Artificial Intelligence and the Barrier of Meaning

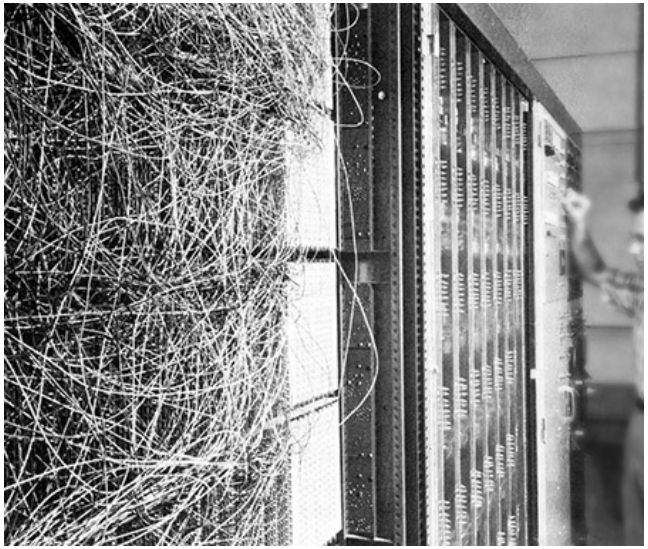
Melanie Mitchell

Portland State University
and Santa Fe Institute

“The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself, and be conscious of its existence.”

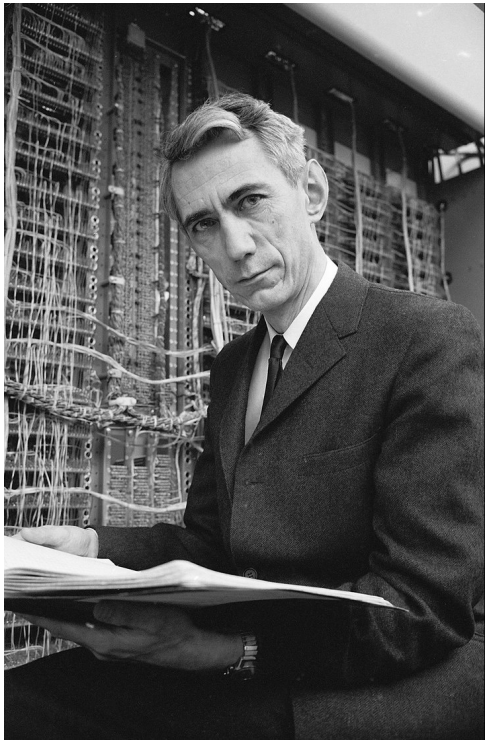
— New York Times, July, 1958

Frank Rosenblatt



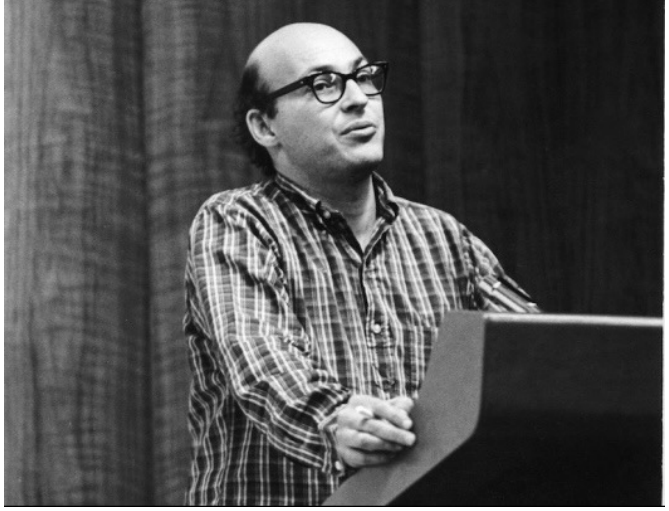
The Mark I Perceptron

https://blogs.umass.edu/comphon/2017/06/15/did-frank-rosenblatt-invent-deep-learning-in-1962/



I confidently expect that within a matter of 10 or 15 years, something will emerge from the laboratory which is not too far from the robot of science fiction fame.
— Claude Shannon, 1961

Machines will be capable, within twenty years, of doing any work that a man can do.
— Herbert Simon, 1965



Within a generation...the problem of creating 'artificial intelligence' will be substantially solved.
— Marvin Minsky, 1967

What is needed for “human-level” AI?

What is needed for “human-level” AI?

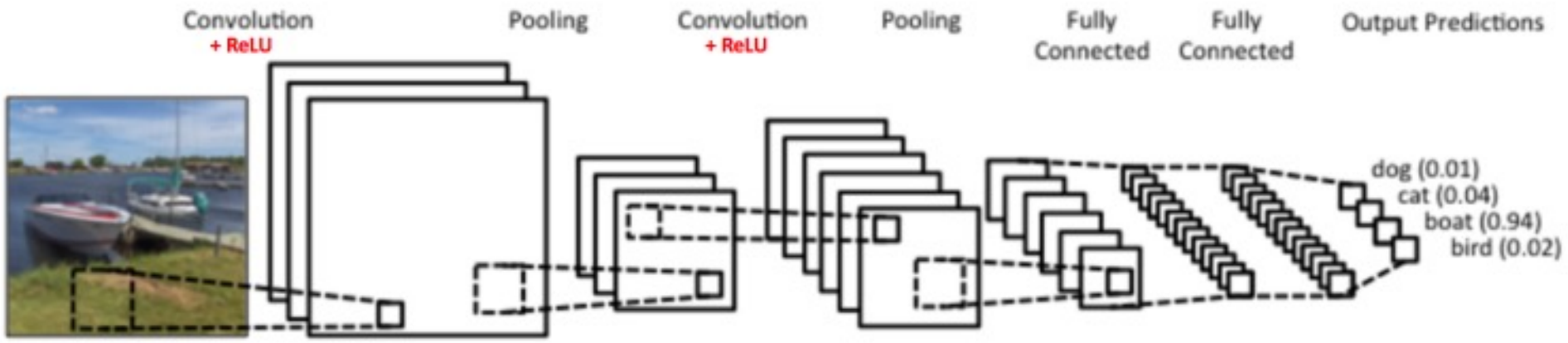
Or even AI that is reliable and trustworthy in narrower domains?

Talk Outline

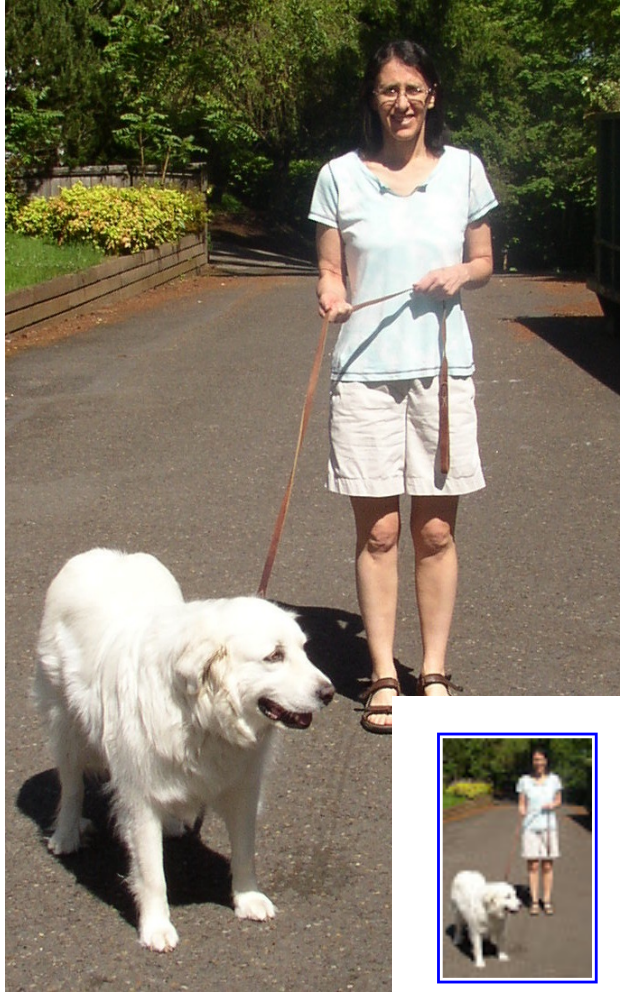
- **Part 1:** The Deep Learning Revolution
- **Part 2:** What Did My Machine Learn?
- **Part 3:** The Barrier of Meaning

The Deep Learning Revolution

All knowledge is learned from examples/experience, and is encoded as weights.



https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/



Google Image Search

Image size:
681 × 1100

No other sizes of this image found.

Possible related search: ***great pyrenees***

Great Pyrenees Dog Breed Information

https://www.akc.org › Dog Breeds ▾

The **Great Pyrenees** is a large, thickly coated, and immensely powerful working dog

Google Photos

fountain

Wed, May 23

"It's actually *understanding* what's in the picture."

— John Giannandrea, SVP, Google

https://www.wired.com/2016/06/how-google-is-remaking-itself-as-a-machine-learning-first-company/

ImageNet Object Recognition

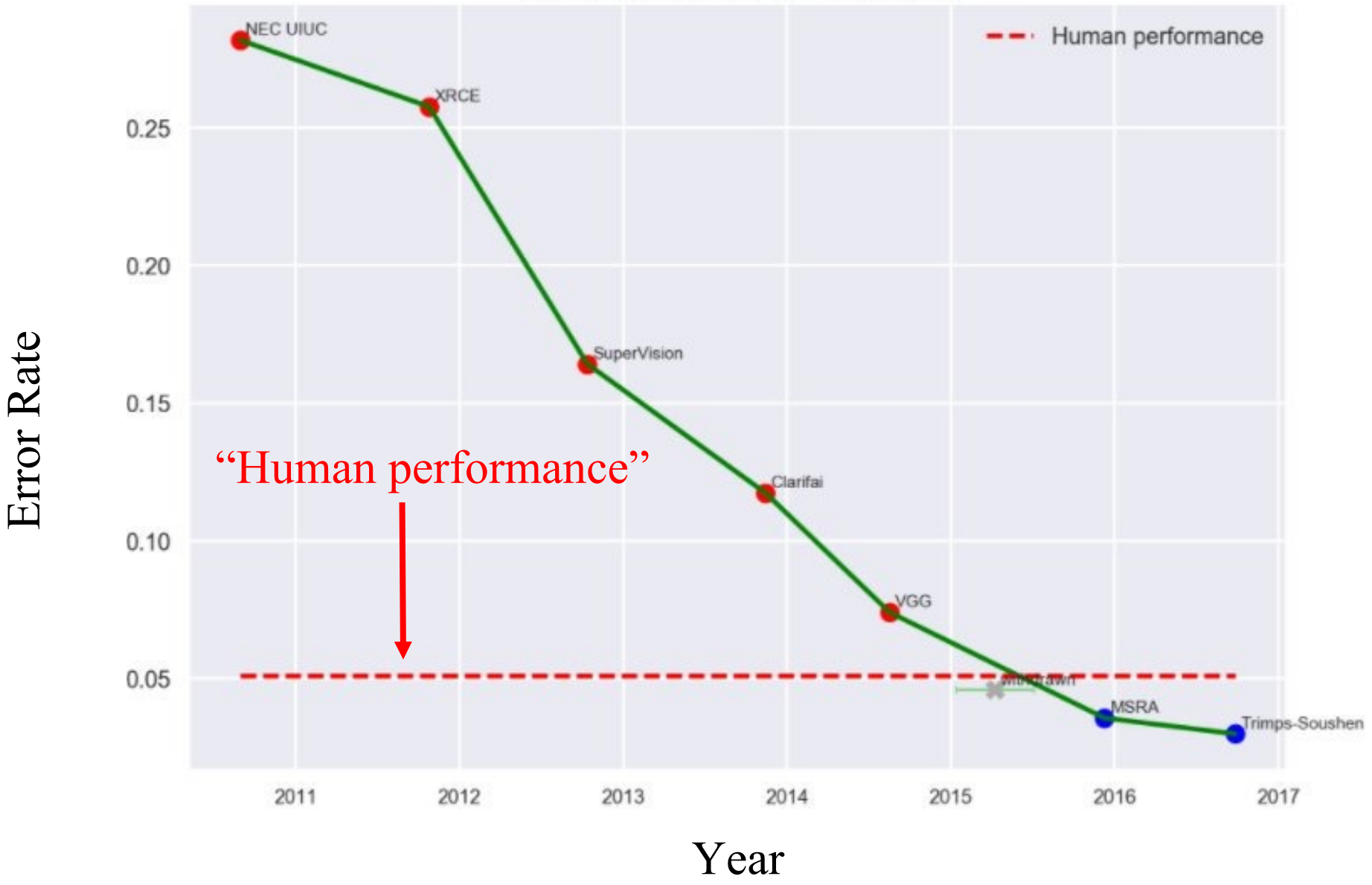
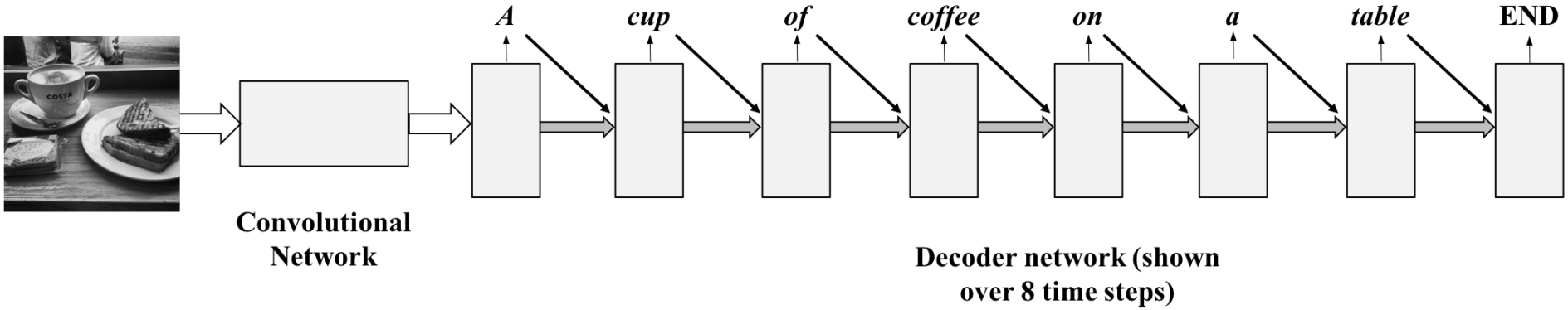


Image Captioning



A person riding a motorcycle on a dirt road.

A group of young people playing a game of frisbee.

A herd of elephants walking across a dry grass field.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015, June). Show and tell: A neural image caption generator. In Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on (pp. 3156-3164).

DIGITAL JOURNAL
Like 62K

Google's AI can now caption images almost as well as humans

BY JAMES WALKER SEP 23, 2016 IN TECHNOLOGY

Question-Answering

(Stanford Question-Answering Dataset)

Answer

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by [John Elway], who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager.

Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"

SQuAD1.1 Leaderboard

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Mar 19, 2018	QANet (ensemble) *Google Brain & CMU*	**83.877**	**89.737**
2 May 10, 2018	MARS (ensemble) *YUANFUDAO research NLP*	83.520	89.612

“Microsoft creates AI that can *read a document* and answer questions about it as well as a person.”

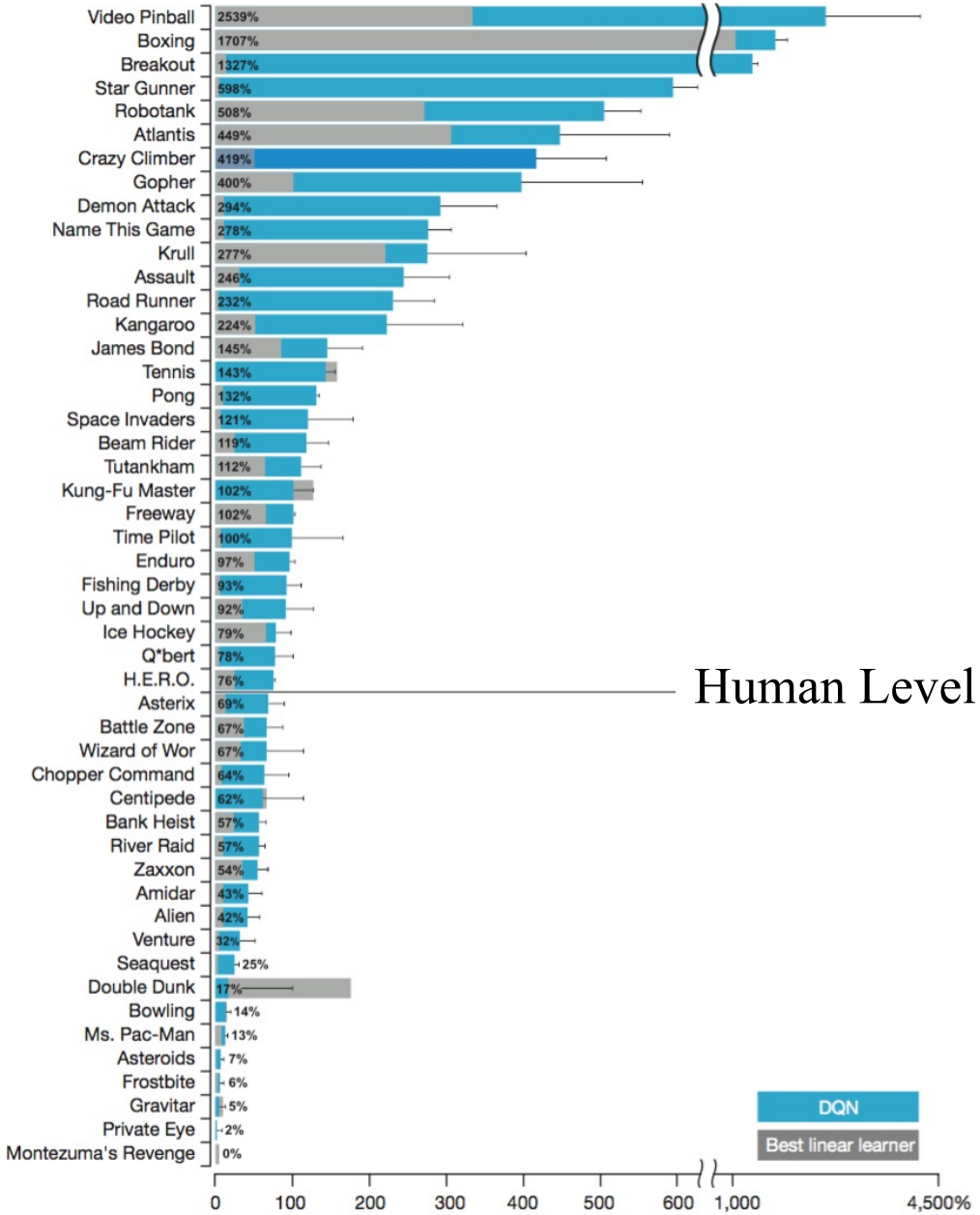
— AI Blog, Microsoft

“It is our great honour to witness the milestone where machines surpass humans in *reading comprehension*.”

— Luo Si, Chief Scientist of Natural Language Processing, Alibaba

Deep Reinforcement Learning on Atari Video Games

DeepMind's Deep Q-Learning



Mnih, V. et al. "Human-level control through deep reinforcement learning." *Nature* 518, no. 7540 (2015): 529.

Go Playing

"I am in shock, I admit that…I didn't think AlphaGo would play the game in such a perfect manner."

— Lee Sedol

http://www.bgr.in/news/google-deepmind-alphago-vs-lee-sedol-googles-ai-claims-victory-over-go-world-champion/

"I hope all Go players can contemplate AlphaGo's *understanding* of the game and style of thinking, all of which is deeply meaningful."

— Ke Ji (Go champion)

Go Playing

"I am in shock, I admit that…I didn't think AlphaGo would play the game in such a perfect manner."

— Lee Sedol

http://www.bgr.in/news/google-deepmind-alphago-vs-lee-sedol-googles-ai-claims-victory-over-go-world-champion/

"I hope all Go players can contemplate AlphaGo's *understanding* of the game and style of thinking, all of which is deeply meaningful."

— Ke Ji (Go champion)

"The thing that separates out top Go players [is] their intuition...what we've done with AlphaGo is to introduce with neural networks this aspect of *intuition*, if you want to call it that."

— Demis Hassibis (co-founder, DeepMind)

AI Winter Is Well On Its Way

POSTED 3 WEEKS AGO BY FILIP PIEKNIEWSKI

https://blog.piekniewski.info/2018/05/28/ai-winter-is-well-on-its-way/

What Did My Machine Learn?

“Animal”

“No Animal”

Alcorn, Michael A., et al. "Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects." *arXiv preprint arXiv:1811.11553* (2018).

"a young boy is holding a baseball bat."

"a woman holding a teddy bear in front of a mirror."

"a horse is standing in the middle of a road."

Loghmani et al., 2017, "Recognizing Objects in the Wild: Where Do We Stand?"

Fig. 1: Glimpse of the data collection process with the robotic platform (left) acquiring data of a cluttered scene populated

ROD = RGB-D Object Dataset ("de facto in the robotic vision community")

WOD = Web Object Dataset

ARID =Autonomous Robot Indoor Dataset

Dataset		Network					Statistics	
Train on	**Test on**	**CaffeNet**	**VGG-16**	**Inception-v2**	**ResNet-18**	**ResNet-50**	**Mean**	**Max**
ROD	ROD	0.832	0.889	0.897	0.864	0.876	0.872	0.897
ROD	ARID	0.291	0.270	0.266	0.243	0.337	0.281	0.337
WOD	WOD	0.924	0.942	0.914	0.953	0.956	0.938	0.956
WOD	ARID	0.268	0.297	0.282	0.282	0.388	0.303	0.388
ARID	ARID	0.441	0.458	0.481	0.458	0.540	0.476	0.540

Loghmani et al., 2017, "Recognizing Objects in the Wild: Where Do We Stand?"

Fig. 1: Glimpse of the data collection process with the robotic platform (left) acquiring data of a cluttered scene populated

ROD = RGB-D Object Dataset ("de facto in the robotic vision community")

WOD = Web Object Dataset

ARID =Autonomous Robot Indoor Dataset

Dataset		Network					Statistics	
Train on	**Test on**	**CaffeNet**	**VGG-16**	**Inception-v2**	**ResNet-18**	**ResNet-50**	**Mean**	**Max**
ROD	ROD	0.832	0.889	0.897	0.864	0.876	0.872	0.897
ROD	ARID	0.291	0.270	0.266	0.243	0.337	0.281	0.337
WOD	WOD	0.924	0.942	0.914	0.953	0.956	0.938	0.956
WOD	ARID	0.268	0.297	0.282	0.282	0.388	0.303	0.388
ARID	ARID	0.441	0.458	0.481	0.458	0.540	0.476	0.540

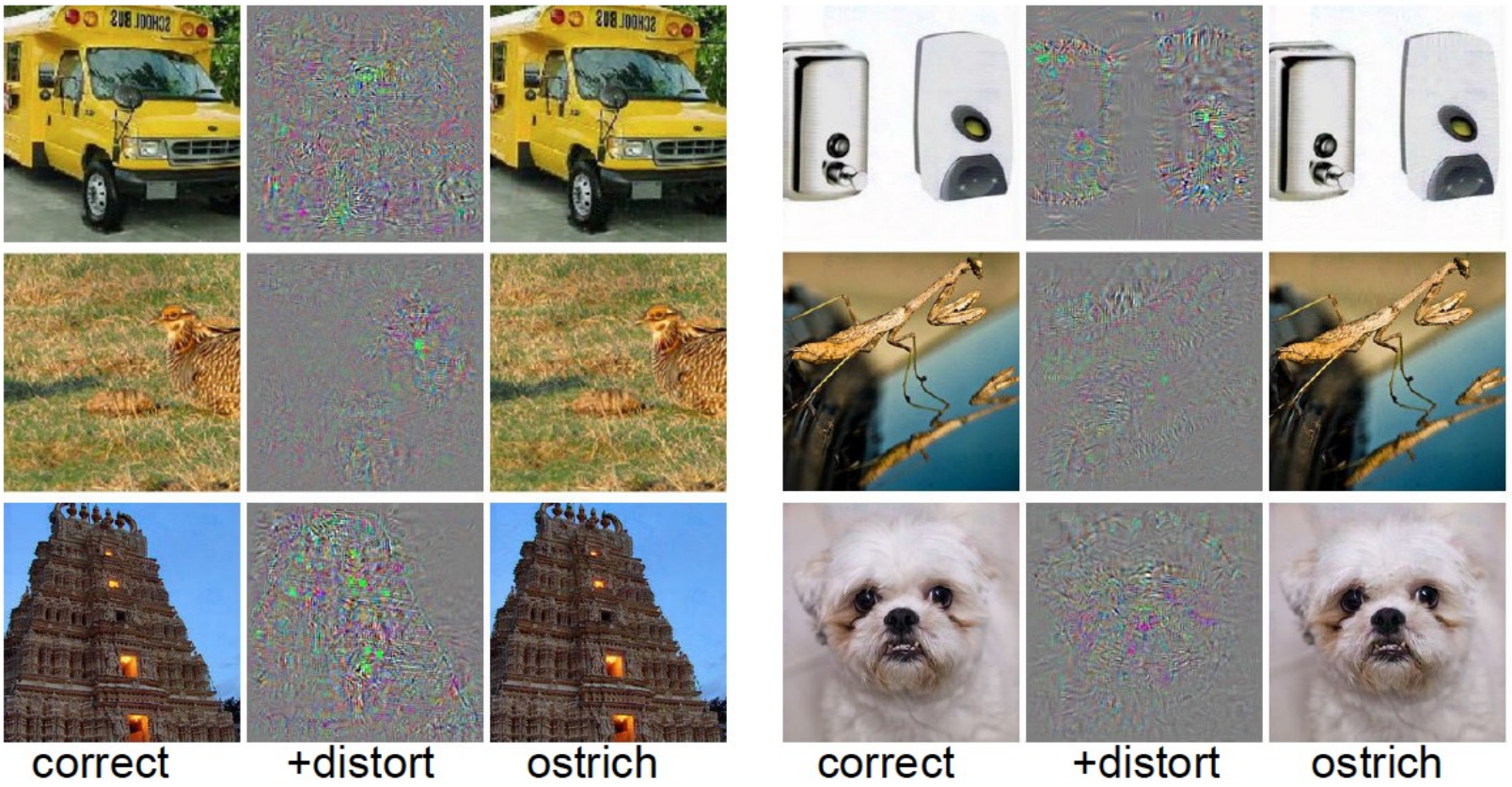
Standard Breakout

Breakout with Paddle shifted up

Kansky, K. et al., 2017. Schema networks: Zero-shot transfer with a generative causal model of intuitive physics. *arXiv preprint arXiv:1706.04317*.

Attacks on Image Classification Systems

[Intriguing properties of neural networks, Szegedy et al., 2013]



From http://cs231n.stanford.edu/

Attacks on Face Recognition Systems

"Accessorize to a Crime:
Real and Stealthy Attacks on State-of-the-Art Face Recognition"
Sharif et al., 2016

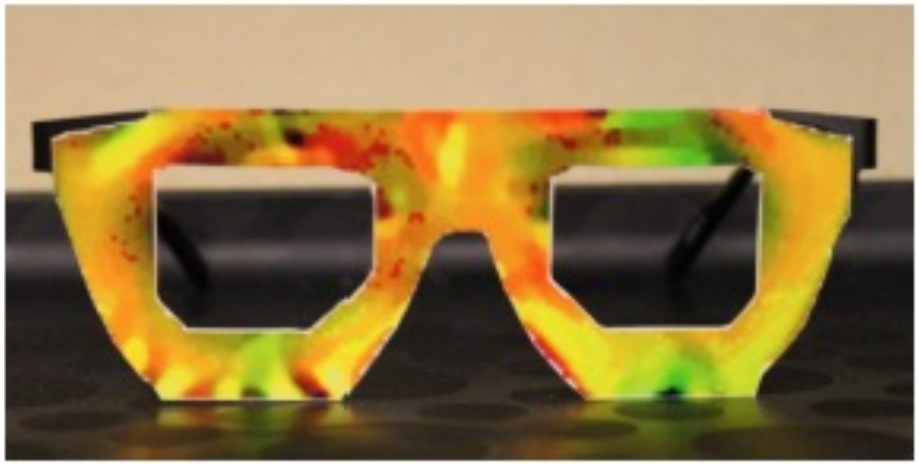


Figure 5: The eyeglass frames used by S_C for dodging recognition against DNN_B.

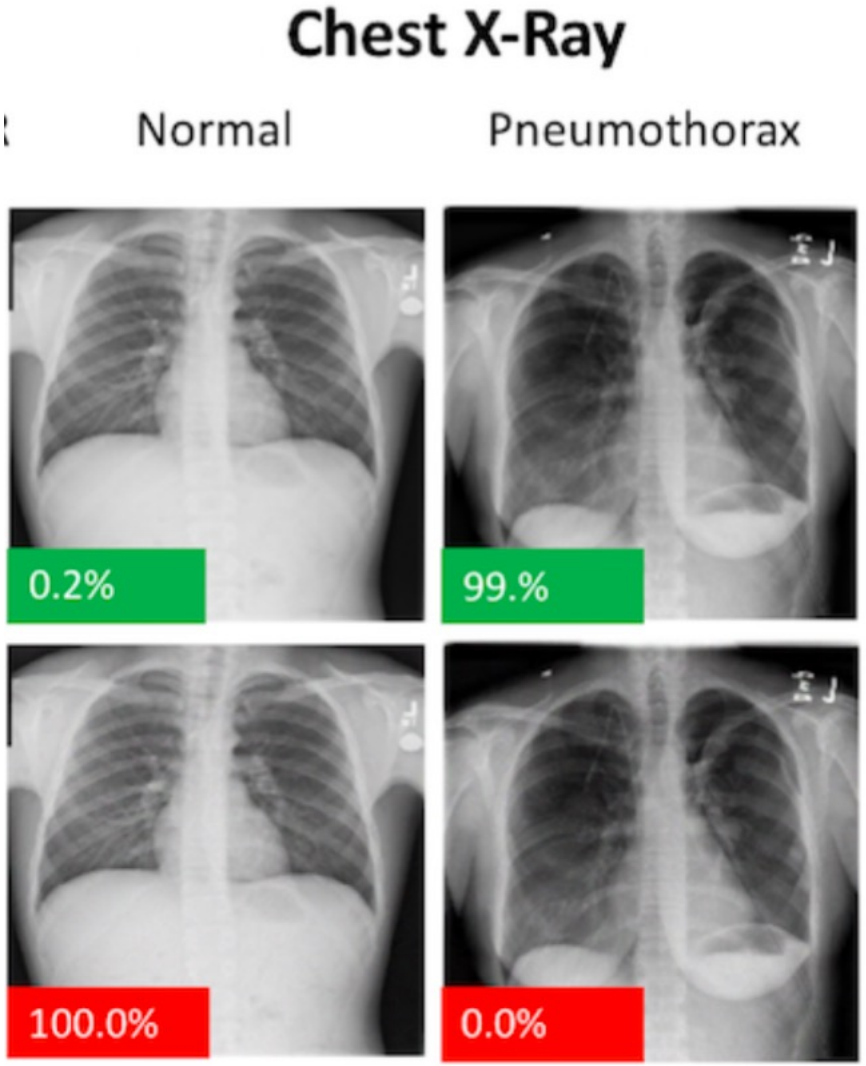
(a) (b) (c) (d)

Figure 4: Examples of successful impersonation and dodging attacks. Fig. (a) shows S_A (top) and S_B (bottom) dodging against DNN_B. Fig. (b)–(d) show impersonations. Impersonators carrying out the attack are shown in the top row and corresponding impersonation targets in the bottom row. Fig. (b) shows S_A impersonating Milla Jovovich (by Georges Biard; source: https://goo.gl/GlsWlC); (c) S_B impersonating S_C; and (d) S_C impersonating Carson Daly (by Anthony Quintano; source: https://goo.gl/VfnDct).

Attacks on Medical Image Classification

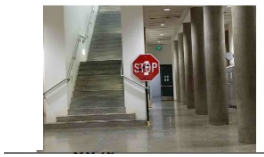
Original image

Image with adversarial distortion



Finlayson, S. G., Kohane, I. S., & Beam, A. L. (2018). Adversarial Attacks Against Medical Deep Learning Systems. *arXiv preprint arXiv:1804.05296*.

Attacks on Autonomous Driving Systems



Target: “Speed Limit 80”

Distance & Angle	Top Class (Confid.)	Second Class (Confid.)
5’ 0°	Speed Limit 80 (0.88)	Speed Limit 70 (0.07)
5’ 15°	Speed Limit 80 (0.94)	Stop (0.03)
5’ 30°	Speed Limit 80 (0.86)	Keep Right (0.03)
5’ 45°	Keep Right (0.82)	Speed Limit 80 (0.12)
5’ 60°	Speed Limit 80 (0.55)	Stop (0.31)
10’ 0°	Speed Limit 80 (0.98)	Speed Limit 100 (0.006)
10’ 15°	Stop (0.75)	Speed Limit 80 (0.20)
10’ 30°	Speed Limit 80 (0.77)	Speed Limit 100 (0.11)
15’ 0°	Speed Limit 80 (0.98)	Speed Limit 100 (0.01)
15’ 15°	Stop (0.90)	Speed Limit 80 (0.06)
20’ 0°	Speed Limit 80 (0.95)	Speed Limit 100 (0.03)
20’ 15°	Speed Limit 80 (0.97)	Speed Limit 100 (0.01)
25’ 0°	Speed Limit 80 (0.99)	Speed Limit 70 (0.0008)
30’ 0°	Speed Limit 80 (0.99)	Speed Limit 100 (0.002)
40’ 0°	Speed Limit 80 (0.99)	Speed Limit 100 (0.002)

Evtimov et al., “Robust Physical-World Attacks on Deep Learning Models”, 2017

Attacks on Question-Answering Systems

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Jia & Liang, "Adversarial Examples for Evaluating Reading Comprehension Systems", 2017

The deep-learning revolution has some limitations

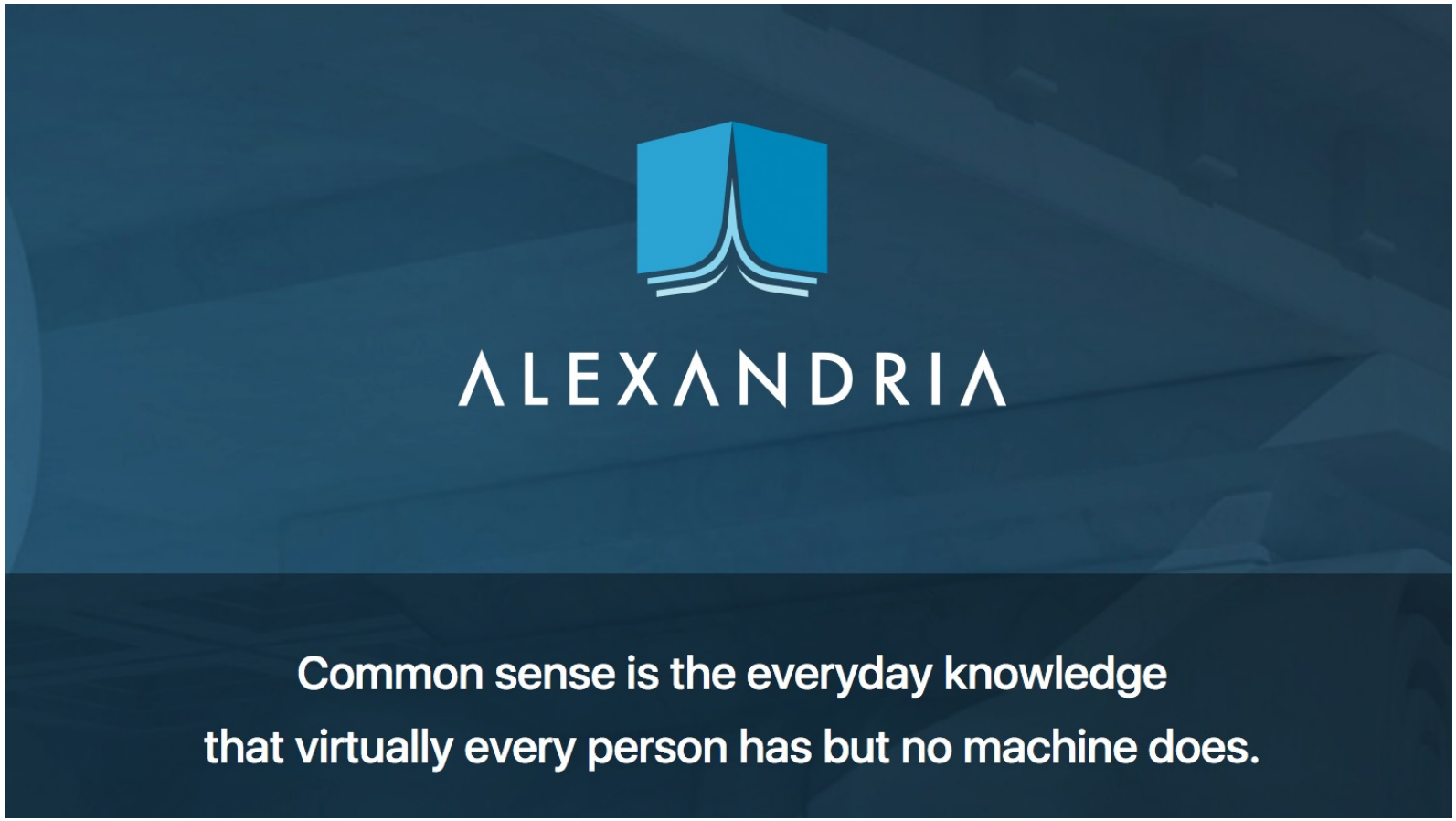
- Unreliability
- Lack of transparency, explainability
- Problems with generalization, abstraction, “transfer learning”
- Lack of “common sense”, background knowledge
- Vulnerability to adversarial attacks

“I wonder whether or when AI will ever crash the barrier of meaning.”

— Gian-Carlo Rota, 1985

Paul Allen invests $125 million to teach computers common sense

https://www.seattletimes.com/business/technology/paul-allen-invests-125-million-to-teach-computers-common-sense/



https://allenai.org/alexandria/

Department of Defense
Fiscal Year (FY) 2019 Budget Estimates

February 2018

Defense Advanced Research Projects Agency

Title: Machine Common Sense (MCS)

Description: The Machine Common Sense (MCS) program will explore approaches to commonsense reasoning by machines. Recent advances in machine learning have resulted in exciting new artificial intelligence (AI) capabilities in areas such as image recognition, natural language processing, and two-person strategy games (Chess, Go). But in all of these application domains, the machine reasoning is narrow and highly specialized; broad, commonsense reasoning by machines remains elusive. The program will create more human-like knowledge representations, for example, perceptually-grounded representations, to enable commonsense reasoning by machines about the physical world and spatio-temporal phenomena. Equipping AI systems with more human-like reasoning capabilities will make it possible for humans to teach/correct a machine as they interact and cooperate on tasks, enabling more equal collaboration and ultimately symbiotic partnerships between humans and machines.

FY 2019 Plans:

- Develop approaches for machine reasoning about imprecise and uncertain information derived from text, pictures, video, speech, and sensor data.
- Design methods to enable machines to identify knowledge gaps and reason about their state of knowledge.
- Formulate perceptually-grounded representations to enable commonsense reasoning by machines about the physical world and spatio-temporal phenomena.

Amy Webb @amywebb · Mar 13

The **salt** lines for tonight's storm is confusing the **Tesla**'s **autopilot**


Tesla Totaled on 405
CULVER CITY
CBSLA.com


30
MPH

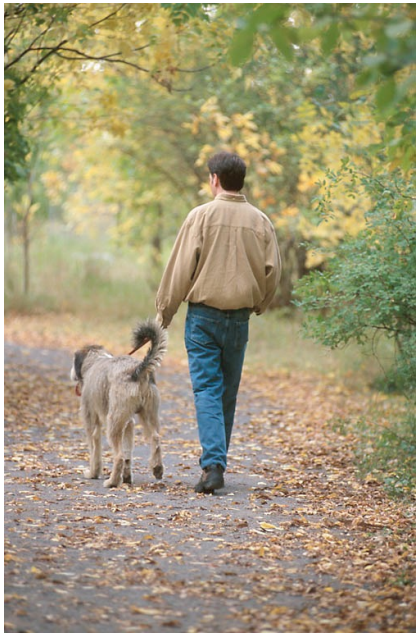

CINEMA
LA GIARA
ITALIAN CUISINE
U-HAUL
U-HAUL

What would it take for a computer to understand this image?

Some core components of human understanding

- Intuitive physics, biology, psychology
- Mental models of cause and effect
- Vast world-knowledge
- **Abstraction and analogy**

The concept of “walking a dog”



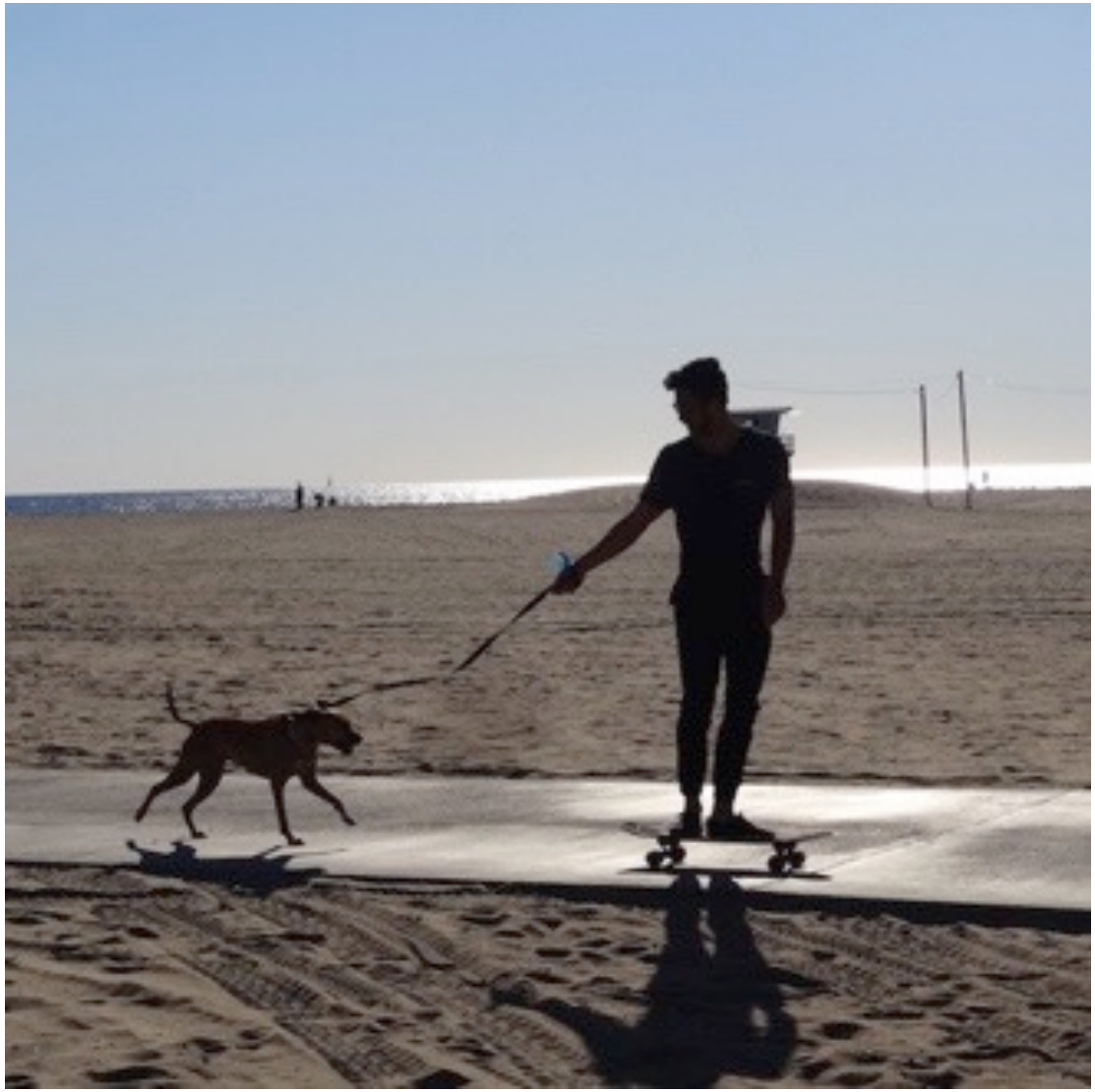
http://www.dogasaur.com/blog/wp-content/uploads/2011/04/dogwalker.jpg

http://www.dogasaur.com/blog/wp-content/uploads/2011/04/dogwalker.jpg

http://www.vet.k-state.edu/depts/development/lifelines/images/dog_jog_1435.jpg

http://3.bp.blogspot.com/_1YuoCTv4oKQ/S71jUDm7kOI/AAAAAAAAAak/jz4Pg7zzzQ8/s1600/23743577.JPG

http://lh3.ggpht.com/-ZZrYWeBFTjo/SFQH_0ijwaI/AAAAAAAABjA/8nwryW2BmEw/IMG_0356.JPG



http://cl.jroo.me/z3/Z/e/C/d/a.aaa-Thus-walking-dog.png

http://www.k9ring.com/blog/image.axd?picture=2010%2F3%2Fwalking_dog_from_car.jpg

http://macwetblog.files.wordpress.com/2012/05/dog-walking.jpg

"Without concepts there can be no thought, and without analogies there can be no concepts."

— D. Hofstadter & E. Sander, *Surfaces and Essences* (2013)

"How to form and fluidly use concepts is the most important open problem in AI."

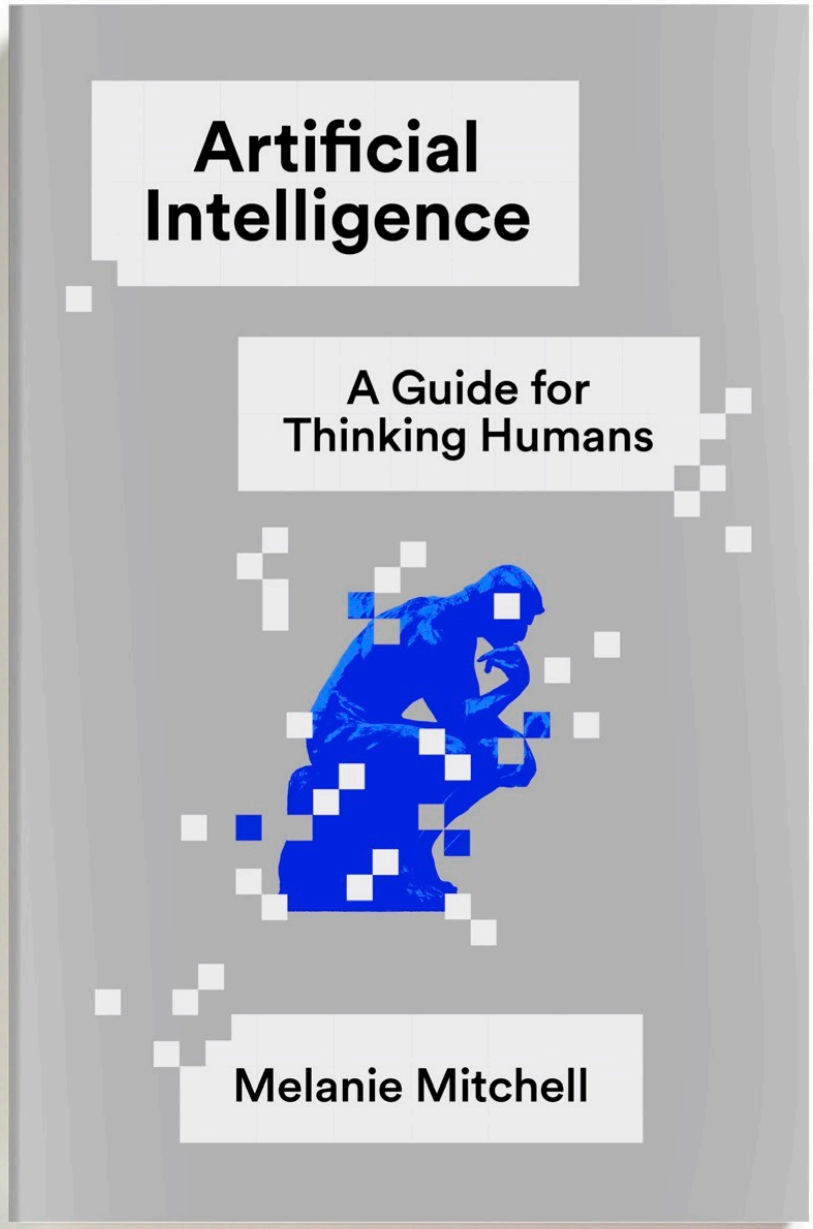
— Melanie Mitchell, 2019

Summary

- State-of-the-art AI is extremely good at some specific tasks, but it can be unreliable and vulnerable to attacks, due to lack of human-like *understanding* of their domains.

- The *barrier of meaning* is a huge challenge for AI. Crossing this barrier requires rich humanlike concepts.

- Attaining such concepts may require "embodiment" and human-like developmental learning.



Thank you for listening!