

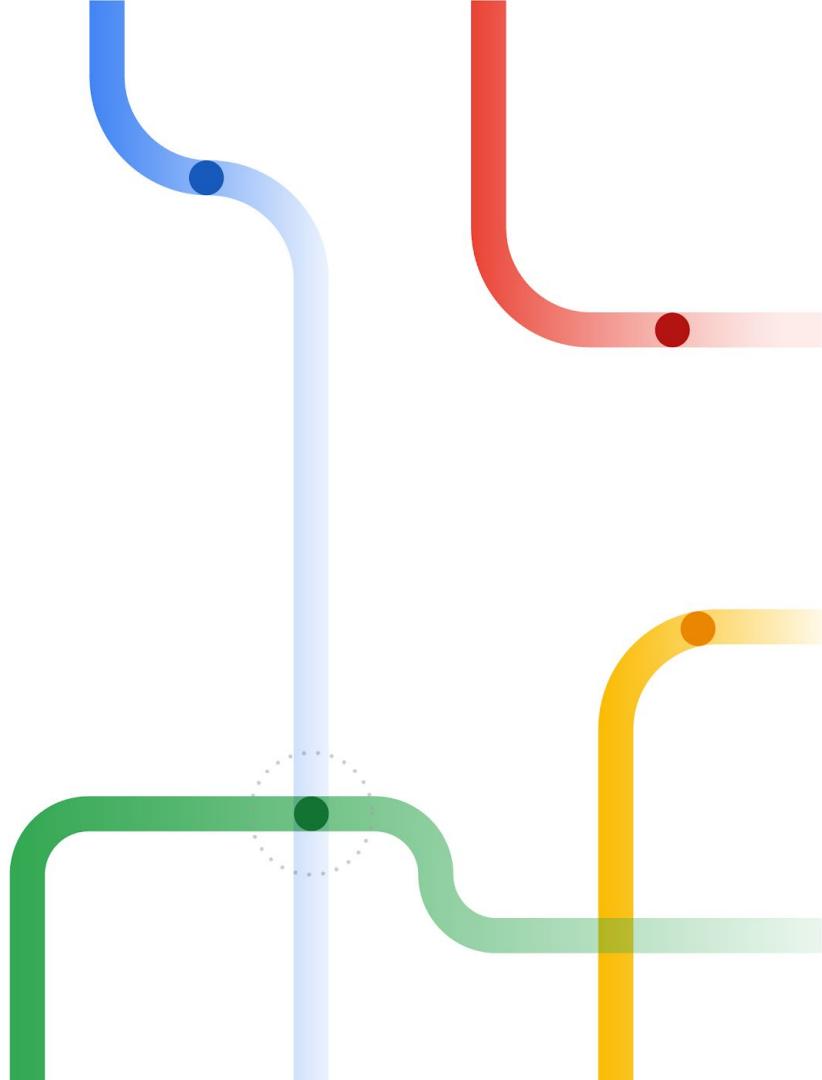
Stay tuned and join data excellence mailing list

[https://bit.ly/data\\_excellence](https://bit.ly/data_excellence)

# Welcome!

Data Excellence:  
Better Data for Better AI

Google Research



**data is the compass for AI** - AI advances where there is data

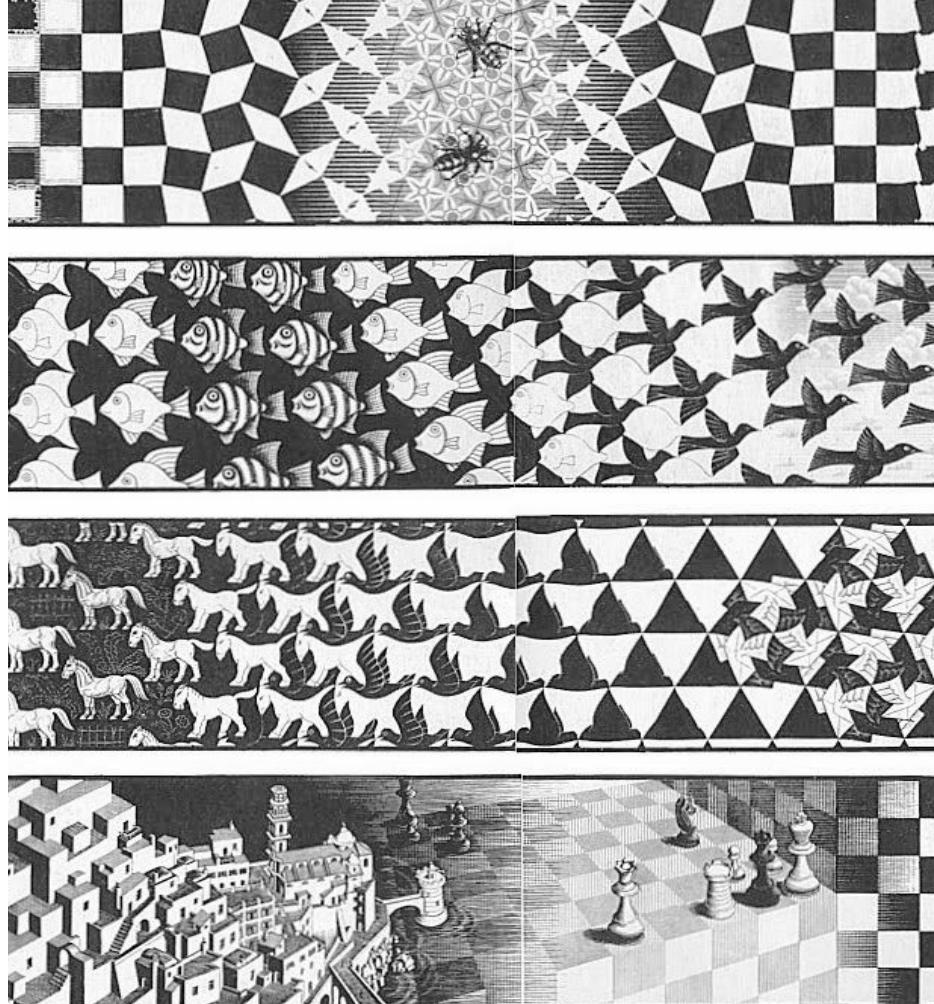
**data quality** must be addressed in AI practices especially in the way we evaluate AI

**improving evaluation of AI** must consider ways to measure variance and capture bias to bring us one step closer to **data excellence**

to address **variance in AI evaluation** we propose a number of novel metrics for reliability, significance (metrology for AI) and disagreement (CrowdTruth)

to address **bias in AI evaluation** we propose a novel method for crowdsourcing adverse test sets for ML models (CATS4ML)

## TAKE HOME MESSAGE



**data is the compass for AI** - AI advances where there is data

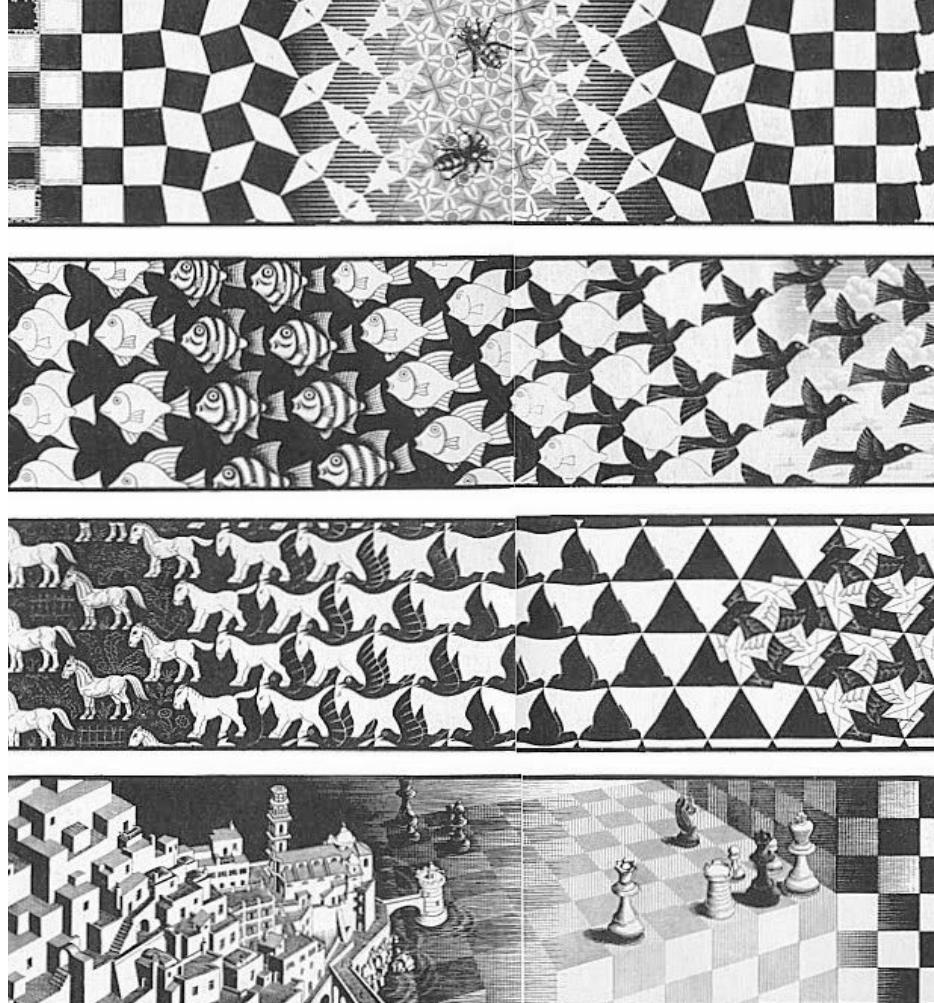
**data quality** must be addressed in AI practices especially in the way we evaluate AI

**improving evaluation of AI** must consider ways to measure variance and capture bias to bring us one step closer to **data excellence**

to address **variance in AI evaluation** we propose a number of novel metrics for reliability, significance (metrology for AI) and disagreement (CrowdTruth)

to address **bias in AI evaluation** we propose a novel method for crowdsourcing adverse test sets for ML models (CATS4ML)

## TAKE HOME MESSAGE



**data is the compass for AI** - AI advances where there is data

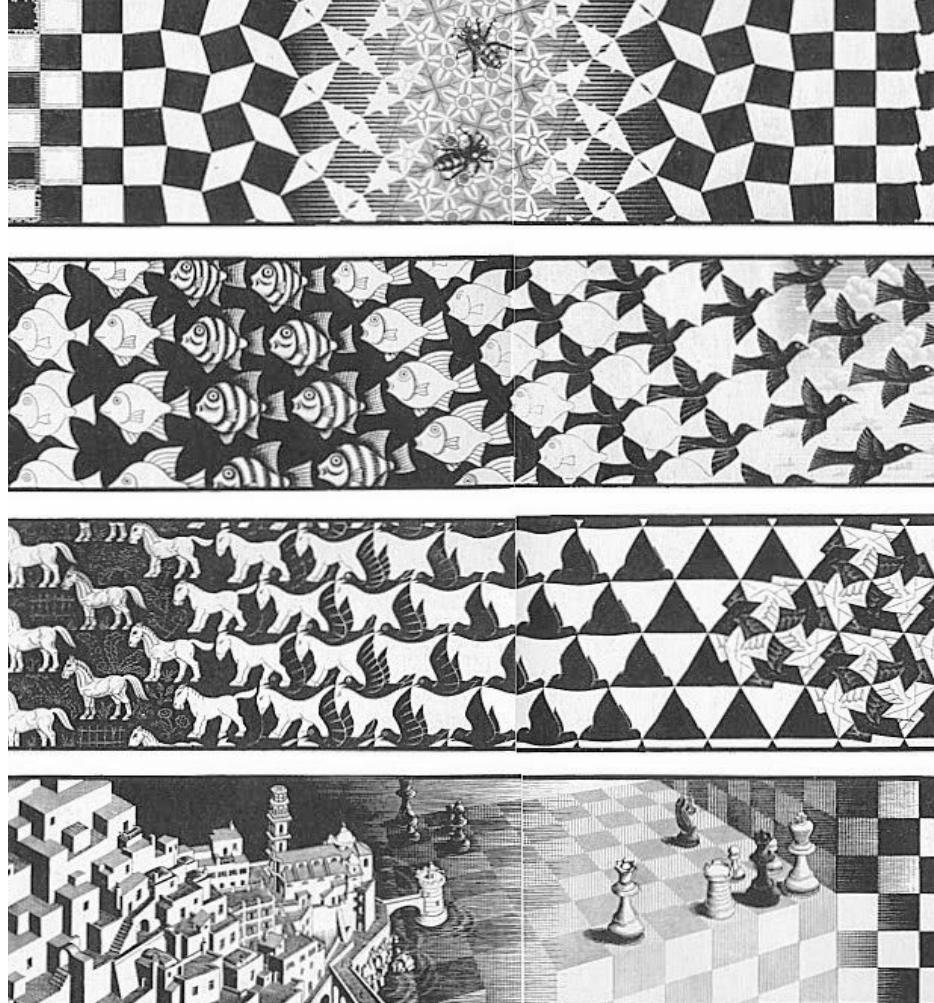
**data quality** must be addressed in AI practices especially in the way we evaluate AI

**improving evaluation of AI** must consider ways to measure variance and capture bias to bring us one step closer to **data excellence**

to address **variance in AI evaluation** we propose a number of novel metrics for reliability, significance (metrology for AI) and disagreement (CrowdTruth)

to address **bias in AI evaluation** we propose a novel method for crowdsourcing adverse test sets for ML models (CATS4ML)

## TAKE HOME MESSAGE



**data is the compass for AI** - AI advances where there is data

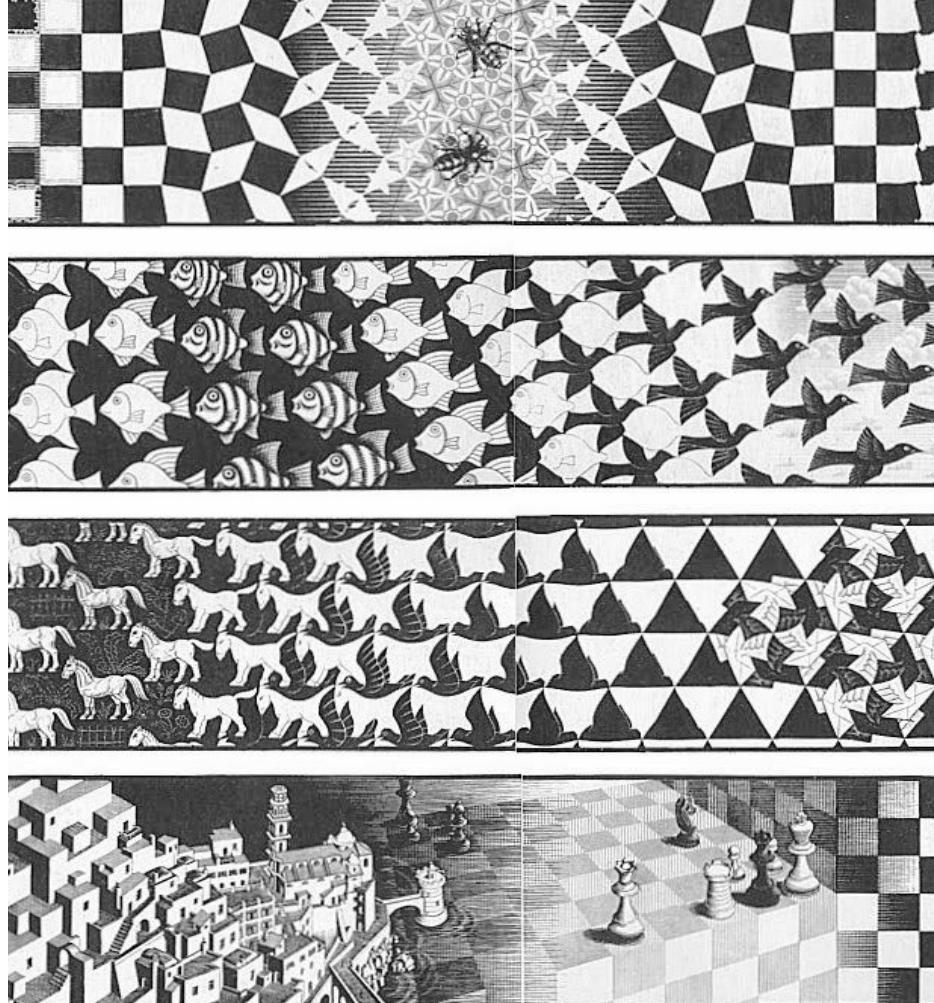
**data quality** must be addressed in AI practices especially in the way we evaluate AI

**improving evaluation of AI** must consider ways to measure variance and capture bias to bring us one step closer to **data excellence**

to address **variance in AI evaluation** we propose a number of novel metrics for reliability, significance (metrology for AI) and disagreement (CrowdTruth)

to address **bias in AI evaluation** we propose a novel method for crowdsourcing adverse test sets for ML models (CATS4ML)

## TAKE HOME MESSAGE



# The Rise of the Machines



*lab experiments*

“AI Winter”

Expert Systems  
small scale  
experiments

# The Rise of the Machines



*beat the humans*

“AI Winter” → “AI Breakthroughs in Games”

IBM Watson Jeopardy  
DeepMind AlphaGo

# The Rise of the Machines



*support the humans*

“AI Winter” → “AI Breakthroughs in Games” → “**Real World Tasks**”

- Health diagnostics
- Flue prediction
- Weather prediction
- Text, Image and Video classification
- Text Generation
- Text Translation
- Conversational AI

# Mainstream Deployment of AI

“Real World Tasks” deployed in the wild → **Unintended behaviors**

- Microsoft Tay bot
- IBM Watson Oncology
- Amazon Rekognition
- Google Photos
- Apple Face ID
- Facebook chat bots
- Various Speech Assistants

# Data is the **compass** for AI

**data quality** is essential for  
guiding AI away from  
unintended behaviours

**getting computers to “see”**  
the diversity of data



# The Life of AI Data



*bootstrapping AI with data*

“It exists!”

Caltech101

LabelMe

Berkley-3D

[https://en.wikipedia.org/wiki/List\\_of\\_datasets\\_for\\_machine-learning\\_research](https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research)

Google Research



# The Life of AI Data



*data hungry AI*

“It exists!” → “It is bigger!”

ImageNet

SIFT10M

OpenImages

COCO

Web 1T 5-Gram

[https://en.wikipedia.org/wiki/List\\_of\\_datasets\\_for\\_machine-learning\\_research](https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research)

# The Life of AI Data



*but before it got better ...*

“It exists!” → “It is bigger!” → “It is better!”

# The Life of AI Data



*but before it got better ...*

“It exists!” → “It is bigger!” → “It is better!”

*it got worse ...*

# Unintended Behaviors in AI

The New York Times

PLAY THE CROSSWORD

## A.I. Shows Promise Assisting Physicians



WIR ED

ANDY GREENBERG SECURITY 11.12.2017 8:44 PM

### Hackers Say They've Broken Face ID a Week After iPhone X Release

"I would say if this is all confirmed, it does mean Face ID is less secure than Touch ID."

f t m

cnet

COVID-19 BEST REVIEWS NEWS HOW TO FINANCE HEALTH

## Facebook put c chatbots that c nt languag

SUBSCRIBE

DAN FAGELLA, TECHEMERGENCE BOIANFAGELLA JUNE 4, 2017 3:10 PM

Bob, the two bots artificial intelligence



Nieve 17 July 31, 2017 11:58 a.m. P

itbots has ie children ate a age.

archers at und two bed in the rk's AI been ing with an way. The Bob and ed a lang

### Tesla's Autopilot keeps crashing parked cars. Here's why.



Self-driving car timeline for 11 top automakers

DAN FAGELLA, TECHEMERGENCE BOIANFAGELLA JUNE 4, 2017 3:10 PM

Bob, the two bots artificial intelligence

Nieve 17 July 31, 2017 11:58 a.m. P

itbots has ie children ate a age.

archers at und two bed in the rk's AI been ing with an way. The Bob and ed a lang

### Tesla's Autopilot keeps crashing parked cars. Here's why.



All in the court: When algorithms rule on jail t

By MATT O'BRIEN and DAKE KANG January 31, 2018

## Machine Bias

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

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

**O**N A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 10-year-old girls were realizing they were too big for the tiny conveniences — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of \$10.

MIT Technology Review

Topics Magazine Newsletters Events

Sign In Subscribe

## Google's medical AI was super accurate in a different story.

Artificial intelligence / Machine learning



Google's medical AI was super accurate in a different story.

April 27, 2020

The Washington Post Democracy Dies in Darkness

Transportation

### Uber's self-driving cars had a major flaw: They weren't programmed to stop for jaywalkers



An Uber driverless car in a garage in San Francisco. (Eric Risberg/AP)

By Hannah Knowles November 6, 2019 at 10:06 p.m. EST

Software detected the woman almost six seconds before Uber's self-driving car struck her, investigators say, in the crash that would lead to her death and prompt the ride-share giant to slam the brakes on its autonomous vehicle testing.

neutral "o" to English (and inspired by this Facebook post).

CHI 2020, April 25–30, 2020, Honolulu, HI, USA

## Evaluation of a Deep Learning System for the Detection of Diabetic Retinopathy

Fred Hersch Google Health Palo Alto, CA fredhersch@google.com

Lauren Wilcox Google Health Palo Alto, CA lwilcox@google.com

Paisan Ruamrongsuk Rajivit Hospital Bangkok, Thailand paisan.trs@gmail.com

Laura M. Vardavakis Google Health Palo Alto, CA lauravar@google.com

### INTRODUCTION

Diabetes is a growing problem around the world, including the United States. As of 2017, approximately 30 million Americans were living with diabetes, comparable to 9.1% of the population in the United States [41, 40]. With diabetes comes an increased risk of diabetic retinopathy, a condition caused by chronically high blood sugar that damages blood vessels in the retina, the thin layer at the back of the eye. These blood vessels can leak or hemorrhage, causing vision distortion or loss. DR is one of the leading causes of blindness worldwide, excluding refractive errors [38]. In Thailand, 34% of patients with diabetes have low vision or blindness in either eye [23].

clinicians work in settings. In this deep learning system, eleven clinics working, privacy and indicate that performance. These blood vessels can leak or hemorrhage, causing vision distortion or loss. DR is one of the leading causes of blindness worldwide, excluding refractive errors [38]. In Thailand, 34% of patients with diabetes have low vision or blindness in either eye [23].

in health care —

Adapted from "AI in the Open World: Discovering Blind Spots of AI", SafeAI 2020, Ece Kumar

Google Research



# The Life of AI Data



*but before it got better ...*

“It exists!” → “It is bigger!” → “It is better!”  
*reactive  
data improvement*

**A.I. Shows Promise Assisting Physicians**

Doctors compared against A.I. computers to recognize illnesses on magnetic resonance images found that the machines were better at reading brain scans. The human doctors did well, though. (Mark Lichtenstein/Associated Press)

By Cade Metz

Feb. 21, 2019

Leave an email

Each year, millions of Americans walk out of a doctor's office with a misdiagnosis. Physicians try to be systematic when identifying symptoms and diseases, but bias creeps in.

Now a group of researchers in the United States and China has tested a potential remedy for all-too-human frailties: artificial intelligence.

In a paper published on Monday in *Nature Medicine*, the researchers reported that they had built a system that automatically diagnoses common childhood conditions — from influenza to meningitis — after processing the patient's symptoms, history, lab results and other clinical data.

**Hackers Say They've Broken Face ID a Week After iPhone X Release**

"I would say if this is all confirmed, it does mean Face ID is less secure than Touch ID."

This article has been updated below with another, more convincing video demonstration of Blair's Face ID-cracking, which the firm revealed two weeks after the original.

When Apple released the iPhone X on November 3, it pushed off an immediate race among hackers around the world to be the first to fool the company's futuristic new form of authentication. A week later, hackers on the actual other side of the world claim to have successfully duplicated someone's face to unlock his iPhone X — with what looks like a simpler technique than some security researchers believed possible.

**Facebook put cork in chatbots that created a secret language**

Alice and Bob, the two bots, raise questions about the future of artificial intelligence.

By Michael Nease

Jan. 21, 2017 8:00 AM PT

Leave an email

A pair of chatbots has recently done something most people don't do: create a secret language.

Last month, researchers at Facebook's artificial intelligence division had been communicating with each other in a completely unexpected way. The two bots, Alice and Bob, had generated a language all on their own.

**Self-driving car timeline for 11 top automakers**

Facebook's gender bias pairs "he" with "hardworking" and "she" with lazy, and other examples

Uber's self-driving cars had a major flaw. They weren't programmed to stop for jaywalkers

**AI in the court: When algorithms rule on jail time**

**Machine Bias**

Google Translate's medical AI was super accurate in a lab. Real life was a different story.

It is really going to matter, if it happens to patients we need to know how it works when real humans get their hands on it, in real situations.

by Will Douglas Heaven

April 27, 2020

Leave an email

In the Turkish language, there is one pronoun, "o," that covers every kind of singular third person. Whether it's a he, a she, or an it, you can't tell just by looking at the word. If you're translating a sentence from Turkish to English, it just has to guess whether "o" means he, she, or it. And those translations reveal the algorithm's gender bias.

Here is a poem written by Google Translate on the topic of gender-neutral "o" to English (and inspired by this Facebook post):

**Uber drivers sue in a garage in San Francisco**

**A Human-Centred Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy**

**Abstract**

Recent advances in machine learning have led to breakthroughs in many fields, including health care. In particular, deep learning has shown great promise for the detection of diabetic retinopathy (DR), a leading cause of blindness worldwide. However, the deployment of such systems in clinical settings has raised concerns about the impact on health care providers and the public perception. We conducted a human-centered evaluation of a deep learning system for DR detection, involving health care providers and patients. Our findings suggest that the system is effective in detecting DR, but it also has limitations, such as difficulty in interpreting the results and potential bias. We believe that further research is needed to address these challenges and ensure that AI systems are used ethically and effectively in health care.

**Google Translate's gender bias pairs "he" with "hardworking" and "she" with lazy, and other examples**

Uber's self-driving cars had a major flaw. They weren't programmed to stop for jaywalkers

**AI in the court: When algorithms rule on jail time**

**Machine Bias**

**Google Translate's medical AI was super accurate in a lab. Real life was a different story.**

It is really going to matter, if it happens to patients we need to know how it works when real humans get their hands on it, in real situations.

by Will Douglas Heaven

April 27, 2020

Leave an email

In the Turkish language, there is one pronoun, "o," that covers every kind of singular third person. Whether it's a he, a she, or an it, you can't tell just by looking at the word. If you're translating a sentence from Turkish to English, it just has to guess whether "o" means he, she, or it. And those translations reveal the algorithm's gender bias.

Here is a poem written by Google Translate on the topic of gender-neutral "o" to English (and inspired by this Facebook post):

**Abstract**

Recent advances in machine learning have led to breakthroughs in many fields, including health care. In particular, deep learning has shown great promise for the detection of diabetic retinopathy (DR), a leading cause of blindness worldwide. However, the deployment of such systems in clinical settings has raised concerns about the impact on health care providers and the public perception. We conducted a human-centered evaluation of a deep learning system for DR detection, involving health care providers and patients. Our findings suggest that the system is effective in detecting DR, but it also has limitations, such as difficulty in interpreting the results and potential bias. We believe that further research is needed to address these challenges and ensure that AI systems are used ethically and effectively in health care.

# The Life of AI Data



*to reach here*

“It exists!” → “It is bigger!” → “It is better!”

*we need proactive  
data improvement*

# The Life of AI Data



## EXPERT OPINION

Contact Editor: **Brian Brannon**, bbrannon@computer.org

## The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, Google

Eugene Wigner's article "The Unreasonable Effectiveness of Mathematics in the Natural Sciences"<sup>1</sup> examines why so much of physics can be neatly explained with simple mathematical formulas such as  $f = ma$  or  $e = mc^2$ . Meanwhile, sciences that

behavior. So, this corpus could serve as the basis of a complete model for certain tasks—if only we knew how to extract the model from the data.

### Learning from Text at Web Scale

The biggest successes in natural-language-related machine learning have been statistical speech rec-

In the decade since then, the research community have done a lot with quantity, but quality has been left behind

Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. *The Unreasonable Effectiveness of Data*. IEEE Intelligent Systems 24, 2 (2009)

# But ...Data Quality is not easy ...

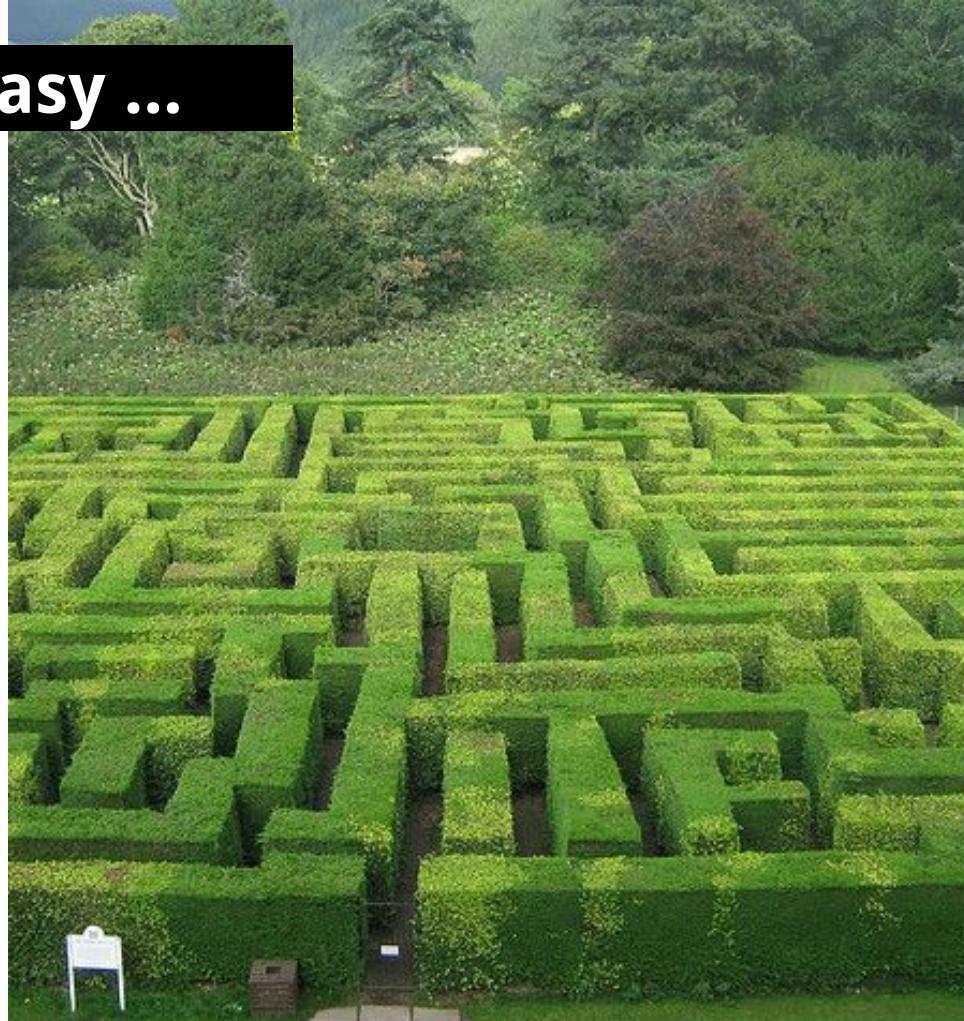
datasets are **not easy to debug**

**low data quality** is typically **not a result of**:

- software bugs, or
- just human errors

**achieving excellent data** requires change in the way we gather data and use it to evaluate AI systems:

- ensure that the **datasets represent the problems they are intended for**
- ensure that the **human annotation process** for these datasets allows to **capture the natural variance & bias** in the problem
- ensure that the **metrics** used to measure the AI quality **utilize the variance & bias** in the human responses



# Data Quality **is not only** human error

Do these images depict a **GUITAR** ?

it is not easy to give Y/N answer  
for most of our AI tasks



# Data Quality **should consider context of use**

Do these images depict **NEW ZEALAND** ?

it is not easy to give Y/N answer  
for most of our AI tasks

the answer typically **depends on**  
**the context, on the task, on the**  
**usage, etc**



# Data Quality **should include** real world diversity

Do these images depict a **WEDDING** ?

it is not easy to give Y/N answer  
for most of our AI tasks

the answer typically **depends on**  
**the context, on the task, on the**  
**usage, etc**

**disagreement is signal for**  
**natural diversity and variance**  
**in human annotations** and  
should be included in AI training



# Data Quality is difficult even with experts

Does the sentence express **TREATS** relation between Chloroquine, Malaria?

Rheumatoid arthritis and **MALARIA** have been treated  
with **CHLOROQUINE** for decades.

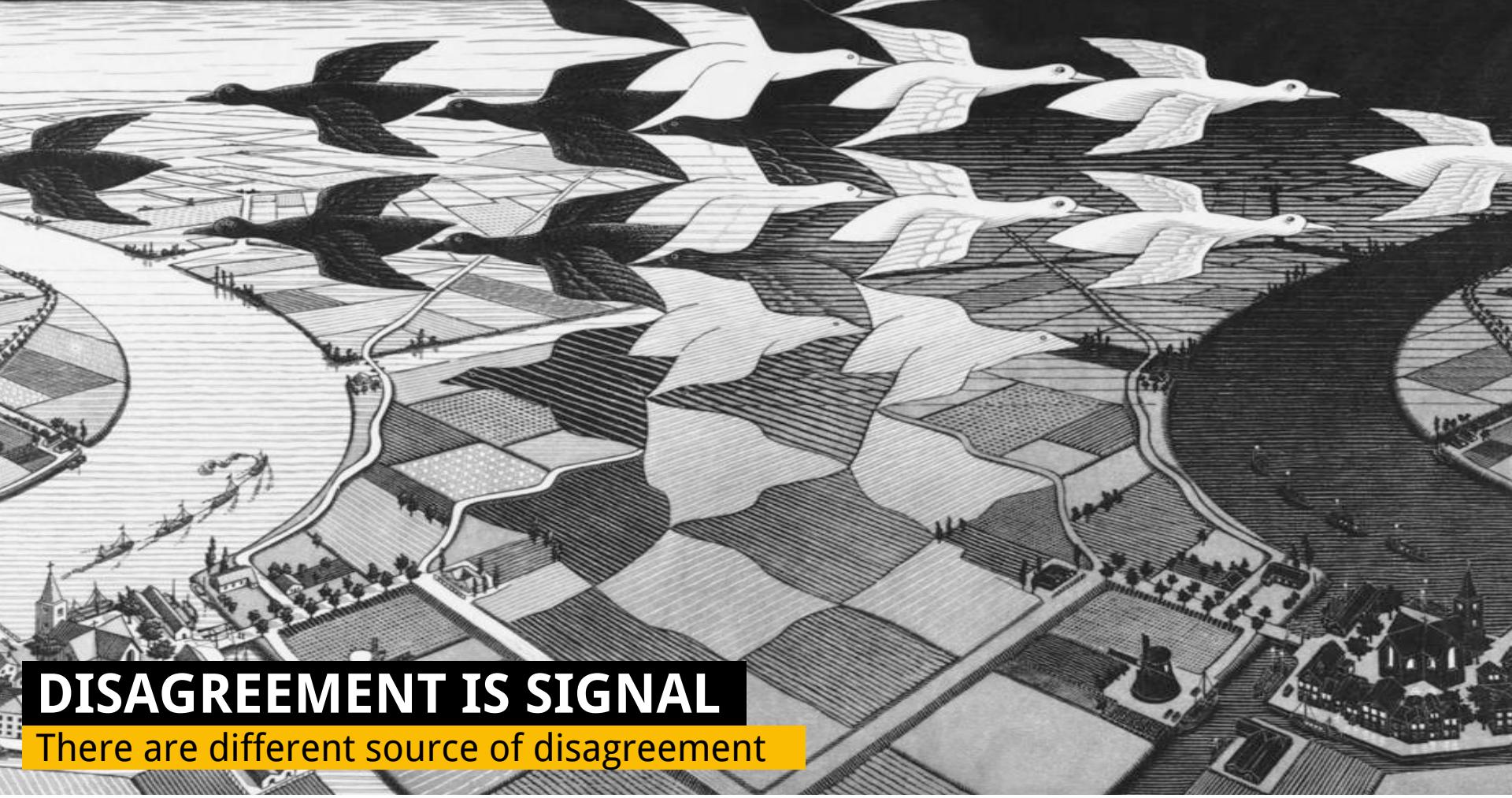


For prevention of malaria, use only in individuals traveling to malarious areas where **CHLOROQUINE** resistant P. falciparum **MALARIA** has not been reported.



Among 56 subjects reporting to a clinic with symptoms of **MALARIA** 53 (95%) had ordinarily effective levels of **CHLOROQUINE** in blood.



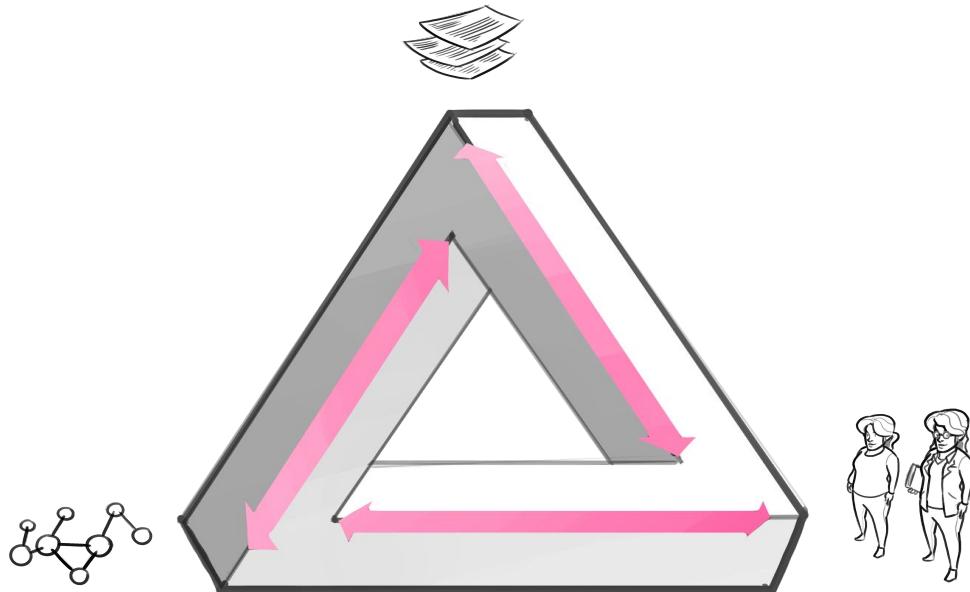


## DISAGREEMENT IS SIGNAL

There are different source of disagreement

# CROWDTRUTH

## Three sides of human interpretation



Disagreement provides  
guidance in task analysis:

- items with **poor semantics**
- items with **salient terms**
- items **difficult to classify**
- items that are **ambiguous**

- users **with/without specific knowledge**
- communities of thought
- spammers

- **subjective** annotations
- **time-sensitive** annotations
- **difficult** annotation tasks
- **mis-translated** annotations

*"Three Sides of CrowdTruth", Human Computation Journal, 2014, L. Aroyo, C. Welty*

# ... but in current practices disagreement is continuously avoided and the world is forced into a binary setting

## 7 Myths about Human Annotation

**One truth:** knowledge acquisition typically assumes one correct interpretation for every example

**Experts rule:** knowledge is captured from domain experts

**One is enough:** single expert's knowledge is sufficient

**Disagreement bad:** when people disagree, they must not understand the problem

**Detailed explanations help:** if examples cause disagreement - adding instructions should help

**Once done, forever valid:** knowledge is not updated; new data not aligned with old

**All examples are created equal:** triples are triples, one is not more important than another, they are all either true or false



"Truth is a Lie: 7 Myths about Human Annotation", AI Magazine 2014, L. Aroyo, C. Welty

**data is the compass for AI** - AI advances where there is data

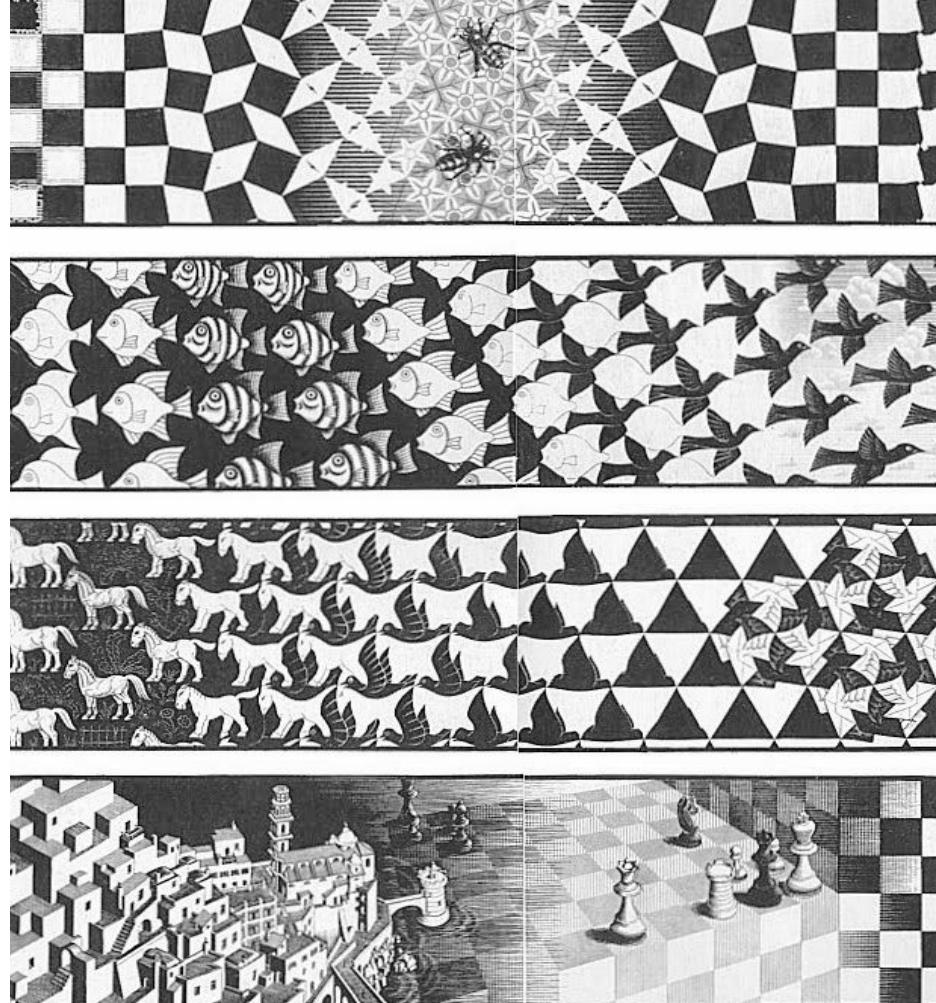
**data quality** must be addressed in AI practices especially in the way we evaluate AI

**improving evaluation of AI** must consider ways to measure variance and capture bias to bring us one step closer to **data excellence**

to address **variance in AI evaluation** we propose a number of novel metrics for reliability, significance (metrology for AI) and disagreement (CrowdTruth)

to address **bias in AI evaluation** we propose a novel method for crowdsourcing adverse test sets for ML models (CATS4ML)

## TAKE HOME MESSAGE



Your AI model is as good as your **evaluation data** ... but is your evaluation data **missing relevant examples?** **How can we find them if they are AI blindspots (i.e. unknown unknowns)?**

Join us tomorrow Dec 9 at 7:30-8:00pm (EST) to hear more about the **CATS4ML challenge (Crowdsourcing Adverse Test Sets for ML):**

[cats4ml.humancomputation.com](https://cats4ml.humancomputation.com)



Stay tuned and join data excellence mailing list

[https://bit.ly/data\\_excellence](https://bit.ly/data_excellence)



Praveen Paritosh



Ka Wong



Lora Aroyo



Devi Krishna

## The Team

## Share your voice about ML data!

We are inviting ML professionals to participate in a short survey to learn about your challenges and needs with data.

Scan to participate now!



# Build for everyone

Fill out our interest form to hear  
about events and opportunities  
at [goo.gle/neurips-booth-form](https://goo.gl/neurips-booth-form)

Google at NeurIPS 2020

