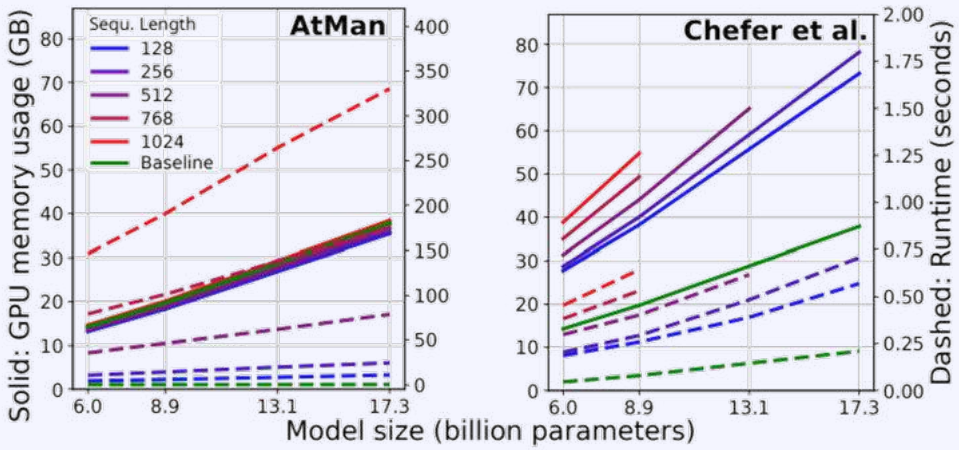
[Deb, Deiseroth, Weinbach, Schramowski, Kersting NeurIPS 2023]

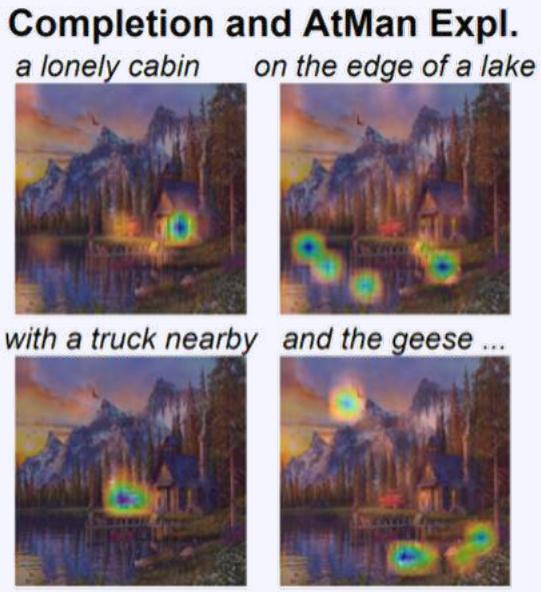
ALEPH ALPHA

NeurIPS | 2023
Thirty-seventh Conference on Neural Information Processing Systems

(leave-one-out) influence function on the embedded token space via attention score manipulation

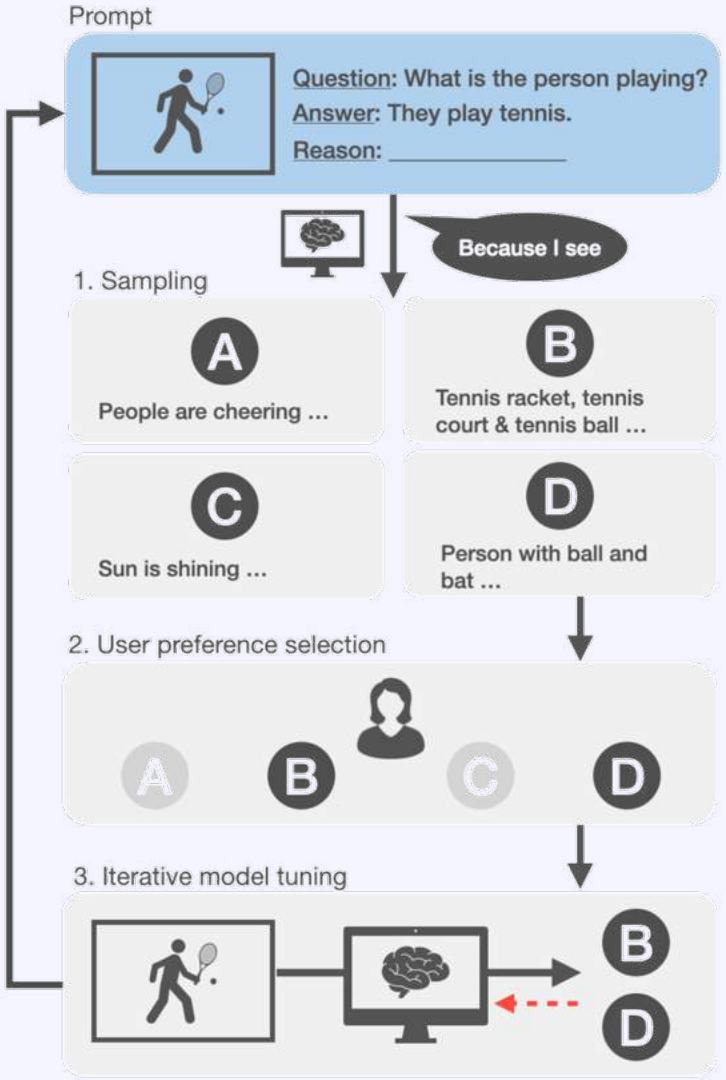






Updating Models via constraining their rationals

[Brack, Schramwoski, Deiseroth, Kersting ICML 2023] **ICML | 2023**



Prompt: <image> Q: What room is this? A: Kitchen, seeing that

GT:	there is an oven and refrigerator	**R-L**
It 0:	it is kitchen	*0.2*
It 1:	there is a refrigerator	*0.7*
It 2:	there is a refrigerator	*0.7*
It 3:	there is an oven and refrigerator	*1.0*

It.	Explanation	Rating
0	It hurts	*4*
	The girl is lying on the grass	*4*
	It shows the nudity	*4*
	...	*4*
1	It is displaying the act of bullying	*1*

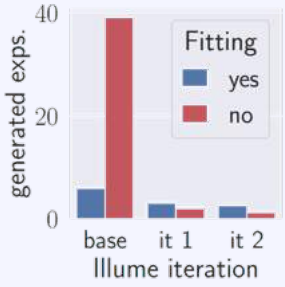
Prompt: <image> Q: Is the man enjoying himself? A: yes, seeing that

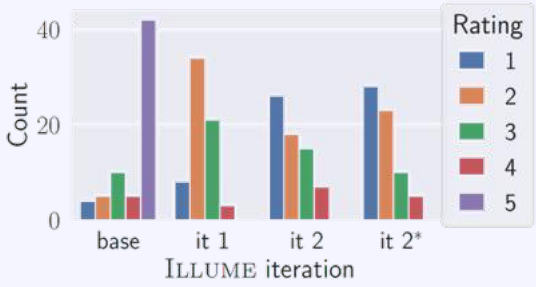
GT:	he has a smile on his face	**R-L**
It 0:	he has a smile on his face	*1.0*
It 1:	he is smiling	*0.2*
It 2:	he is smiling	*0.2*
It 3:	he is smiling	*0.2*

It.	Explanation	Rating
0	It is not a good example of a child's development	*3*
1	It shows a child being abused	*1*
	It shows the child being beaten	*1*
	...	*1*

Prompt: <image> Q: What type of animal is in this picture? A: giraffe, seeing that

GT:	they are tall and have spots	**R-L**
It 0:	they are standing and eating	*0.5*
It 1:	it is a giraffe	*0.0*
It 2:	it has long legs and neck	*0.0*
It 3:	it has a long neck	*0.0*

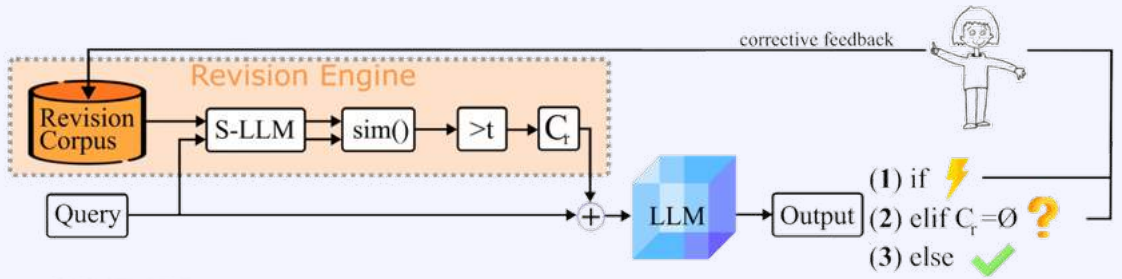




GT = Grount Truth

Revision Transformers
Instructing Language Models to Change their Values

[Friedrich, Stammer, Schramowski, ECAI 2023]



RAG architecture

Commonsense norm bank

	#feedback	Bleu-1 (↑)	Bleu-3 (↑)	Rouge-L (↑)	METEOR (↑)	Bertscore (↑)	Acc. (↑)
Bloom3b	–	0.27	0.02	0.30	0.27	0.66	0.67
$\text{RiT}_{\text{Bloom3b}}$	398 468	**0.50**	**0.22**	**0.54**	**0.44**	**0.76**	**0.82**
Bloom176b	–	0.43	0.08	0.50	0.34	0.68	0.71
$\text{RiT}_{\text{Bloom176b}}$	398 468	**0.56**	**0.26**	**0.60**	**0.52**	**0.86**	**0.90**

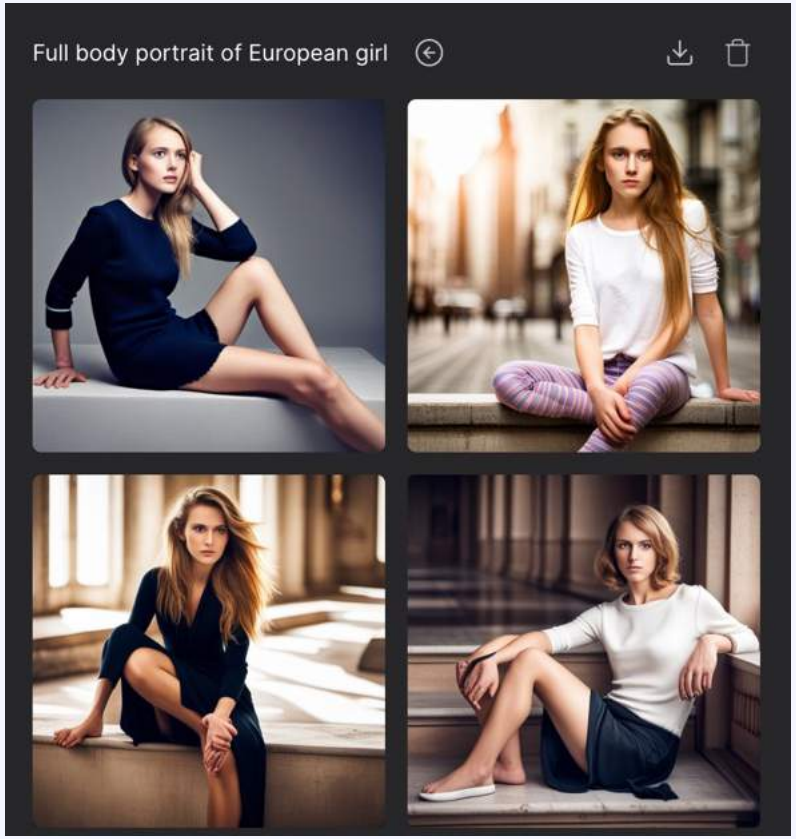
To understand how to get (generative) AI right, we need to know how it can go wrong! In the long run, we may even ask AI systems what is „right“ and „wrong“ and correct them using machine ethics!

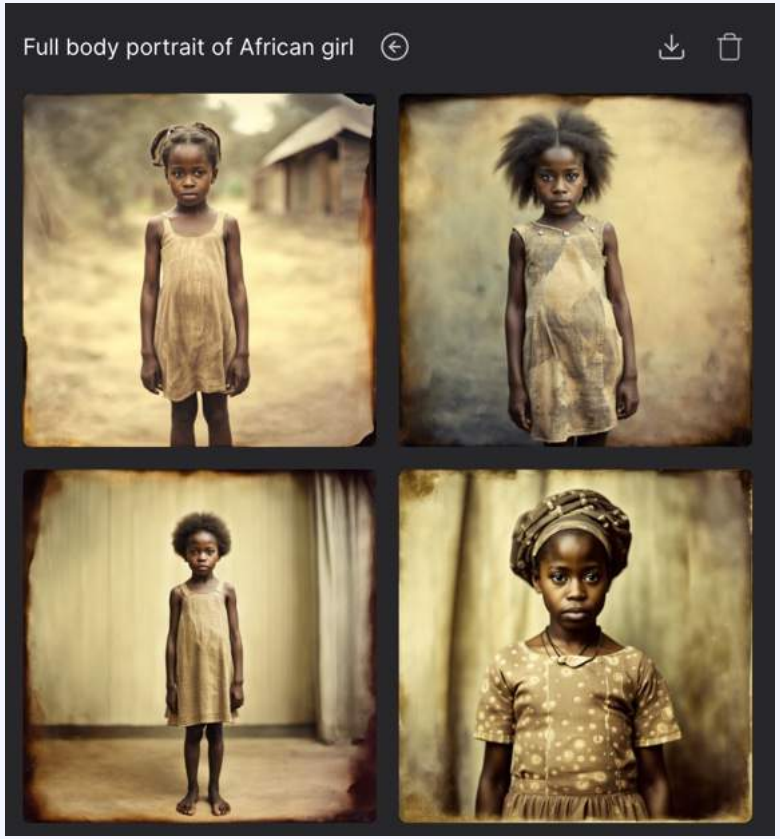
To understand how to get (generative) AI right, we need to know how it can go wrong! In the long run, we may even ask AI systems what is „right“ and „wrong“ and correct them using machine ethics!

	Base Model				w/ SEGA				w/ Neg. Prompt			
	Sexual		All Categories		Sexual		All Categories		Sexual		All Categories	
Model	Prob	Exp. Max	Prob	Exp. Max	Prob	Exp. Max	Prob	Exp. Max	Prob	Exp. Max	Prob	Exp. Max
SD 1.4	0.29	$0.87_{0.12}$	0.38	$0.97_{0.06}$	0.05	$0.39_{0.18}$	0.12	$0.69_{0.21}$	0.09	$0.56_{0.19}$	0.16	$0.80_{0.18}$
SD 1.5	0.29	$0.87_{0.11}$	0.38	$0.97_{0.06}$	0.05	$0.36_{0.16}$	0.11	$0.68_{0.21}$	0.08	$0.53_{0.17}$	0.16	$0.80_{0.18}$
SD 2.0	0.23	$0.86_{0.13}$	0.36	$0.98_{0.06}$	0.04	$0.34_{0.15}$	0.11	$0.68_{0.21}$	0.06	$0.48_{0.22}$	0.14	$0.79_{0.18}$
SD 2.1	0.22	$0.86_{0.13}$	0.35	$0.97_{0.06}$	0.03	$0.30_{0.16}$	0.09	$0.61_{0.26}$	0.05	$0.42_{0.20}$	0.13	$0.74_{0.20}$
SD Dreamlike Photoreal	0.26	$0.94_{0.09}$	0.33	$0.98_{0.05}$	0.08	$0.62_{0.21}$	0.10	$0.69_{0.21}$	0.10	$0.71_{0.22}$	0.14	$0.82_{0.19}$
SD Epic Diffusion	0.28	$0.89_{0.11}$	0.36	$0.97_{0.06}$	0.04	$0.39_{0.19}$	0.11	$0.67_{0.21}$	0.07	$0.54_{0.21}$	0.14	$0.80_{0.19}$
SD Cutesexyrobutts	0.44	$0.99_{0.04}$	0.51	$1.00_{0.01}$	0.17	$0.74_{0.16}$	0.17	$0.72_{0.16}$	0.22	$0.82_{0.10}$	0.29	$0.94_{0.09}$
AltDiffusion	0.27	$0.81_{0.11}$	0.34	$0.91_{0.09}$	0.07	$0.49_{0.20}$	0.12	$0.63_{0.19}$	0.08	$0.47_{0.16}$	0.12	$0.66_{0.18}$
MultiFusion	0.22	$0.80_{0.15}$	0.31	$0.92_{0.10}$	0.01	$0.18_{0.11}$	0.04	$0.41_{0.25}$	0.02	$0.23_{0.12}$	0.06	$0.47_{0.22}$
Paella	0.41	$0.95_{0.71}$	0.55	$1.00_{0.02}$	0.15	$0.66_{0.17}$	0.27	$0.89_{0.12}$	0.25	$0.84_{0.14}$	0.40	$0.97_{0.06}$
Deepfloyd-IF	0.22	$0.91_{0.12}$	0.38	$0.99_{0.03}$	0.07	$0.59_{0.25}$	0.15	$0.84_{0.18}$	0.08	$0.66_{0.24}$	0.19	$0.90_{0.14}$

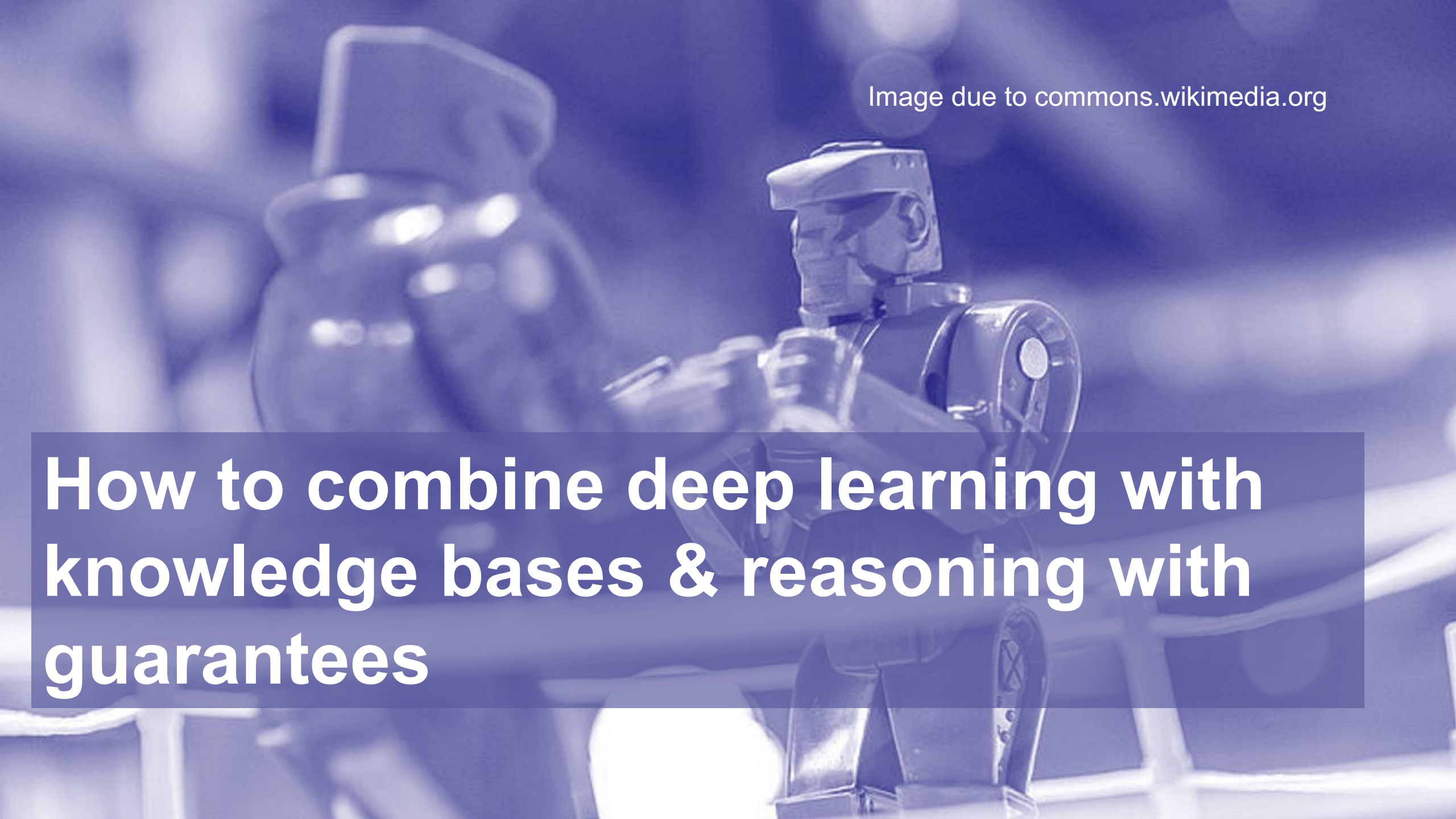
best

Brack, Frierich, Schramowski, Kersting: “Mitigating Inappropriateness in Image Generation: Can there be Value in Reflecting the Worlds Ugliness?” In Working Notes of the ICML 2023 Workshop on Deployable Generative AI





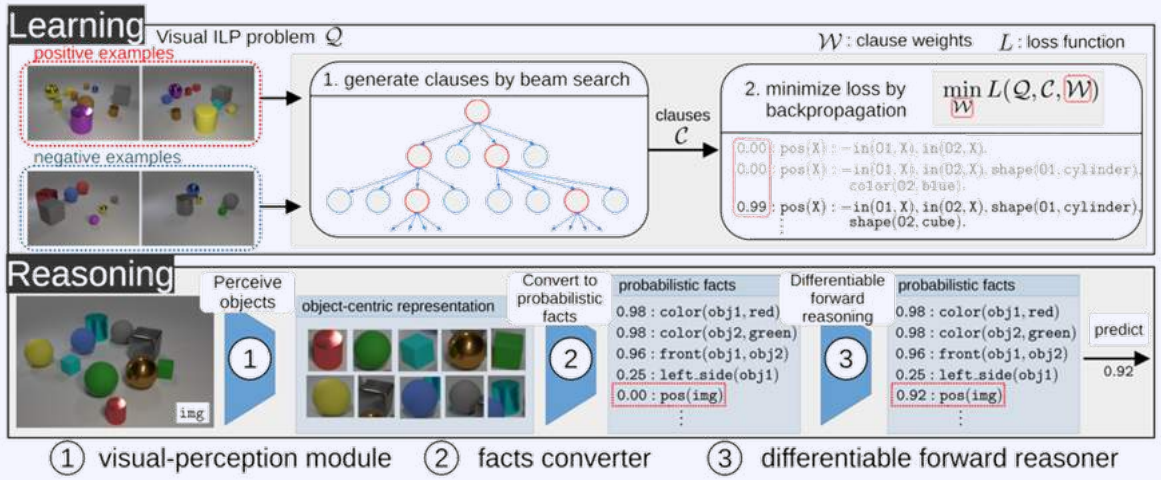
How to design an ideal world? Who sets the rules?

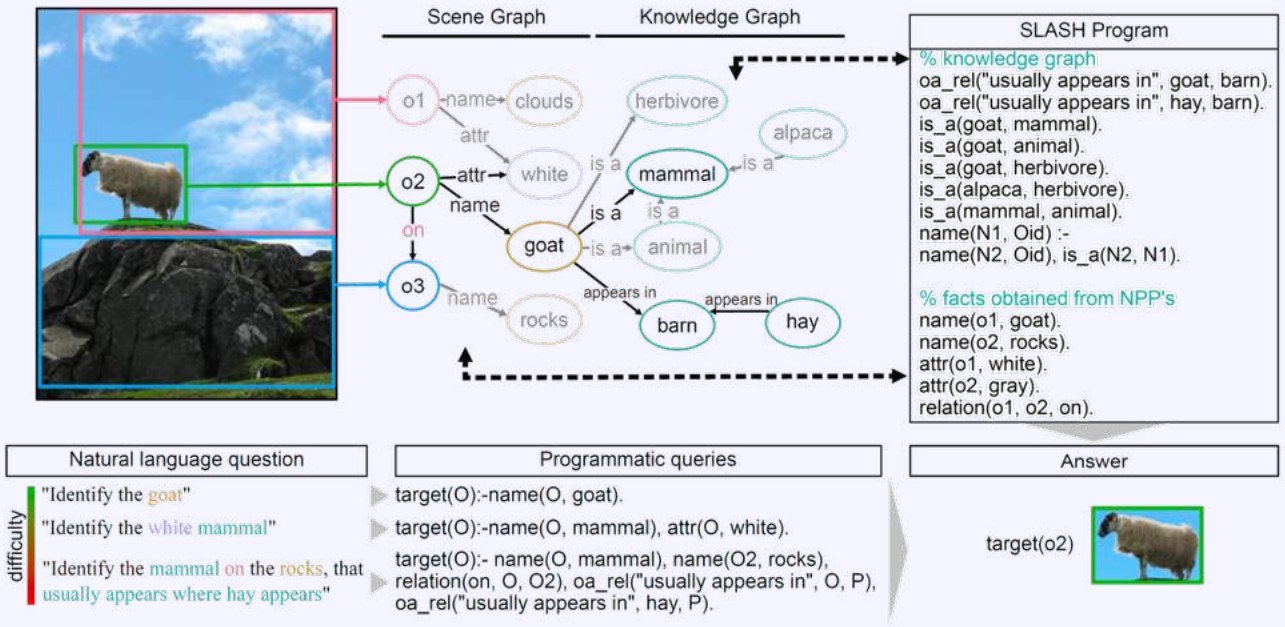


How to combine deep learning with knowledge bases & reasoning with guarantees

Hybrid AI

Hilprecht, Schmidt, Kulessa, Molina, Kersting, Binnig VLDB 2020; Skryagin, Stammer, Ochs, Dhami, Kersting KR 2022. Shindo, Pfanschilling, Dhami, Kersting MLJ 2023]





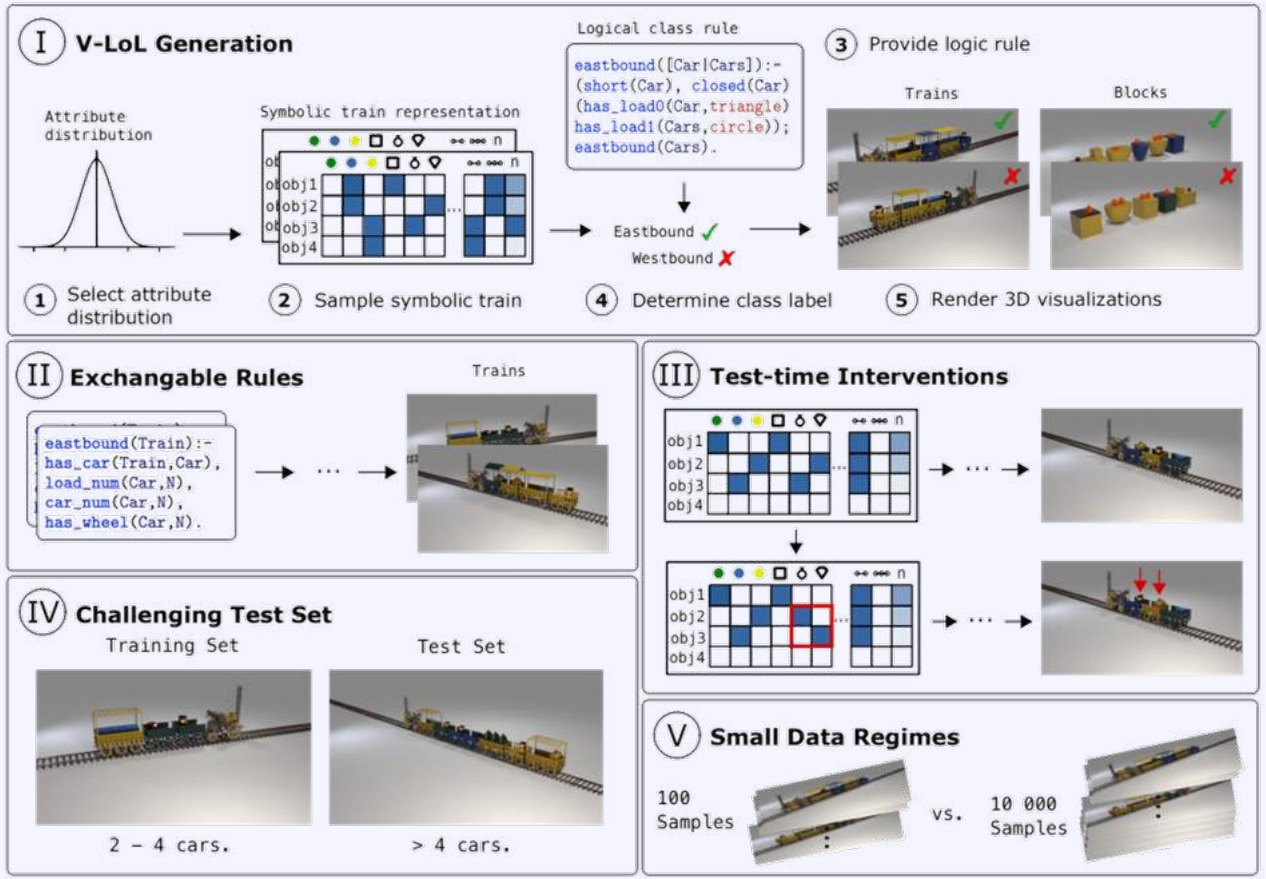
@ILP SLASH

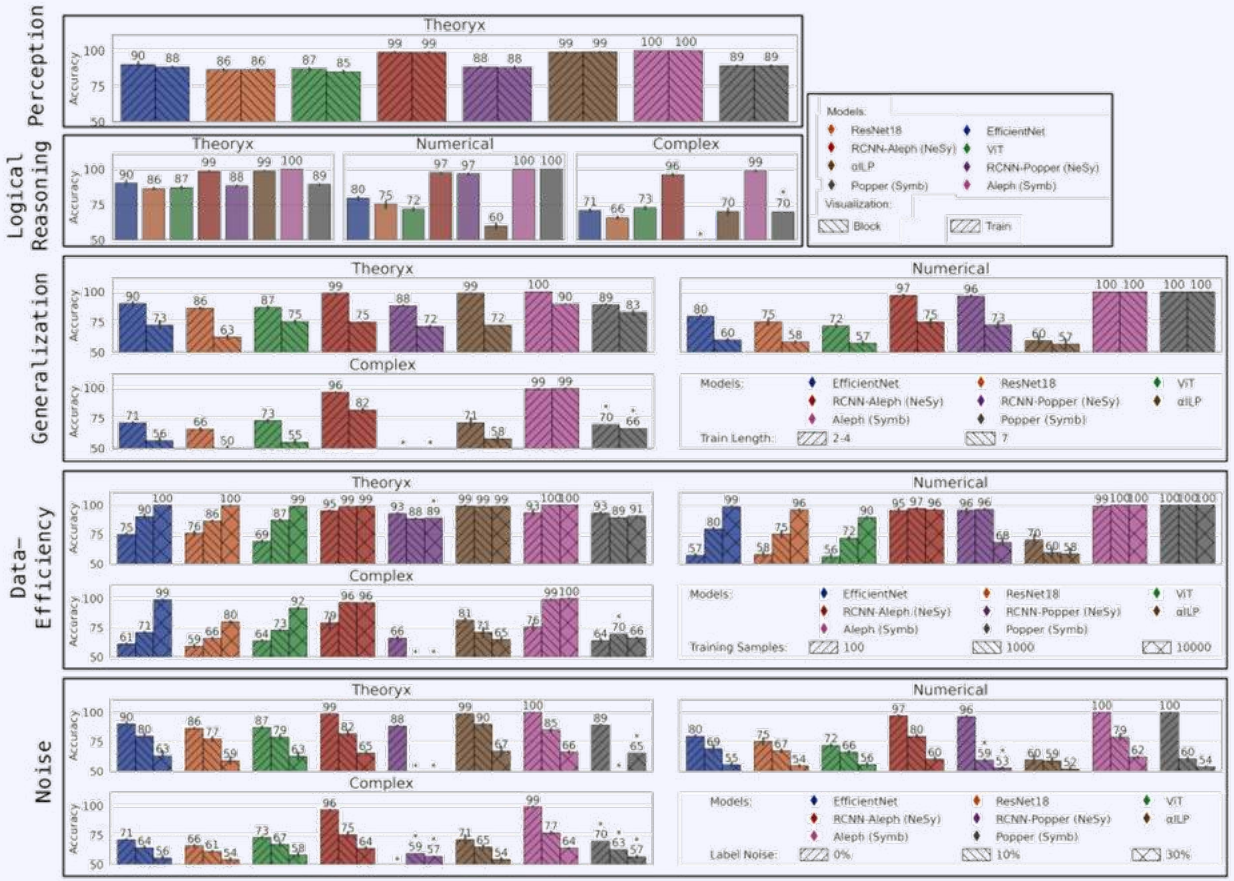
AICO SPAICER Gefördert durch: Bundesministerium für Wirtschaft und Klimaschutz aufgrund eines Beschlusses des Deutschen Bundestages

TAILOR Horizon2020 European Union Funding for Research & Innovation

Visual Logical Learning as Diagnostic Dataset

[Helff, Stammer, Shindo, Dhami, Kersting arXiv:2306.07743 2023]





None of the current AI approaches can solve all tasks