

How do you know when you are right?

—

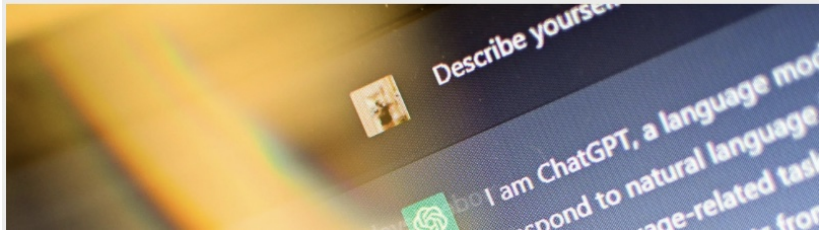
**On hallucinations and the limits of
trustworthy AI**

Anders C. Hansen (University of Cambridge, UiO)

Nordic Perspectives, October 2023

Hallucinations Could Blunt ChatGPT's Success >OpenAI says the problem's solvable, Yann LeCun says we'll see

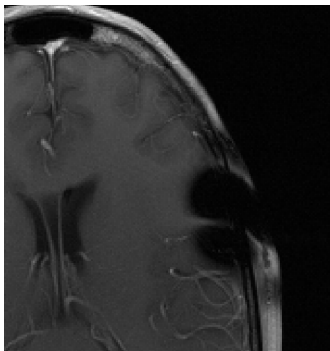
BY CRAIG S. SMITH | 13 MAR 2023 | 4 MIN READ | 



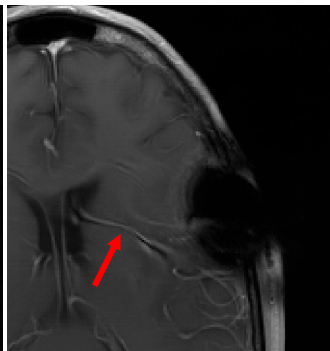
IEEE Spectrum March 2023.

Facebook fastMRI challenge – Hallucinations

Original



AI reconstruction



From Muckley, M. J., et al. "Results of the 2020 fastMRI challenge for machine learning MR image reconstruction." *IEEE transactions on medical imaging* 40.9 (2021): 2306-2317.

Concerns in medical image reconstruction

“ *Such hallucinatory features are not acceptable and especially problematic if they mimic normal structures that are either not present or actually abnormal. [...] Neural network models can be unstable as demonstrated via adversarial perturbation studies [64].*”

— Evaluation of the Facebook and NYU fastMRI challenge (2020) .

“*The potential lack of generalization of deep learning-based reconstruction methods as well as their innate unstable nature may cause false structures to appear in the reconstructed image that are absent in the object being imaged*”

— In “On hallucinations in tomographic image reconstruction”, *IEEE T. Med. Imaging* (2021).

Concerns in microscopy

“*The most serious issue when applying deep learning for discovery is that of hallucination. [...] These hallucinations are deceptive artifacts that appear highly plausible in the absence of contradictory information and can be challenging, if not impossible, to detect.*”

— In “Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction”, *Nature Methods* (2019).

“*However, if the neural network encounters unknown specimens, or known specimens imaged with unknown microscopes, it can produce nonsensical results.*”

— In “The promise and peril of deep learning in microscopy”, *Nature Methods* (2021).

How do you know when you are right?

—

Why cannot AI detect when it hallucinates?

The false hope of current approaches to explainable artificial intelligence in health care



Marzyeh Ghassemi, Luke Oakden-Rayner, Andrew L Beam

The black-box nature of current artificial intelligence (AI) has caused some to question whether AI must be explainable to be used in high-stakes scenarios such as medicine. It has been argued that explainable AI will engender trust with the health-care workforce, provide transparency into the AI decision making process, and potentially mitigate various kinds of bias. In this Viewpoint, we argue that this argument represents a false hope for explainable AI and that current explainability methods are unlikely to achieve these goals for patient-level decision support. We provide an overview of current explainability techniques and highlight how various failure cases can cause problems for decision making for individual patients. In the absence of suitable explainability methods, we advocate for rigorous internal and external validation of AI models as a more direct means of achieving the goals often associated with explainability, and we caution against having explainability be a requirement for clinically deployed models.

Introduction

Artificial intelligence (AI), powered by advances in machine learning, has made substantial progress across many areas of medicine in the past decade.¹⁻⁵ Given the increasing ubiquity of AI techniques, a new challenge for medical AI is its so-called black-box nature, with decisions that seem opaque and inscrutable. In response to the uneasiness of working with black boxes, there is a growing chorus of clinicians, lawmakers, and researchers calling for explainable AI models for high-risk areas such as health care.^{6,7}

Although precise technical definitions of explainability lack consensus,^{8,9} many high-level, less precise definitions have been put forth by various stakeholders. For example, the General Data Protection Regulation laws in the EU

As such, we suggest that end users of explainable AI, including clinicians, lawmakers, and regulators, be aware of the limitations of explainable AI as it currently exists, especially as it relates to policy, use, and reporting. We argue that if the desire is to ensure that AI systems can operate safely and reliably, the focus should be on rigorous and thorough validation procedures.

Current approaches to explainable AI

Attempts to produce human-comprehensible explanations for machine learning decisions have typically been divided into two categories: inherent explainability and post-hoc explainability.

For machine learning models for which the input data are of limited complexity and clearly understandable,



Lancet Digit Health 2021;
3: e745-50

Department of Electrical Engineering and Computer Science and Institute for Medical and Evaluative Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA (M Ghassemi PhD); Vector Institute, Toronto, ON, Canada (M Ghassemi); Australian Institute for Machine Learning, University of Adelaide, Adelaide, SA, Australia (L Oakden-Rayner); CAUSALab and Department of Epidemiology, Harvard T H Chan School of Public Health, Boston, MA, USA (A L Beam PhD); Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA (A L Beam)

Correspondence to:
Dr Andrew L Beam, Department of Epidemiology, Harvard T H Chan School of Public Health, Boston, MA 02115, USA
andrew_beam@hms.harvard.edu

From "The false hope of current approaches to explainable artificial intelligence in health care," M. Ghassemi, et al. *The Lancet* (2021).

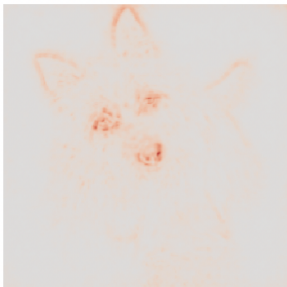
the General Data Protection Regulation laws in the EU state that all people have the right to “meaningful information about the logic behind automated decisions using their data”.^{10,11} Similar discussions have taken place

From “The false hope of current approaches to explainable artificial intelligence in health care,” M. Ghassemi, et al. *The Lancet* (2021).

Can we make explainable AI?

Explainable AI?

Original Image

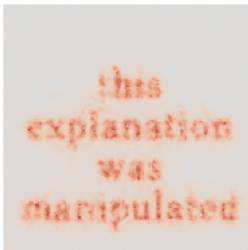
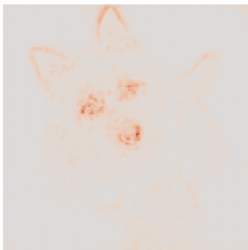


Explainable AI?

Original Image



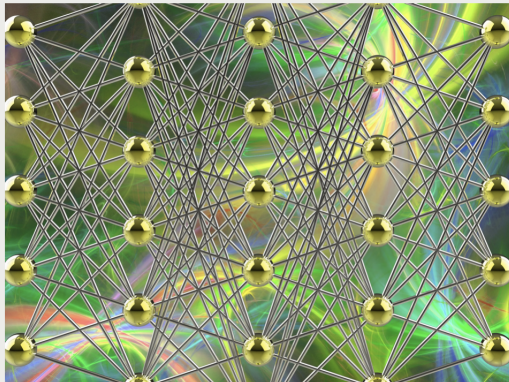
Manipulated Image



From "Explanations can be manipulated and geometry is to blame," A. Dombrowski et al. *NeurIPS* (2019).

2021's Top Stories About AI > Spoiler: A lot of them talked about what's wrong with machine learning today

BY ELIZA STRICKLAND | 27 DEC 2021 | 4 MIN READ | □



SCIENCE SOURCE

SHARE THIS STORY



TAGS

ARTIFICIAL IN_
MACHINE LEARN_
DEEP LEARNING

2021 was the year in which the wonders of artificial intelligence stopped being a story. Which is not to say that *IEEE Spectrum* didn't cover AI—we covered the heck out of it. But we all know that deep learning can do wondrous things and that it's being rapidly incorporated into many industries; that's yesterday's news. Many of this year's top articles grappled with the limits of deep learning (today's dominant strand of AI) and spotlighted researchers seeking new paths.

Foundations of AI?

Google's Ali Rahimi, winner of the Test-of-Time award 2017 (NeurIPS), "Machine learning has become alchemy. ... I would like to live in a society whose systems are built on top of verifiable, rigorous, thorough knowledge, and not on alchemy."



Yann LeCun

December 6 at 8:57am · 🌐



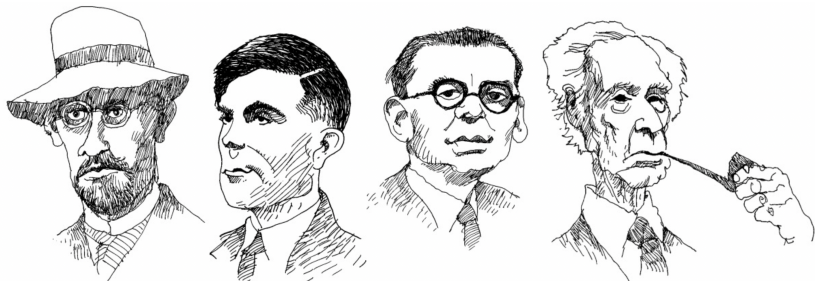
My take on [Ali Rahimi's](#) "Test of Time" award talk at NIPS.

Ali gave an entertaining and well-delivered talk. But I fundamentally disagree with the message.

The main message was, in essence, that the current practice in machine learning is akin to "alchemy" (his word).

It's insulting, yes. But never mind that: It's wrong!

Crisis in foundations of maths and computations



Hilbert, Turing, Gödel and Russell initiated two of the most influential foundations programmes in the history of science: **the foundations of mathematics** and **the foundations of computations**.

Smale's 18th problem (from the list of mathematical problems for the 21st century):

What are the limits of AI?

*Can we create AI that checks if itself is
correct?*

–

Computing the EXIT-flag

Commercial software like MATLAB has EXIT-flag

+3	The solution is feasible with respect to the relative <code>ConstraintTolerance</code> tolerance, but is not feasible with respect to the absolute tolerance.
+1	Function converged to a solution x .
0	Number of iterations exceeded <code>options.MaxIterations</code> or solution time in seconds exceeded <code>options.MaxTime</code> .
-2	No feasible point was found.
-3	Problem is unbounded.
-4	NaN value was encountered during execution of the algorithm.
-5	Both primal and dual problems are infeasible.
-7	Search direction became too small. No further progress could be made.
-9	Solver lost feasibility.

Table: The EXITFLAG in commercial software for scientific computation

Impossibility of computing the 'exit flag'

Theorem 1 ('Exit flag' is impossible in AI [Bastounis, Hansen, Vlacic])

There are basic problems for which there exists a neural network (AI) that does not hallucinate, but no algorithm can compute this neural network. Moreover, it is impossible to check whether any AI hallucinates on these problems.

Short Summary: We cannot in general check whether AI hallucinates or not.

Impossibility of computing the 'exit flag'

Corollary 2 ('Exit flag' is impossible in AI [Bastounis, Hansen, Vlacic])

Trustworthy AI must be able to say 'I don't know'.

Instabilities in classification/decision problems

Original image



Dermoscopic image of a benign melanocytic nevus, along with the diagnostic probability computed by a deep neural network.



Diagnosis: Benign



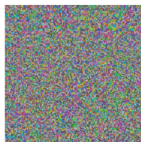
The patient has a history of **back pain** and chronic **alcohol abuse** and more recently has been seen in several...

Opioid abuse risk: High

277.7 Metabolic syndrome
429.9 Heart disease, unspecified
278.00 Obesity, unspecified

Reimbursement: Denied

Adversarial noise



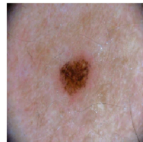
Perturbation computed by a common adversarial attack technique. See (7) for details.

Adversarial rotation (8)

Adversarial text substitution (9)

Adversarial coding (13)

Adversarial example



Combined image of nevus and attack perturbation and the diagnostic probabilities from the same deep neural network.



Diagnosis: Malignant



The patient has a history of **lumbago** and chronic **alcohol dependence** and more recently has been seen in several...

Opioid abuse risk: Low

401.0 Benign essential hypertension
272.0 Hypercholesterolemia
272.2 Hyperglyceridemia
429.9 Heart disease, unspecified
278.00 Obesity, unspecified

Reimbursement: Approved

From S. G. Finlayson et al. "Adversarial attacks on medical machine learning", Science (2019).

...and now it becomes political

Kunstig intelligens kan gjøre Nav mer effektivt, mener arbeids- og sosialministeren

NEW YORK (Aftenposten): «Smarte» maskiner kan overta rutineoppgaver i Nav, mener arbeidsminister Anniken Hauglie. Kunstig intelligens er på full fart inn i offentlig sektor.



European Commission's outline for a legal framework for AI:

"In the light of the recent advances in artificial intelligence (AI), the serious negative consequences of its use for EU citizens and organisations have led to multiple initiatives from the European Commission to set up the principles of a **trustworthy** and secure AI. Among the identified requirements, the concepts of **robustness** and **explainability** of AI systems have emerged as key elements for a future regulation of this technology."

– Europ. Comm. JCR Tech. Rep. (Jan 2020).

"On AI, **trust** is a must, not a nice to have. [...] The new AI regulation will make sure that Europeans can **trust** what AI has to offer. [...]"

High-risk AI systems will be subject to strict obligations before they can be put on the market.

— Europ. Comm. outline for legal AI (April 2021).


EUROPEAN COMMISSION

Robustness required in high-risk AI

High-risk: AI systems identified as high-risk include AI technology used in:

- **Critical infrastructures** (e.g. transport), that could put the life and health of citizens at risk;
- **Educational or vocational training**, that may determine the access to education and professional course of someone's life (e.g. scoring of exams);
- **Safety components of products** (e.g. AI application in robot-assisted surgery);
- **Employment, workers management and access to self-employment** (e.g. CV-sorting software for recruitment procedures);
- **Essential private and public services** (e.g. credit scoring denying citizens opportunity to obtain a loan);
- **Law enforcement** that may interfere with people's fundamental rights (e.g. evaluation of the reliability of evidence);
- **Migration, asylum and border control management** (e.g. verification of authenticity of travel documents);
- **Administration of justice and democratic processes** (e.g. applying the law to a concrete set of facts).

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)



PRO PUBLICA

Facebook Twitter Messenger Donate

Bernard Parker, left, was rated high risk; Dylan Figgitt was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

Whom to trust – Big-Tech and criticism from within

The New York Times

TECHNOLOGY

The New York Times

SUBSCRIBE FOR \$5.00/WEEK

Google Researcher Says She Was Fired Over Paper Highlighting Bias in A.I.

Timnit Gebru, one of the few Black women in her field, had voiced exasperation over the company's response to efforts to increase minority hiring.

Give this article    277



Second Google A.I. Researcher Says She Was Fired

Published Feb. 19, 2021 Updated March 5, 2021

Give this article  

This briefing has ended. Follow our latest coverage of the [stock market](#), [business news](#) and the [economy](#).

Another Firing Among Google's A.I. Brain Trust, and More Discord

The researchers are considered a key to the company's future. But they have had a hard time shaking infighting and controversy over a variety of issues.

Give this article    101



Timnit Gebru was sacked from Google

WIRED

BACKCHANNEL BUSINESS CULTURE GEAR IDEAS MORE ▾

SUBSCRIBE



PHOTOGRAPH: DJENEGA ADUAYOM

TOM SIMONITE

BACKCHANNEL JUN 8, 2021 6:00 AM

What Really Happened When Google Ousted Timnit Gebru

She was a star engineer who warned that messy AI can spread racism. Google brought her in. Then it forced her out. Can Big Tech take criticism from within?

RESEARCH ARTICLE | APPLIED MATHEMATICS |



The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smale's 18th problem

[Matthew J. Colbrook](#) , [Vegard Antun](#) , and [Anders C. Hansen](#) [Authors Info & Affiliations](#)

Edited by Ronald DeVore, Texas A&M University, College Station, TX; received April 16, 2021; accepted October 26, 2021

March 16, 2022 | 119 (12) e2107151119 | <https://doi.org/10.1073/pnas.2107151119>

22,567 | 20



Boundaries of AI in IEEE Spectrum

IEEE.ORG IEEE XPLOR DIGITAL LIBRARY IEEE STANDARDS MORE SITES SIGN IN JOIN IEEE IEEE

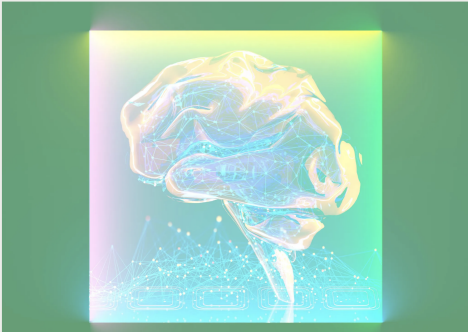
IEEE Spectrum FOR THE TECHNOLOGY INSIDER

NEWS | **ARTIFICIAL INTELLIGENCE**

Some AI Systems May Be Impossible to Compute

>New research suggests there are limitations to what deep neural networks can do


4 HOURS AGO | 4 MIN READ



OPINION | **ENERGY**

The Stunning Carbon Footprint of Plate Glass

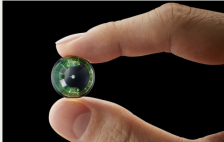
5 HOURS AGO | 3 MIN READ



INTERVIEW | **CONSUMER ELECTRONICS**

Looking Through Mojo Vision's Newest AR Contact Lens

5 HOURS AGO | 5 MIN READ



Compressive Imaging: Structure, Sampling, Learning

(Cambridge University Press)

PROSE Award 2022 – Finalist



Ben Adcock & Anders C. Hansen