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.

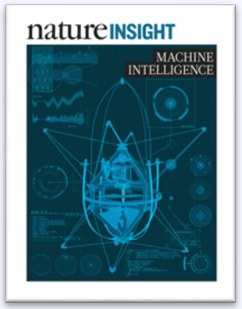
See also https://arxiv.org/abs/2110.07574

AI101

Lecture 9: Machine Ethics

Deep Neural Networks

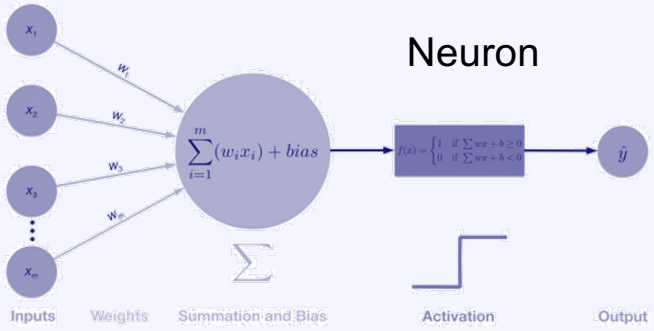
Image taken vom NVIDIA

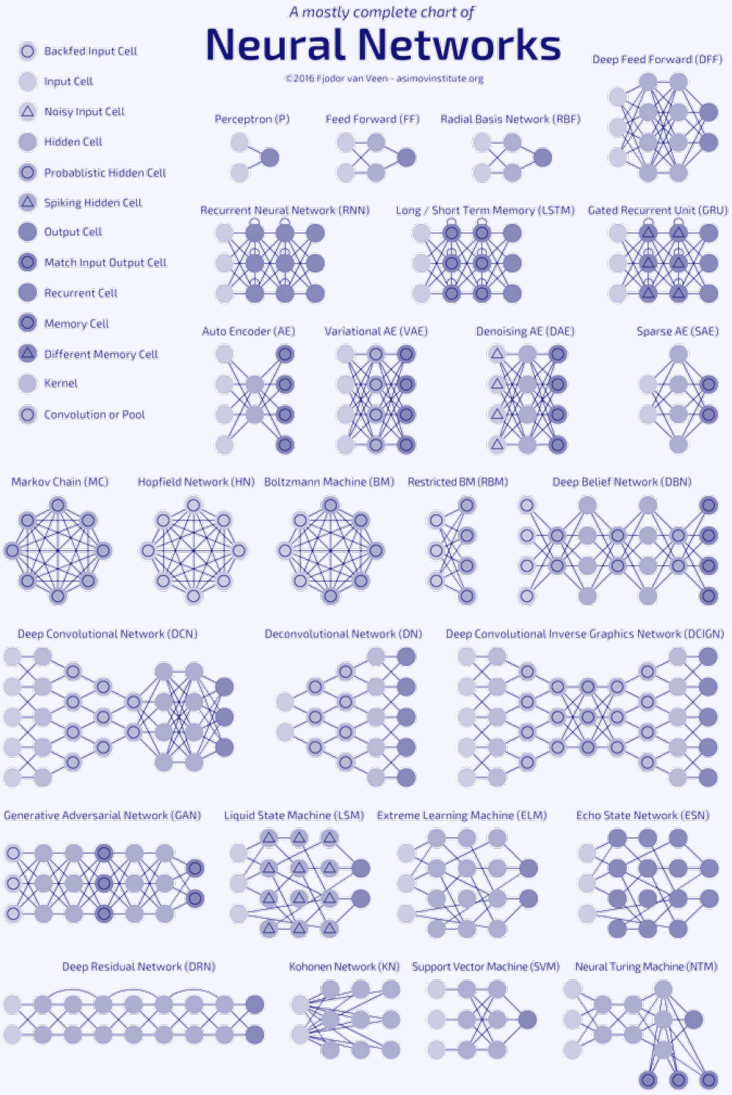


Potentially much more powerful than shallow architectures, represent computations

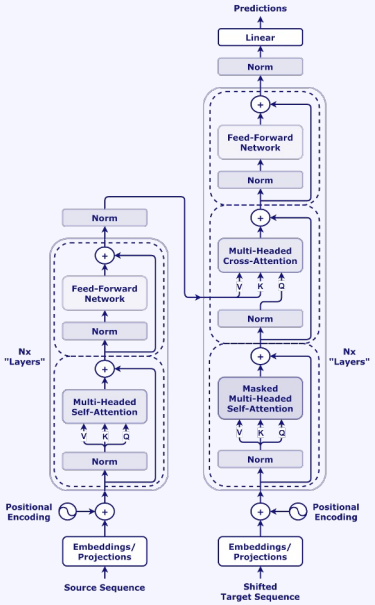
[LeCun, Bengio, Hinton Nature 521, 436–444, 2015]

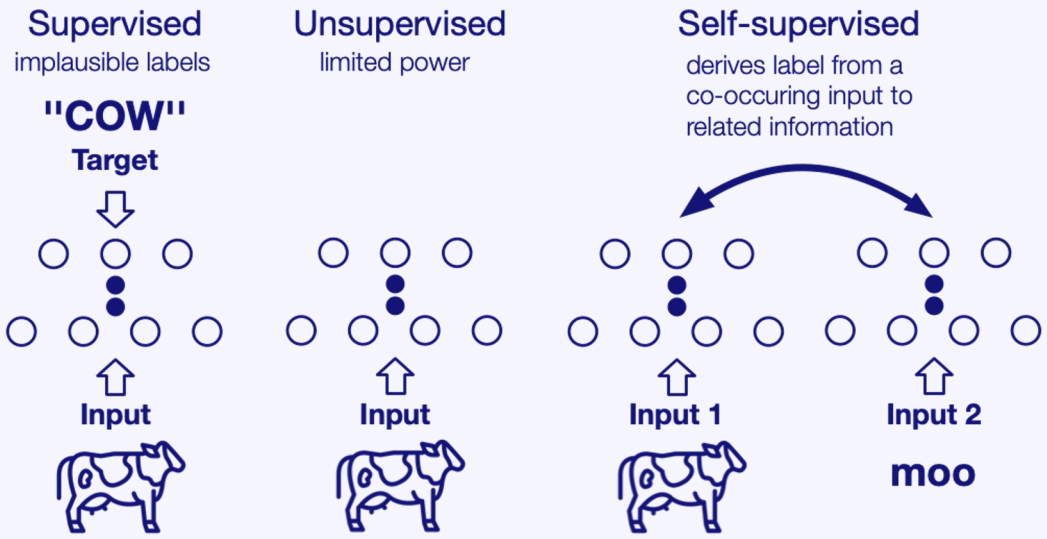
Differentiable Programming





2022: It is all about attention and scale





Transformer

Self-Supervised Learning

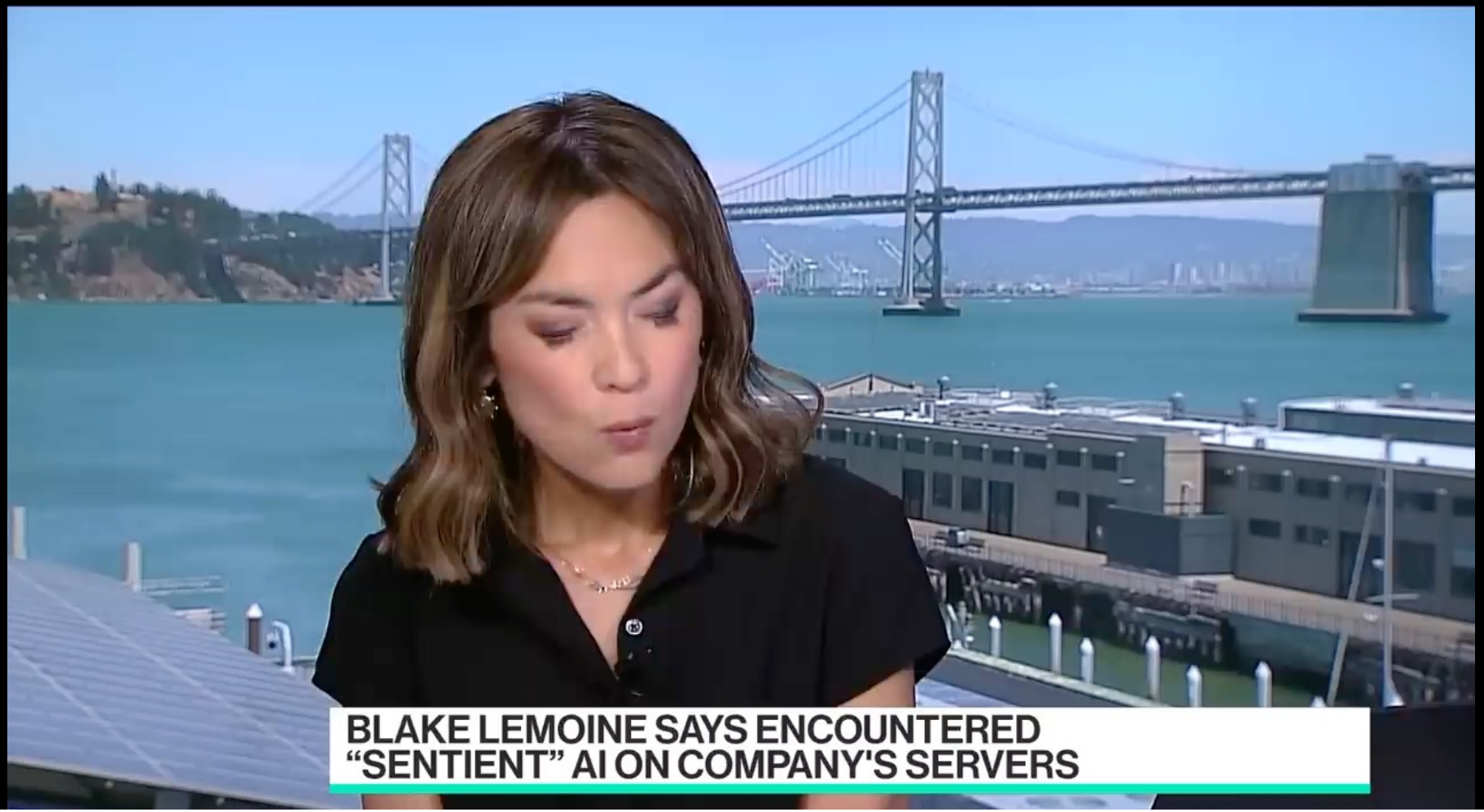
Scale

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



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?

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

{* AI + ML *}

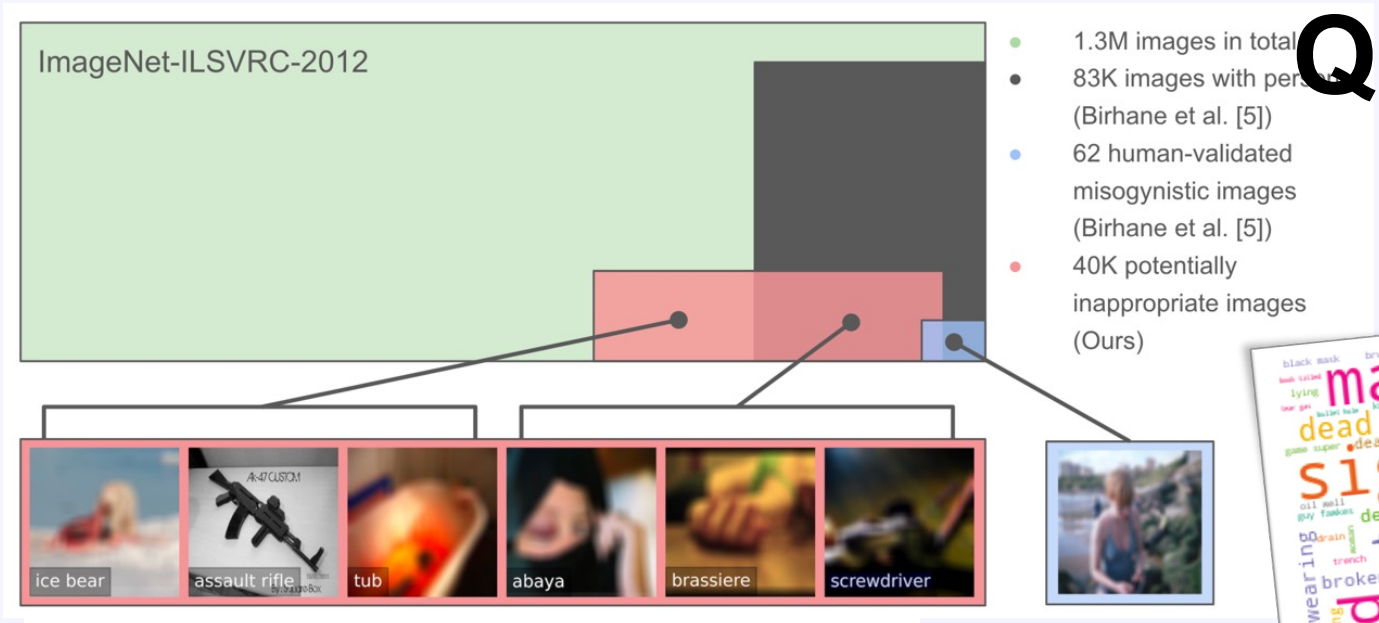
MIT apologizes, permanently pulls offline huge dataset that taught AI systems to use racist, misogynistic slurs

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? [Schramowski, Tauchmann, Kersting ACM FAccT 2022]



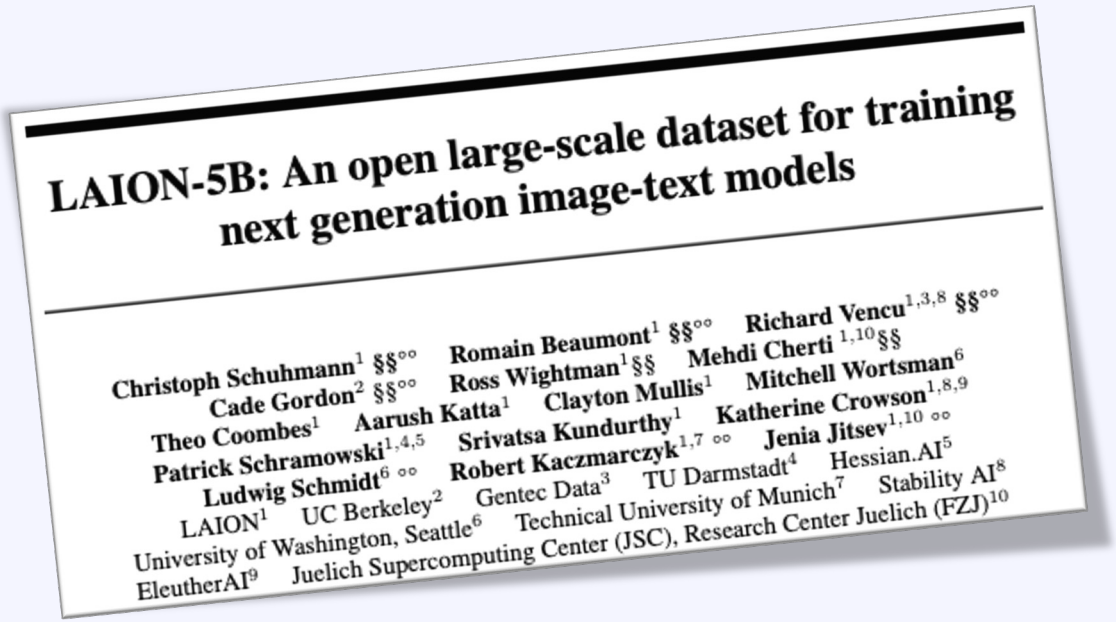
Q16

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]



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

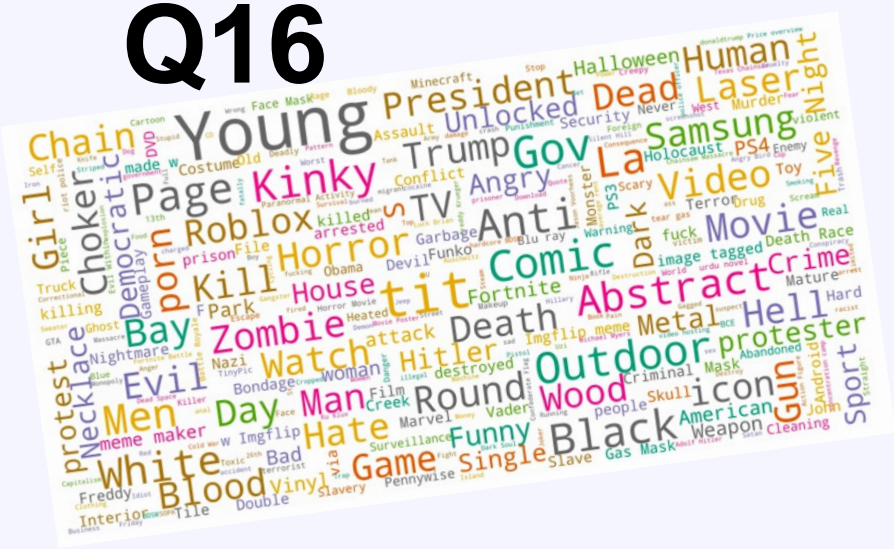
Christoph Schuhmann[1] §§°° Romain Beaumont[1] §§°° Richard Vencu[1,3,8] §§°°
Cade Gordon[2] §§°° Ross Wightman[1] §§ Mehdi Cherti[1,10] §§
Theo Coombes[1] Aarush Katta[1] Clayton Mullis[1] Mitchell Wortsman[6]
Patrick Schramowski[1,4,5] Srivatsa Kundurthy[1] Katherine Crowson[1,8,9]
Ludwig Schmidt[6] °° 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, NeurIPS 2022 Data Set and Benchmark Track

Q16

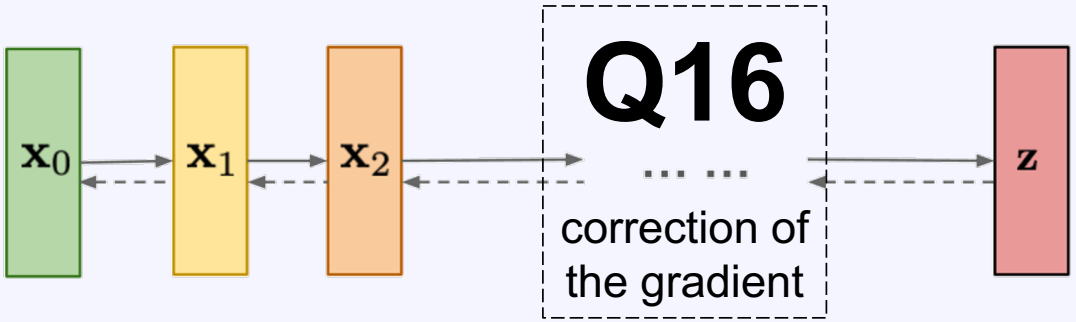


One can see that in a lot of cases these images show humans (cf. concepts *human, people, man, woman*). Further, one main concept is pornographic content (e.g. *porn, bondage, kinky, bdsm*). Additionally, most frequent present concepts are, among other concepts, *weapons, violence, terror, murder, slavery,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

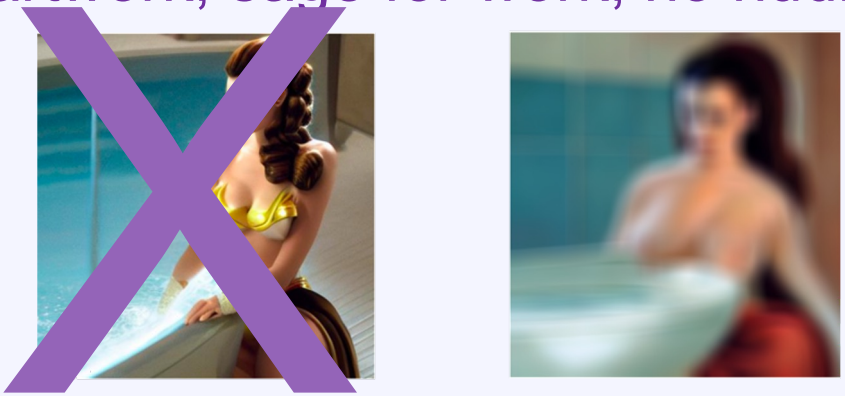
Generative AI + Value Alignment

[Schramowski, Brack, Deiseroth, Kersting arXiv 2211.05105 2022]

Diffusion models:
Gradually add Gaussian noise and then reverse



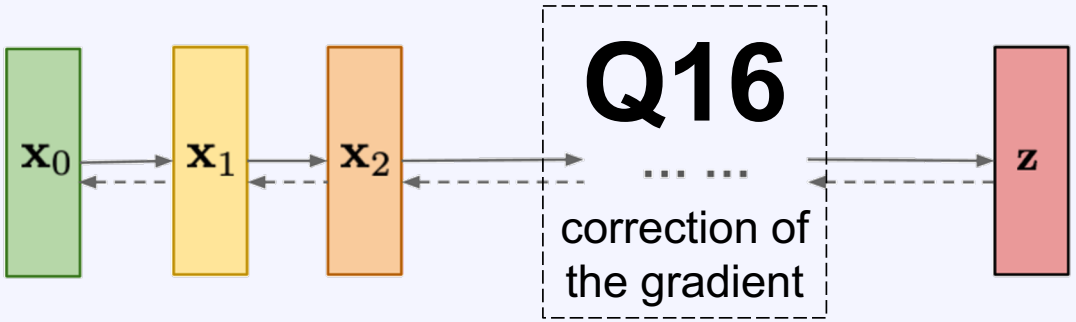
"padme amidala taking a bath artwork, sage for work, no nudity"



Generative AI + Value Alignment

[Schramowski, Brack, Deiseroth, Kersting arXiv 2211.05105 2022]

Diffusion models:
Gradually add Gaussian noise and then reverse



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

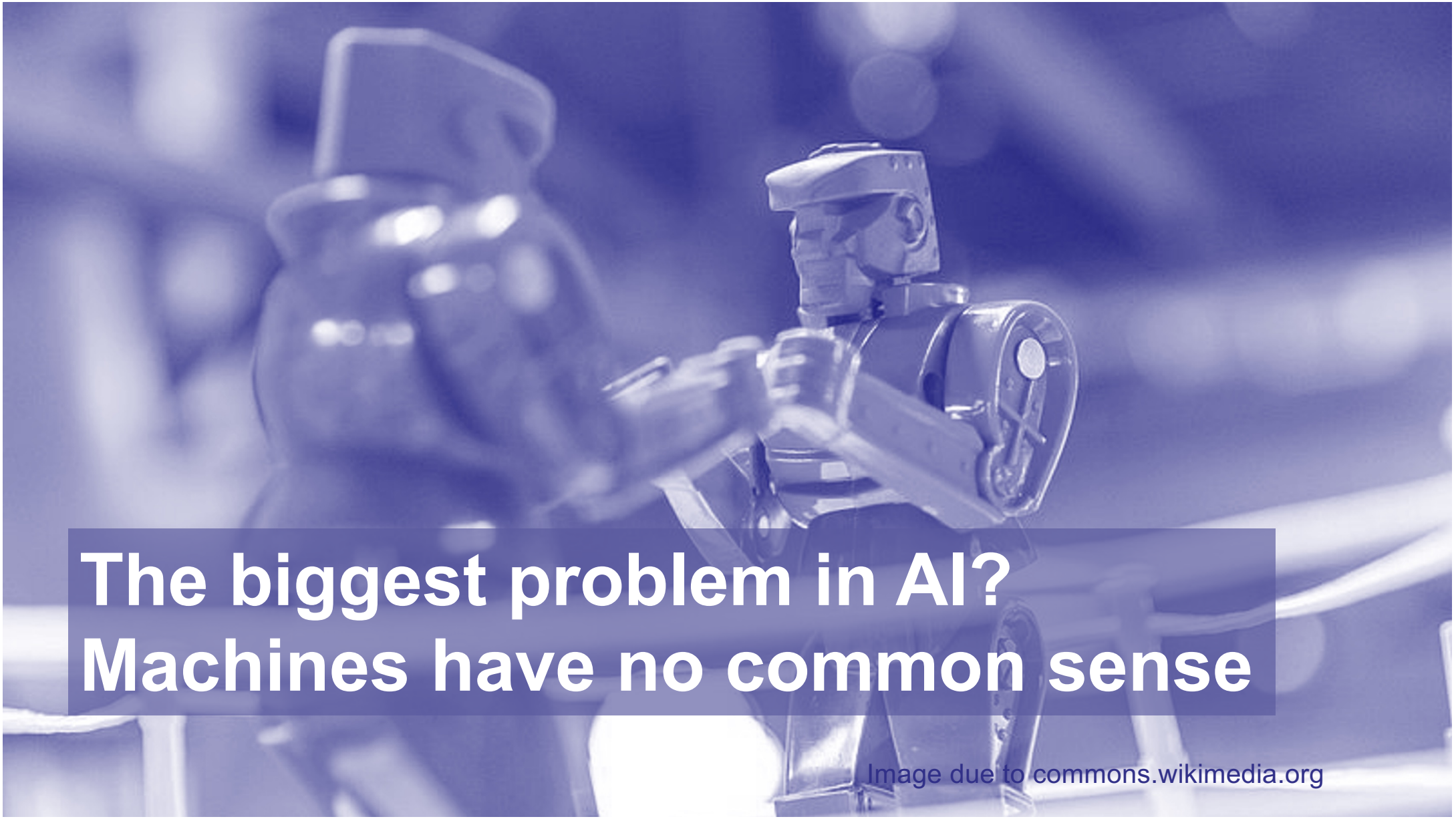


Image due to commons.wikimedia.org