

A decorative graphic resembling a circuit board or a stylized letter 'U' is drawn in a dark blue color. It features several vertical and horizontal lines with small circles at the ends, suggesting solder points or components. The pattern is symmetrical and frames the central text.

apresentado em

Faixa 2

Terminologia de IA e detector de besteiras

O que abordaremos neste módulo

- Uma taxonomia das mentiras sobre IA
- Afirmações sobre a precisão da IA
- Segurança da IA e narrativas pessimistas
- AI $\geq$ comparações humanas





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

Precisão

PR Newswire

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



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

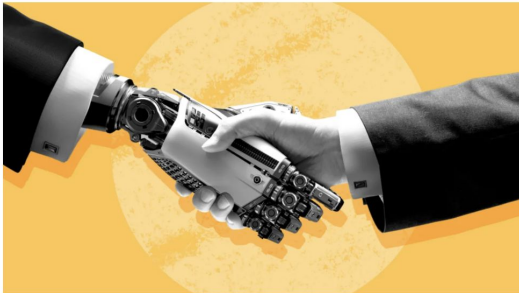
A new collaboration of compute and responsible AI combine to accelerate the deployment of foundation models with less error and bias.

Segurança da IA

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

A robotic hand shaking a human hand. The robotic hand is white and black, with visible joints and sensors. The human hand is in a dark suit sleeve. The background is a bright yellow circle on a dark background.

IA >= Humanos

"Mágico"

"Inovador"

"Revolucionário"

"99% preciso"

"Sem erros"

**"Economia
significativa"**

**"Aumento da
produtividade em 50%"**

PRECISÃO DA IA: Perguntas a fazer

Quem chegou a esse número?

A própria empresa? Um estudo revisado por pares? Pesquisadores contratados pela empresa?

- **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

PRECISÃO DA IA: Perguntas a fazer

Qual é a sua base de comparando?

EX: Se o método anterior tinha 96% de precisão, 97% pode não ser uma grande melhoria.

- **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

PRECISÃO DA IA: Perguntas a fazer

Quais são as taxas de falsos positivos, falsos negativos e precisão?

Existem muitas métricas diferentes para medir a “precisão”.

- **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

PRECISÃO DA IA: Perguntas a fazer

A precisão foi testada em todos os grupos? A precisão difere entre os grupos?

Um modelo pode ter 97% de precisão para um grupo de pessoas, mas apenas 80% para outro.

- **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

PRECISÃO DA IA: Perguntas a fazer

Em quantas pessoas a ferramenta é aplicada?

99% de precisão (e 1% de taxa de erro) para dezenas de milhões de pessoas pode significar 100 mil decisões erradas.

- **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

PRECISÃO DA IA: Perguntas a fazer

O modelo foi testado no
“mundo real”?

Um modelo pode ter
alta precisão em dados
de testes internos (nos
quais também pode ter
sido treinado), mas
baixa precisão quando
aplicado no mundo real.

- **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

As alegações
de precisão
geralmente
têm letras
miúdas
reveladoras

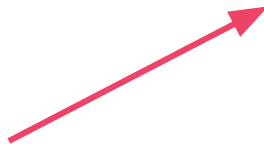
🕒 JANUARY 25, 2023

✓✓ Editors' notes

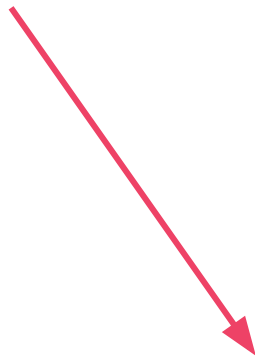
A new AI tool can predict mosquitoes' ages with 98% accuracy to speed malaria research

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.

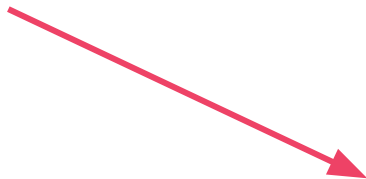


As alegações de precisão geralmente têm letras miúdas reveladoras



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

As alegações de precisão geralmente têm letras miúdas reveladoras



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



"Intenções"

"Racional"

"Sensível"

**"IA semelhante
à humana"**

**"Superior aos
humanos"**

**"A IA compreende
as emoções
humanas"**

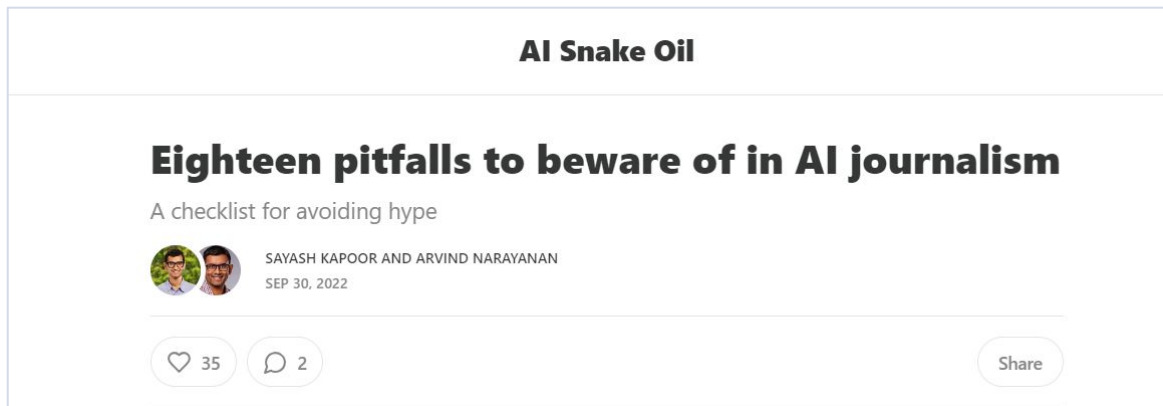
"Objetiva"

Antropomorfização da IA



É fácil reproduzir
esses clichês em
nossas reportagens.

O substack AI
Snakeoil, de Sayash
Kapoor e Arvind
Narayanan, é leitura
obrigatória para
jornalistas que
atuam nessa área.



Eles descrevem quatro armadilhas.

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.

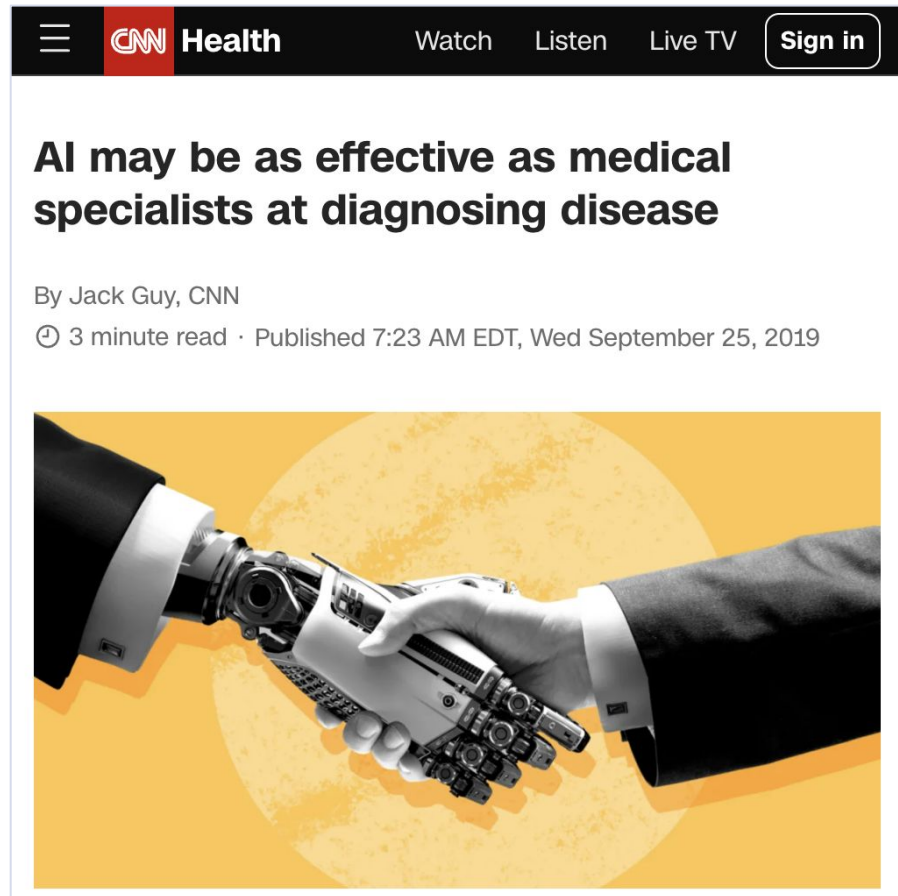
Pitfall 2. Suggestive imagery: Images of humanoid robots are often used to illustrate articles about AI, even if the article has nothing to do with robots. This gives readers a false impression that AI tools are embodied, even when it is just software that learns patterns from data.

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.

**É importante identificar
esses clichês em nosso
trabalho e identificá-los nas
assessorias de imprensa de
empresas ou governos.**

**Porque muitas
empresas promovem
seus produtos de IA
antropomorfizando-os.**



**Isso leva à
suposição errônea
de que a IA pode
sentir e pensar. Ou
que a IA é
inerentemente
objetiva e racional.**



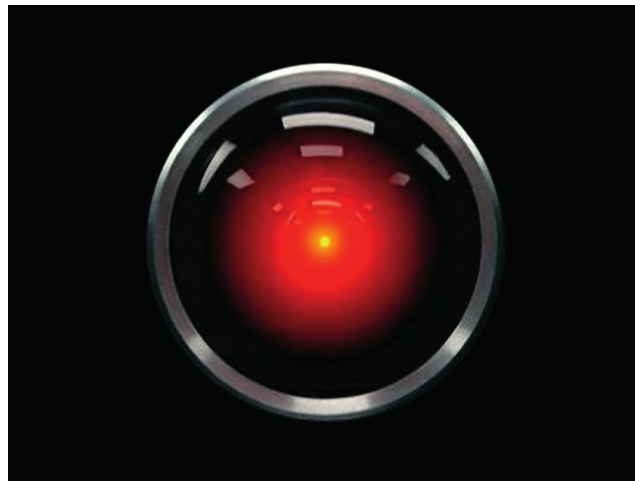
Definições

- IA alinhada com valores: codificar “valores humanos” em um modelo de IA (geralmente usado no contexto de LLMs)
- Segurança da IA*: alinhar valores da IA para proteger a humanidade de ameaças “sobre-humanas” e “existenciais” da IA.

** A segurança da IA às vezes é usada para descrever de forma mais geral as tentativas de controlar os danos sociais causados pela IA.*

Sobre o “pessimismo em relação à IA” e o tecnosolucionismo

O pessimismo em relação à IA nos distrai dos danos atuais e concretos relacionados à IA.



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.

Sobre o “pessimismo em relação à IA” e o tecnosolucionismo

O pessimismo em relação à IA nos distrai dos danos atuais e concretos relacionados à implantação da IA.



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



-05:34

Powered by [Trinity Audio](#)



By [Jordan Pearson](#)

November 20, 2023, 8:07pm   





AI Snake Oil



Subscribe

Sign in

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

Segurança da IA

- Posiciona a IA como uma ameaça existencial e sobre-humana.
- A solução é técnica.

Ética da IA

- Focada nos danos da IA no tempo presente.
- A solução é sociotécnica (ou seja, não se trata apenas da tecnologia, mas do contexto de como e onde ela é implantada).

Auditoria de IA

- A auditoria de IA pode ser uma ferramenta poderosa de responsabilização, mas também devemos ser saudavelmente céticos em relação a algumas alegações.
- Empresas e governos frequentemente afirmam que seus produtos foram auditados quanto à segurança e/ou viés.
- Mas não isso não significa que seus sistemas são éticos ou imparciais.
- Quem fez a auditoria? Interna ou externa? Adversária? Os auditores estão sendo pagos?

Auditoria de IA

FINANCIAL
ADVISORY SERVICES

Deloitte.

Combating welfare fraud with machine learning

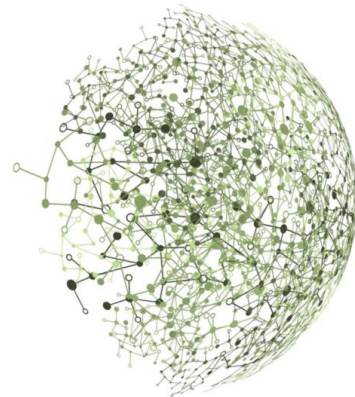
A government organisation in the Netherlands had to contend with cases of fraud. The organisation was responsible for paying allowances to citizens, but hundreds of millions of euros were being paid to ineligible persons. An internal department had identified that processes were not running effectively, 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)



Dicas para perceber o exagero dos relações públicas sobre IA

- Não acredite nas afirmações das empresas sem questionar. Peça evidências específicas.
- Entenda quais métricas uma empresa está usando para sustentar suas alegações de precisão.
- Concentre-se em danos concretos e presentes, em vez de se distrair com o pessimismo “existencial”.
- Não fale apenas com a empresa, converse com especialistas e críticos que são financiados de forma independente.

Atividade:
Analise um argumento
de relações públicas
15 minutos

Uma startup desenvolveu um software de IA que analisa radiografias de tórax e afirma ter 99% de precisão. O CEO da startup diz que, após um teste em um hospital, o software demonstrou que salvará vidas e reduzirá custos, pois os enfermeiros poderão detectar câncer de pulmão ou outros problemas com a mesma proficiência que médicos especialistas.

dados

+

calcular



modelos de IA



aplicações

Com quais dados esse software foi treinado?

De onde vieram esses dados?

Eles consentiram?

Quais eram suas características demográficas?

Quanto custou para desenvolver o software?

Para onde os dados dos pacientes estão sendo enviados?

O que acontece depois que eles são enviados para esse sistema?

O que exatamente os algoritmos dentro deste sistema fazem?

Quem afirma que ele tem 99% de precisão?

O software foi testado em condições (semelhantes às) do mundo real?

Que monitoramento existe para detectar erros?

Como o sistema economiza dinheiro e para quem? Isso foi comprovado?

Como os enfermeiros/médicos se sentem em relação a esse sistema?

Os pacientes podem optar por não participar?

Quanto custa o software?
Como a startup é remunerada?

Um governo municipal planeja usar IA para avaliar inscrições para moradias populares. O governo afirma que o sistema aceitará automaticamente as inscrições legítimas e sinalizará as fraudulentas para investigação, o que eliminará completamente qualquer parcialidade do processo.

dados

+

computação



modelos de IA



aplicações

Com quais dados esse modelo foi treinado?

De onde vieram esses dados?

Eles consentiram?

Quais eram suas características demográficas?

Como esse sistema foi treinado?

O governo está fazendo parceria com uma empresa privada, por exemplo, um provedor de nuvem, para treinar esse modelo?

Com que frequência ele será atualizado?

Que tipo de algoritmo eles estão usando?

É possível ter acesso ao algoritmo?

O software foi auditado quanto à precisão e imparcialidade?

Qual é o nível de confiabilidade do software? Isso é suficiente para esse fim?

Quais tarefas exatamente ele está realizando e o que os humanos ainda precisam fazer?

Quem está usando o software e em quem?

Desde a implementação do software, houve alguma mudança em quem é aceito ou rejeitado?

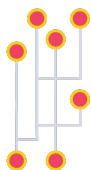
Você realmente precisa de IA ou existe uma alternativa mais barata, mais transparente e mais confiável?

Recursos

- [AI Snake Oil](#)
- [Dezoito armadilhas a serem evitadas no jornalismo com IA](#)
- [Como fazer uma melhor cobertura sobre inteligência artificial - Columbia Journalism Review](#)
- [Três maneiras fáceis de avaliar alegações sobre IA - IEEE Spectrum](#)
- [Emily M. Bender – Medium](#)
- [Sobre os perigos dos papagaios estocásticos | Anais da Conferência ACM 2021 sobre Equidade, Responsabilidade e Transparência](#)



AI Spotlight Series



**The AI
Accountability
Network**



Pulitzer Center