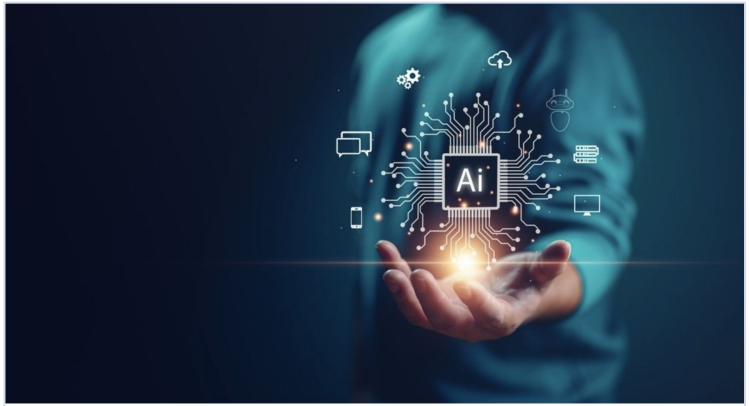
presented in

Track 2

AI Terminology and B.S. Detector

What we will cover in this module

- A taxonomy of AI B.S.
- Claims of AI accuracy
- AI Safety and Doomerism Narratives
- AI >= Human Comparisons



Gen-AI tools deliver 'unprecedented, near-perfect' data accuracy

Latest results deliver 99% accuracy, ready for enterprise decisioning following benchmark test.

3 months ago Posted in

Accuracy

PR Newswire

Seekr and Intel Collaborate to Deliver Trusted, Responsible AI Solutions for Enterprise Customers

seekr | intel

NEWS PROVIDED BY
Seekr Technologies →
Apr 16, 2024, 22:13 ET

A new collaboration of compute and responsible AI combine to accelerate the deployment of foundation models with less error and bias.

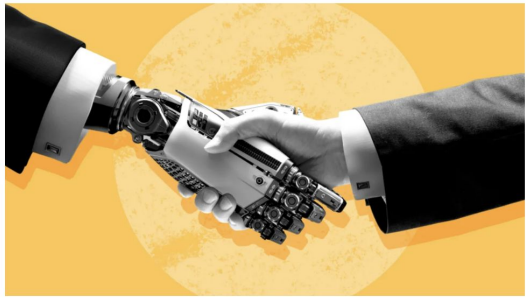
AI Safety

CNN Health Watch Listen Live TV Sign in

AI may be as effective as medical specialists at diagnosing disease

By Jack Guy, CNN

3 minute read · Published 7:23 AM EDT, Wed September 25, 2019



AI >= Humans


“Magic”
“Groundbreaking”
“Revolutionary”
“99% accurate”
“Error-free”
“Huge Savings”
“Boosted Productivity by 50%”

AI ACCURACY: Questions to ask

Who came up with this number?

The company itself? A peer-reviewed study? Researchers contracted by the company?

- ***Hippo:*** Hippo is a home insurance company that uses AI to identify fraudulent claims with 99% accuracy.

Forbes

FORBES > INNOVATION

BREAKING

New Tool Can Tell If Something Is AI-Written With 99% Accuracy

Arianna Johnson Forbes Staff

Johnson is a reporter on the Forbes news desk who covers explainers.

Follow

AI ACCURACY: Questions to ask

What is the baseline that you are comparing against?

EX: If the previous method was 96% accurate, 97% may not be much of an improvement.

- ***Hippo:*** Hippo is a home insurance company that uses AI to identify fraudulent claims with 99% accuracy.

Forbes

FORBES > INNOVATION

BREAKING

New Tool Can Tell If Something Is AI-Written With 99% Accuracy

Arianna Johnson Forbes Staff

Johnson is a reporter on the Forbes news desk who covers explainers.

Follow

AI ACCURACY: Questions to ask

What are the false positive, false negative rates, and precision?

There are many different metrics to measure ‘accuracy.’

- ***Hippo:*** Hippo is a home insurance company that uses AI to identify fraudulent claims with 99% accuracy.

Forbes

FORBES > INNOVATION

BREAKING

New Tool Can Tell If Something Is AI-Written With 99% Accuracy

Arianna Johnson Forbes Staff

Johnson is a reporter on the Forbes news desk who covers explainers.

Follow

AI ACCURACY: Questions to ask

Has accuracy been tested across groups? Does the accuracy differ across groups?

A model may be 97% accurate for one group of people while only being 80% accurate for another.

- ***Hippo:*** Hippo is a home insurance company that uses AI to identify fraudulent claims with 99% accuracy.

Forbes

FORBES > INNOVATION

BREAKING

New Tool Can Tell If Something Is AI-Written With 99% Accuracy

Arianna Johnson Forbes Staff

Johnson is a reporter on the Forbes news desk who covers explainers.

Follow

AI ACCURACY: Questions to ask

How many people is the tool applied against?

99% accuracy (and 1% error rate) for tens of millions of people can means 100,000 thousands of wrong decisions.

- ***Hippo:*** Hippo is a home insurance company that uses AI to identify fraudulent claims with 99% accuracy.

Forbes

FORBES > INNOVATION

BREAKING

New Tool Can Tell If Something Is AI-Written With 99% Accuracy

Arianna Johnson Forbes Staff

Johnson is a reporter on the Forbes news desk who covers explainers.

Follow

AI ACCURACY: Questions to ask

Has the model been tested in the ‘real world’?

A model may have high accuracy on internal testing data (which it may also be trained on) but low accuracy once applied in the real world.

- ***Hippo:*** Hippo is a home insurance company that uses AI to identify fraudulent claims with 99% accuracy.

Forbes

FORBES > INNOVATION

BREAKING

New Tool Can Tell If Something Is AI-Written With 99% Accuracy

Arianna Johnson Forbes Staff

Johnson is a reporter on the Forbes news desk who covers explainers.

Follow

Accuracy claims often have revealing fine print

JANUARY 25, 2023

Editors' notes

A new AI tool can predict mosquitoes' ages with 98% accuracy to speed malaria research

by SciDev.Net

However, the scientists stress that further research is needed since the study looked at only one specific type of mosquito, Anopheles arabiensis, obtained from only two countries.

Accuracy claims often have revealing fine print

ShotSpotter collects are actual gunfire. ShotSpotter, Inc. continues to claim a supposed 97% "accuracy" rate. To that end, they paid a consultant to "audit" this number. But the audit actually confirmed that this is not an accuracy rate at all, reflecting no actual testing of the system. Instead, it showed that ShotSpotter just assumes that all of its alerts are gunshots unless local police happen to file a complaint with the company reporting a potential error. ShotSpotter's supposed "accuracy" claims are actually just a tally of customer complaints. And, according to Manes, CPD does not report any of

Accuracy claims often have revealing fine print

Amr Awadallah, the chief executive of Vectara, warns that its chatbot software doesn't always tell the truth. Cayce Clifford for The New York Times

Chatbots May 'Hallucinate' More Often Than Many Realize

When summarizing facts, ChatGPT technology makes things up about 3 percent of the time, according to research from a new start-up. A Google system's rate was 27 percent.

Share full article

By Cade Metz
Cade Metz has been watching chatbots hallucinate since 2017.

Published Nov. 6, 2023 Updated Nov. 16, 2023

Leer en español

FORTUNE

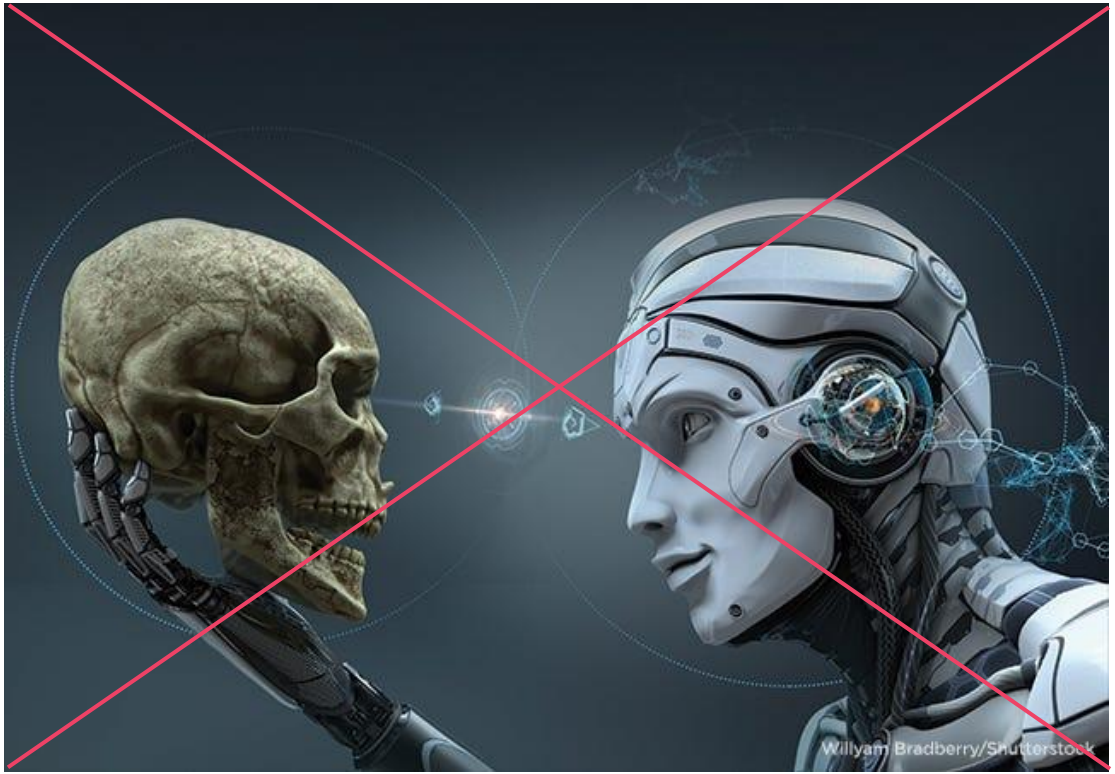
CONFERENCES · A.I.

AI hallucinations will be solvable within a year, ex-Google AI researcher says—but that may not be a good thing: 'We want them to propose things that are weird and novel'

BY ORIANNA ROSA ROYLE
April 16, 2024 at 3:10 PM GMT+3

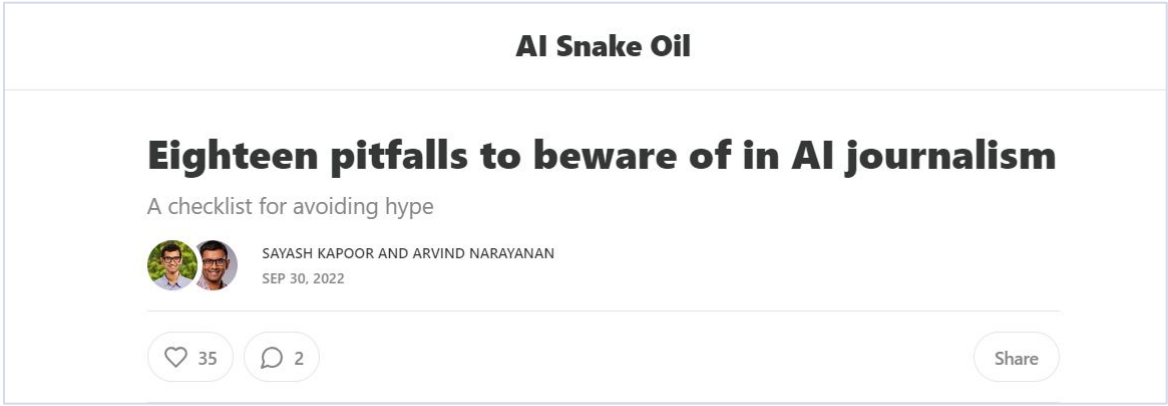

“Intentions ”
“Rational”
“Sentient”
“Human-like AI”
“Outperforms Humans”
“AI understands human emotions”
“Objective”

Anthropomorphizing AI



It’s easy to reproduce these tropes in our reporting.

Sayash Kapoor and Arvind Narayanan’s *AI Snakeoil* substack is a must read for journalists in this space.



They outline 4 pitfalls.

Pitfall 1. Attributing agency to AI: Describing AI systems as taking actions independent of human supervision or implying that they may soon be able to do so.

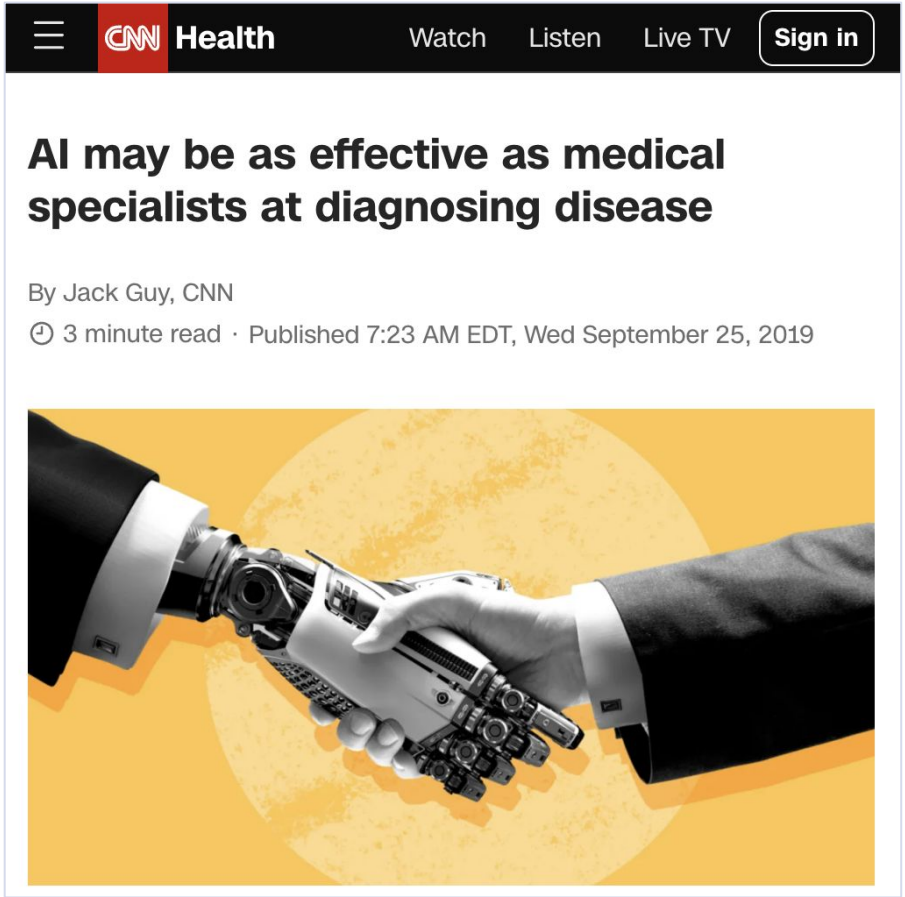
Pitfall 2. Suggestive imagery: Images of humanoid robots are often used to illustrate articles about AI, even if the article has nothing to do with robots. This gives readers a false impression that AI tools are embodied, even when it is just software that learns patterns from data.

Pitfall 3. Comparison with human intelligence: In some cases, articles on AI imply that AI algorithms learn in the same way as humans do. For example, comparisons of deep learning algorithms with the way the human brain functions are common. Such comparisons can lend credence to claims that AI is "sentient", as Dr. Timnit Gebru and Dr. Margaret Mitchell note in their recent op-ed.

Pitfall 4. Comparison with human skills: Similarly, articles often compare how well AI tools perform with human skills on a given task. This falsely implies that AI tools and humans compete on an equal footing—hiding the fact that AI tools only work in a narrow range of settings.

Catching these tropes in our work, and catching them in company or government PR is important.

Because many companies hype their AI products by anthropomorphizing them.



CNN Health Watch Listen Live TV Sign in

AI may be as effective as medical specialists at diagnosing disease

By Jack Guy, CNN

3 minute read · Published 7:23 AM EDT, Wed September 25, 2019

Leading to the flawed assumption that AI can feel and think. Or that AI is innately objective and rational.


“Friendly AI”
“Existential Threat”
“Interpretable ”
“Responsible AI”
“AI Safety”
“Red Teamed”
“Aligned LLM”

Definitions

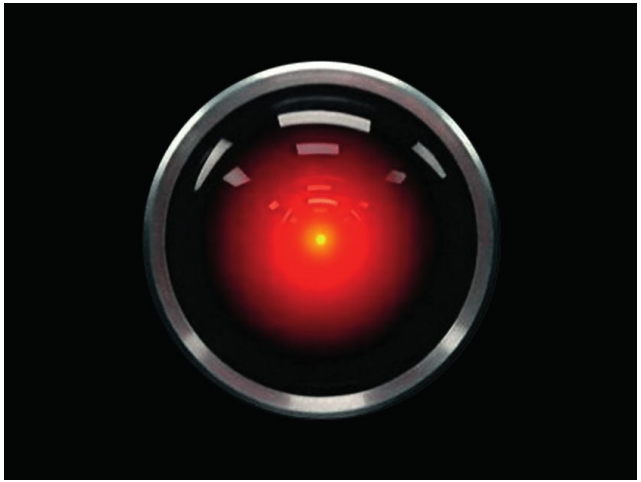
- Value Aligned AI: encoding “human values” into an AI model (usually used in the context of LLMs)

- AI Safety*: Value aligning AI to protect humanity from ‘superhuman’ and ‘existential’ threat of AI.

** AI safety is sometimes used to more generally describe attempts to reign in societal harms from AI.*

Of “AI Doomerism” and Techno Solutionism

AI Doomerism distracts from the current and concrete harms related to AI deployments.

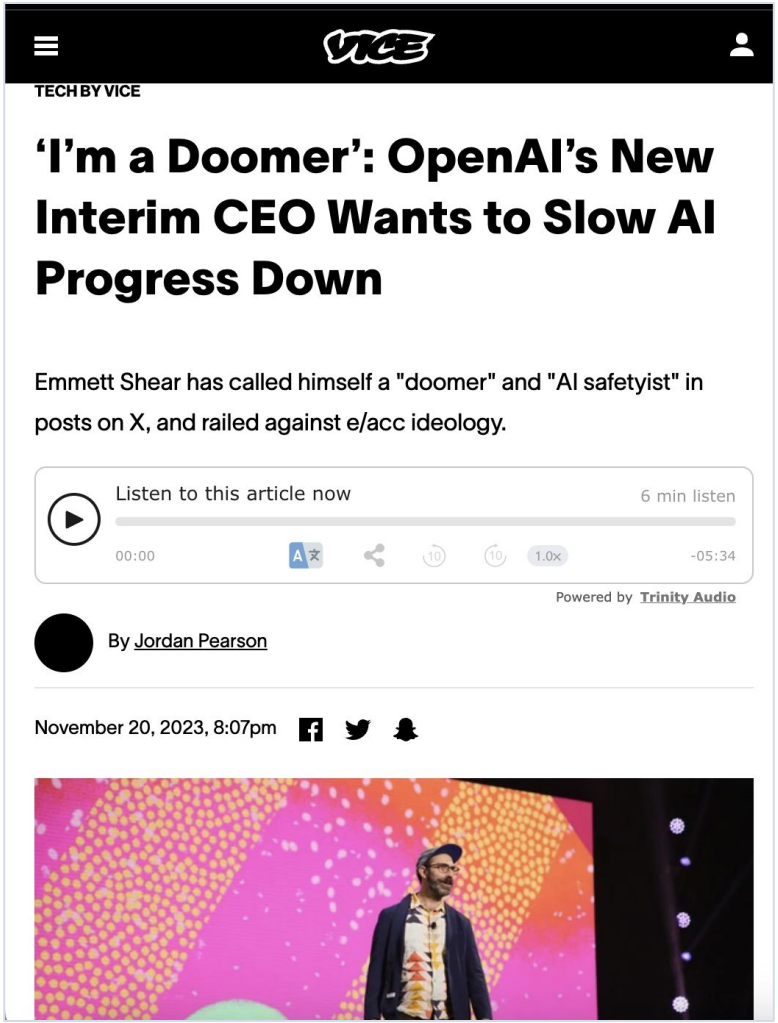


An open letter with signatures from hundreds of the biggest names in tech, including Elon Musk, has urged the world’s leading artificial intelligence labs to pause the training of new super-powerful systems for six months, saying that recent advances in AI present “profound risks to society and humanity.”

The letter comes just two weeks after the public release of OpenAI’s GPT-4, the most powerful AI system ever released, which has led researchers to slash their expectations for when AGI—or artificial general intelligence that surpasses human cognitive ability—will arrive. Many experts fear that, as an AI arms race heats up, humanity is sleepwalking into catastrophe.

Of “AI Doomerism” and Techno Solutionism

AI Doomerism distracts from the current and concrete harms related to AI deployments.



VICE

TECH BY VICE

‘I’m a Doomer’: OpenAI’s New Interim CEO Wants to Slow AI Progress Down

Emmett Shear has called himself a "doomer" and "AI safetyist" in posts on X, and railed against e/acc ideology.

Listen to this article now 6 min listen

00:00 1.0× -05:34

Powered by Trinity Audio

By Jordan Pearson

November 20, 2023, 8:07pm

AI safety is not a model property

Trying to make an AI model that can't be misused is like trying to make a computer that can't be used for bad things

ARVIND NARAYANAN AND SAYASH KAPOOR

MAR 12, 2024

AI Safety

- Positions AI as an existential, superhuman threat.
- Solution is technical.

AI Ethics

- Focused on present-tense AI harms.
- Solution is sociotechnical (ie. not just about the technology, but the context of how and where it is deployed.)

AI Auditing

- AI auditing can be a powerful accountability tool, but we should also be healthily skeptical of some claims.
- Companies and governments will often claim that their product has been audited for safety and/or bias.
- But we shouldn’t take these claims to mean that their systems are ethical or unbiased.
- Who has done the auditing? Internal or external? Adversarial? Are the auditors being paid?

AI Auditing

FINANCIAL ADVISORY SERVICES

Deloitte.

Combating welfare fraud with machine learning

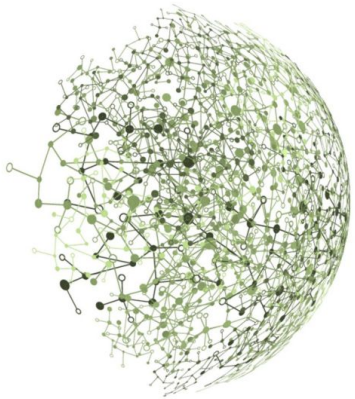
A government organisation in the Netherlands had to contend with cases of fraud. The organisation was responsible for paying allowances to citizens, but hundreds of millions of euros were being paid to ineligible persons. An internal department had identified that processes were not running effectively, and in the aftermath of negative publicity it was clear that the method of tackling fraud needed to be improved.

NEWS > TECHNOLOGY

Dutch scandal serves as a warning for Europe over risks of using algorithms

The Dutch tax authority ruined thousands of lives after using an algorithm to spot suspected benefits fraud — and critics say there is little stopping it from happening again.

Deloitte.



Municipality of Amsterdam

Reporting Analysis Algorithm with the application for a Living allowance

Published version (shortened version for the algorithm register)

MAKING AN IMPACT THAT MATTERS since 1845

Tips for seeing through PR AI hype

- Do not take company claims at face value. Ask for specific evidence.
- Understand which metrics a company is using to support claims of accuracy.
- Focus on concrete, present-tense harms rather than distracting “existential” doomerism.
- Don’t just speak to the company, speak to experts and critics who are independently funded.

Activity:
Pick apart a PR pitch
15 minutes

A startup has developed AI software that reads chest X-rays and claims to be 99 percent accurate. The startup CEO says that after running a trial at a hospital, the software demonstrated that it will save lives and cut costs because nurses will be able to detect lung cancer or other problems with the same proficiency as expert doctors.

data	+	compute	→	AI models	→	applications

What data was this software trained on?

Who did that data come from?

Did they consent?

What were their demographics?

How much did it cost to develop the software?

Where is the patient data being uploaded?

What happens once it is uploaded into this system?

What exactly do the algorithms inside this system do?

Who says it's 99 percent accurate?

Was the software tested in (similar to) real world conditions?

What monitoring is in place to detect errors?

How does the system save money, and for whom? Has that been proven?

How do nurses/doctors feel about this system?

Can patients opt out?

How much does the software cost? How does the startup get paid?

A city government plans to use AI to evaluate applications for low-income housing. The government says the system will automatically accept those that are legitimate and flag those that are fraudulent for investigation; this will completely remove any bias from the process.

data + compute →	AI models →	applications
What data was this model trained on? Who did that data come from? Did they consent? What were their demographics?	What kind of algorithm are they using? Can you get access to the algorithm? Has the software been audited for accuracy and bias? How reliable is the software? Is that enough for this purpose?	Exactly what tasks is it performing and what do humans still have to do? Who is using it on whom? Since implementing the software, has there been a change in who gets accepted or rejected? **Do you really need AI, or is there an alternative that would be cheaper, more transparent, more reliable?**

Resources

- AI Snake Oil
- Eighteen pitfalls to beware of in AI journalism
- How to report better on artificial intelligence - Columbia Journalism Review
- 3 Easy Ways to Evaluate AI Claims - IEEE Spectrum
- Emily M. Bender – Medium
- On the Dangers of Stochastic Parrots | Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency

AI Spotlight Series

The AI Accountability Network

Pulitzer Center