Taking a closer look at the I in AI

Erik Barzagar-Nazari
Kassel Data Science Meetup
18 February 2020

Meet Clever Hans

https://upload.wikimedia.org/wikipedia/commons/5/57/Osten_und_Hans.jpg

Das Talent des klugen Hans blieb so lange verborgen, weil man beim Pferd suchte, was man beim Menschen hätte finden können.

The talent of Clever Hans remained hidden for so long, because one looked for in the horse what one could have found in man.

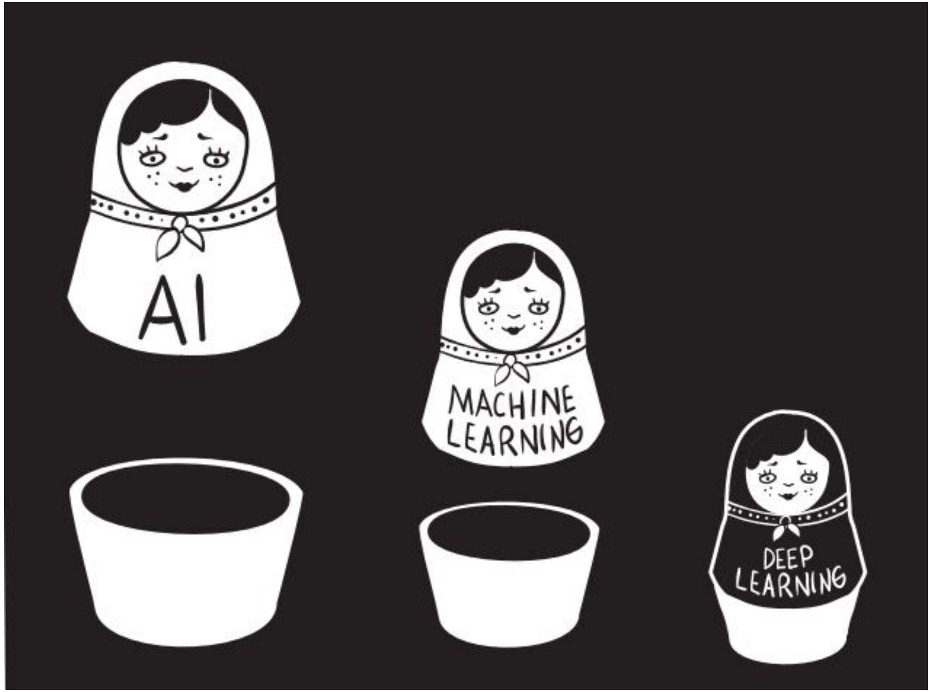
Oskar Pfungst, 1907

Das Talent der Künstlichen Intelligenz blieb so lange verborgen, weil man beim Neuronalen Netz suchte, was man beim Menschen hätte finden können.

The talent of Artificial Intelligence remained hidden for so long, because one looked for in the Neural Network what one could have found in man.

Some Data Scientist, 2020

Artificial Intelligence



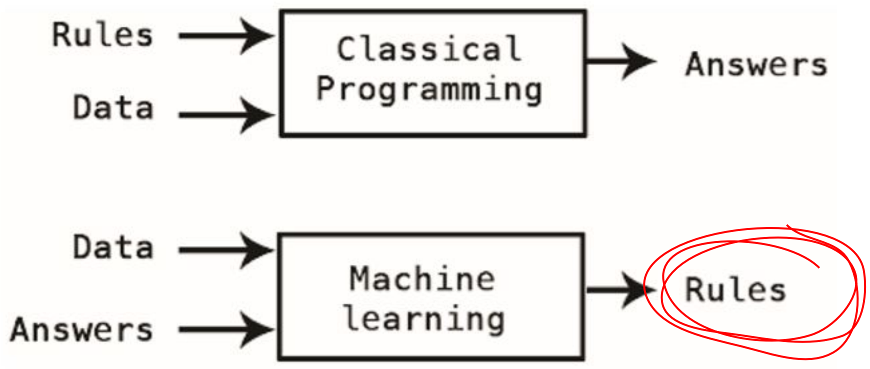
Screenshot from https://weneedtotalk.ai/

- Dominance of Deep Learning
- Overhyped? Probably not.
- Biological terminology and analogies...
- ...lead to some misconceptions

Self-Learning Machines

Learn to map **data** to **answers** by looking for statistical patterns

Machine Learning outputs static **rules**, not intelligent agents



Chollet (2017, 2019)

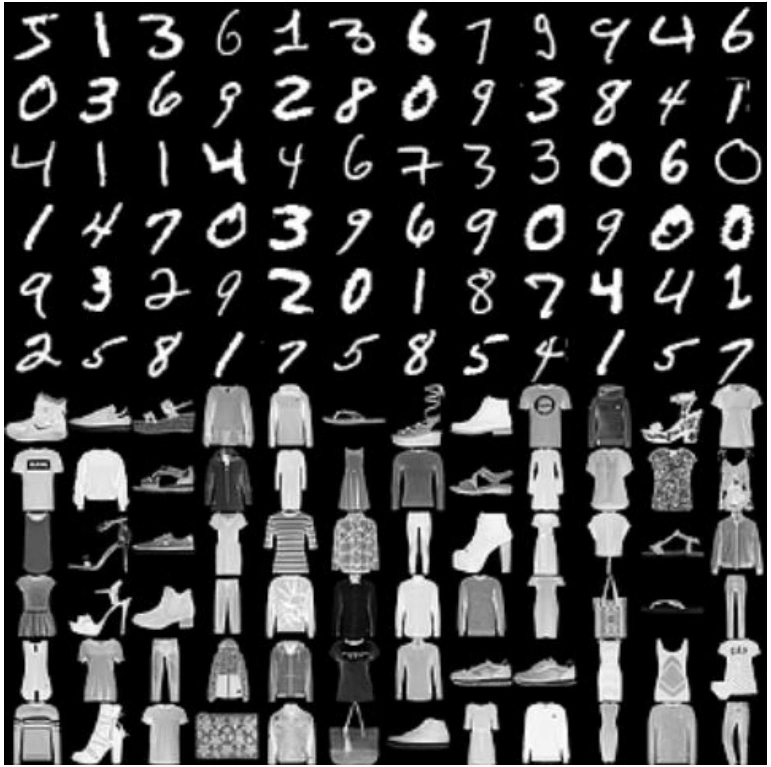
Catastrophic Forgetting

Catastrophic interference, also known as catastrophic forgetting, is the tendency of an artificial neural network to completely and abruptly forget previously learned information upon learning new information.

Wikipedia, 15 February 2020

Catastrophic Forgetting Experiment

1. Define a Convolutional Neural Network
2. Train the network to classify digits using MNIST
3. Train the network to classify fashion categories using Fashion-MNIST



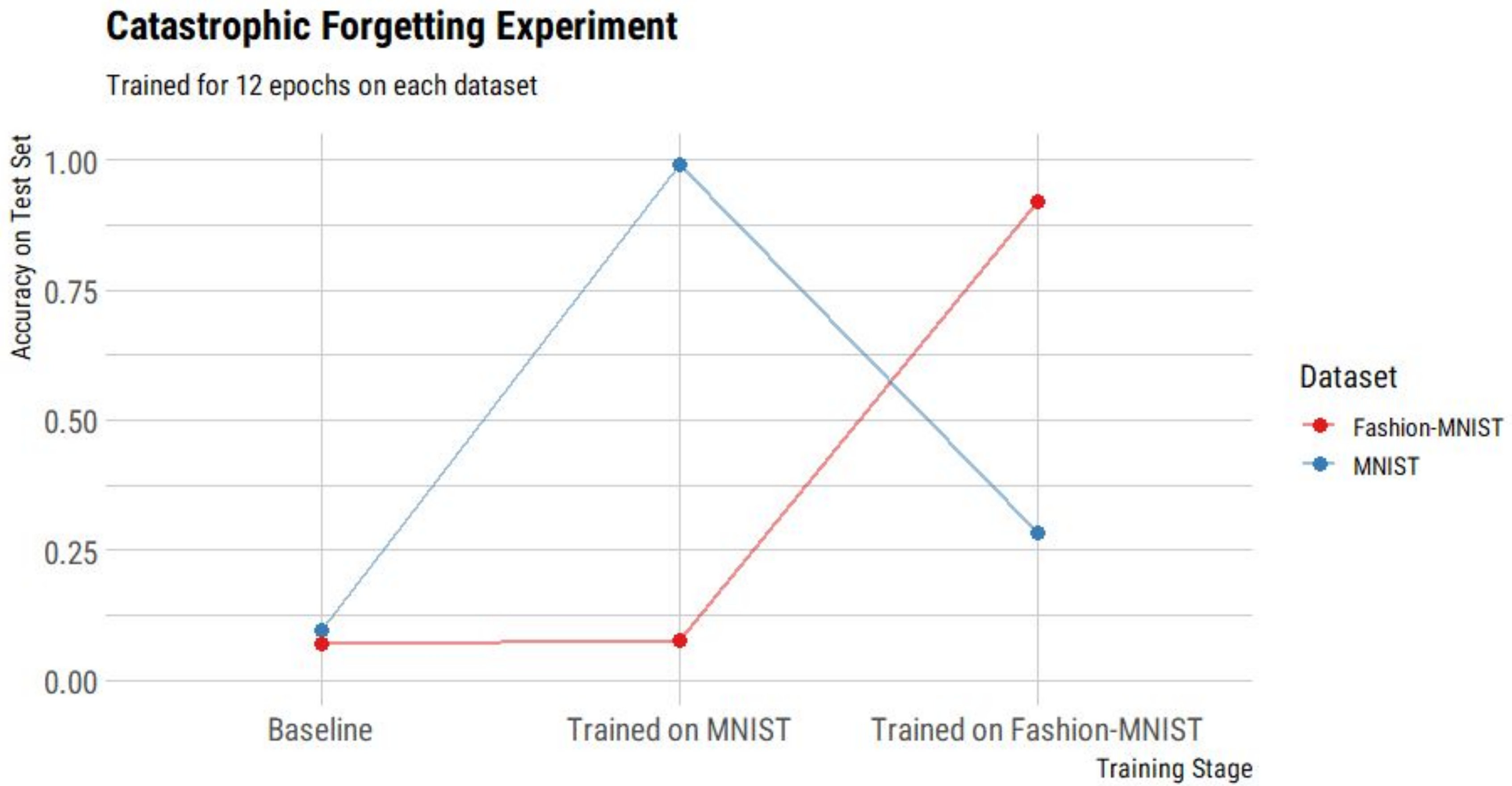
Catastrophic Forgetting: Example

```r
# define convnet
t800 <- keras_model_sequential(name = 'T-800') %>%
  layer_conv_2d(filters = 32, kernel_size = c(3,3), activation = 'relu', input_shape = input_shape) %>%
  layer_conv_2d(filters = 64, kernel_size = c(3,3), activation = 'relu') %>%
  layer_max_pooling_2d(pool_size = c(2, 2)) %>%
  layer_dropout(rate = 0.25) %>%
  layer_flatten() %>%
  layer_dense(units = 128, activation = 'relu') %>%
  layer_dropout(rate = 0.5) %>%
  layer_dense(units = num_classes, activation = 'softmax')

# compile model
t800 %>% compile(
  loss = loss_categorical_crossentropy,
  optimizer = optimizer_adadelta(),
  metrics = 'accuracy'
)
```

Catastrophic Forgetting: Example

```
Model
Model: "T-800"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 26, 26, 32)        320
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 24, 24, 64)        18496
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 12, 12, 64)        0
_________________________________________________________________
dropout (Dropout)            (None, 12, 12, 64)        0
_________________________________________________________________
flatten (Flatten)            (None, 9216)              0
_________________________________________________________________
dense (Dense)                (None, 128)               1179776
_________________________________________________________________
dropout_1 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_1 (Dense)              (None, 10)                1290
=================================================================
Total params: 1,199,882
Trainable params: 1,199,882
Non-trainable params: 0
_________________________________________________________________
```


Catastrophic Forgetting Experiment
Trained for 12 epochs on each dataset
Accuracy on Test Set
1.00
0.75
0.50
0.25
0.00
Baseline
Trained on MNIST
Trained on Fashion-MNIST
Training Stage
Dataset
Fashion-MNIST
MNIST

Catastrophic Forgetting

- Stability-Plasticity-Dilemma
- Active research area
- Current remedy: retrain and redeploy

Taking Shortcuts

Remember Clever Hans?

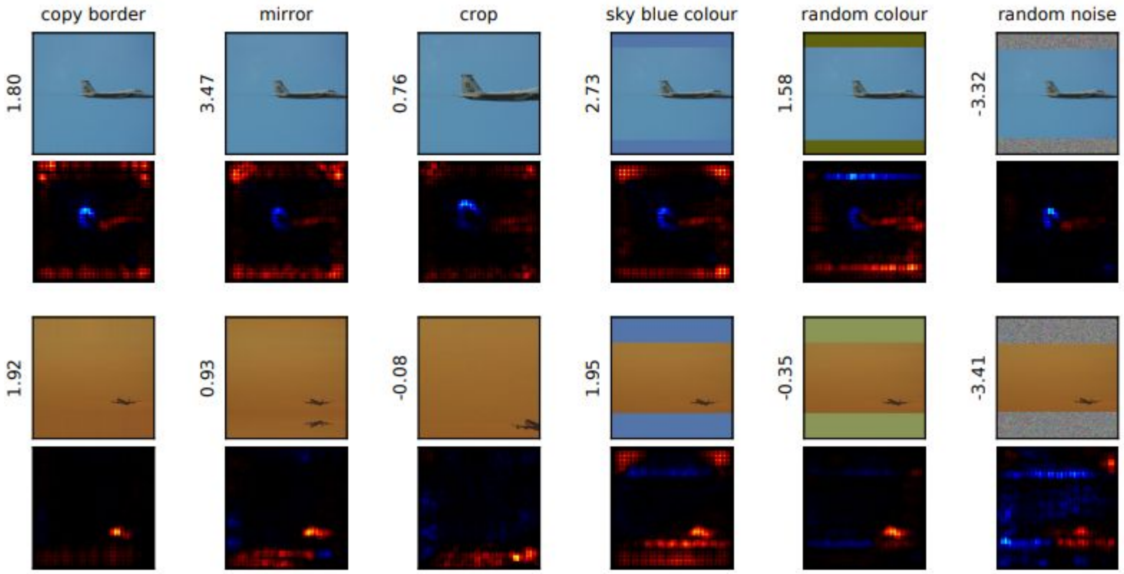
- AI systems do not truly understand what they are doing
- Deep Neural Networks look for a function, that maps Inputs to Outputs
- Sometimes, they take shortcuts
- “Clever Hans Features”

Unusual Suspects

Lapuschkin et al. (2019) discover that even typical image preprocessing steps might introduce Clever Hans Features

Clever Hans Feature: image padding

https://doi.org/10.1038/s41467-019-08987-4



Supplementary Figure 27: Samples from class "aeroplane" and predicted scores for class "aeroplane", with corresponding relevance maps, as affected by different preprocessing strategies to obtain square images. Padding with (high frequency) random noise effectively decreases the predictor output and removes the "border artifact". Using low frequency areas (of the right color) for padding increases the predictor output for class "aeroplane" and may even introduce the "border artifact" in the first place.

Criminals don’t smile

Two researchers claim to be able to predict a person’s “criminality” from its portrait using convnets.

Fortunately, other researchers quickly pointed out this study’s flaws.

Clever Hans Feature: most likely the network learned to distinguish relaxed from tense facial expressions.

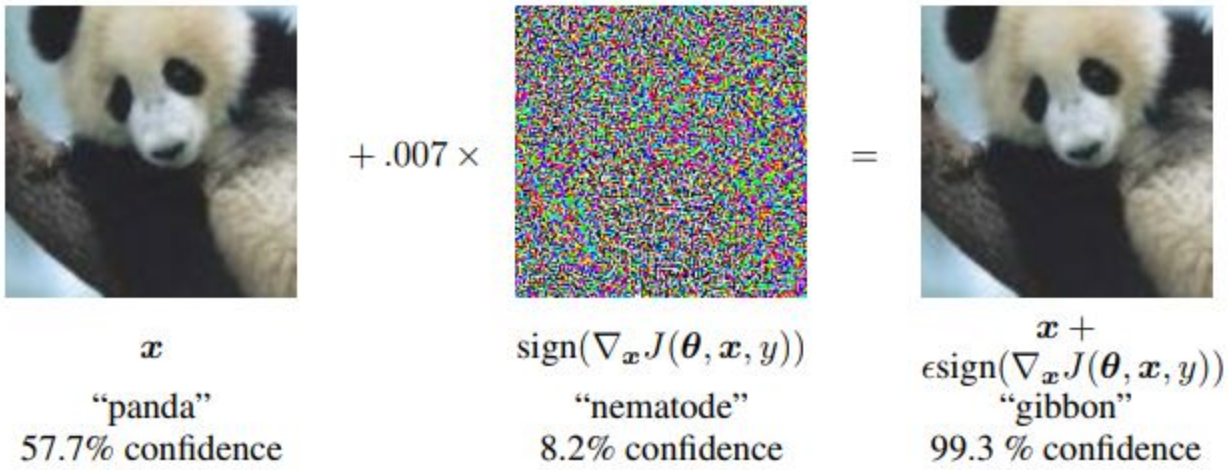
https://arxiv.org/abs/1611.04135

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Adversarial Attacks

Attacking AI Systems



Goodfellow, Shlens & Szegedy (2015)
https://arxiv.org/pdf/1412.6572.pdf

Targeted Attack Example

Class	Label	Score
n02123045	tabby	0.400
n02124075	Egyptian_cat	0.400
n02123159	tiger_cat	0.175
n02127052	lynx	0.007
n04553703	washbasin	0.002

Classification based on ResNet50 with ImageNet weights; Original resized to 224px * 224 px

Targeted Attack Example

```
import foolbox
[...]
attack = LBFGSAttack(model=fmodel, criterion=TargetClassProbability(849, p=.50))
adversarial = attack(image, label, maxiter=500)
```

Index of target class
In ImageNet “849” is “teapot”

Index of true label

Targeted Attack Example

Original (resized)
Tabby (40%)

Adversarial Image
Teapot (97,4%)

Attacking from the Physical World

- Printed fake signs
- Adversarial Patches
- Evaluated on different Architectures
- Evaluated in different scenarios (e.g. drive-by)

Table 1: Sample of physical adversarial examples against LISA-CNN and GTSRB-CNN.

Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5′ 0°					
5′ 15°					
10′ 0°					
10′ 30°					
40′ 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%

Eykholdt et al. (2018)
https://arxiv.org/pdf/1707.08945.pdf

Adversarial Attacks

- White Box Attacks
- Black Box Attacks
- Training with Adversarial Examples improves robustness
- Implications for cybersecurity

Intelligence

Let's ask Open AI's GPT-2

What is intelligent about Artificial Intelligence?

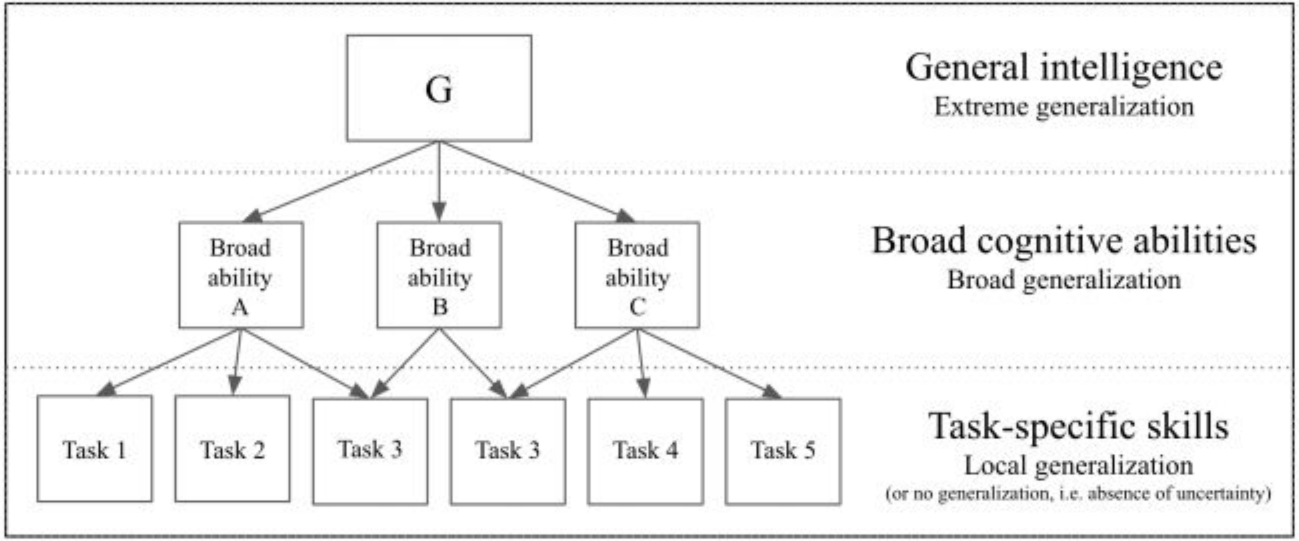
Why can't I describe it without crying?

Well you really can't and I think that's the key. There is a lot of unintelligent and odd behavior that I feel it's very easy to generalise about AI but is really specific to a group of personal behavior.

I think the key to this problem is behavioural universality. Inherently different people react to the same situations in different ways and this leads to confusion. I think people cannot help but general…

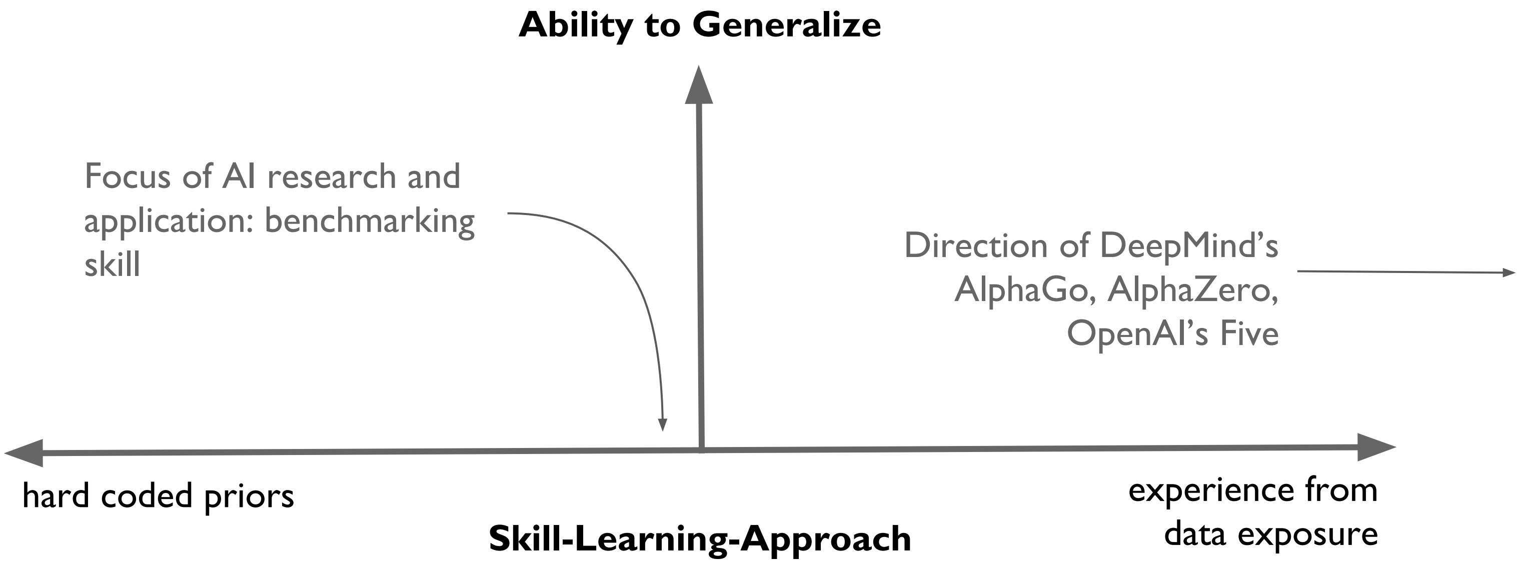
http://talktotransformer.com, accessed on December 2nd, 2019

Hierarchical Model of Intelligence



Chollet (2019)

Skills and Abilities



Chollet (2019)

"The intelligence of a system is a measure of its skill-acquisition efficiency over a scope of tasks, with respect to priors, experience, and generalization difficulty."

Chollet (2019)

Abstraction and Reasoning Corpus (ARC)

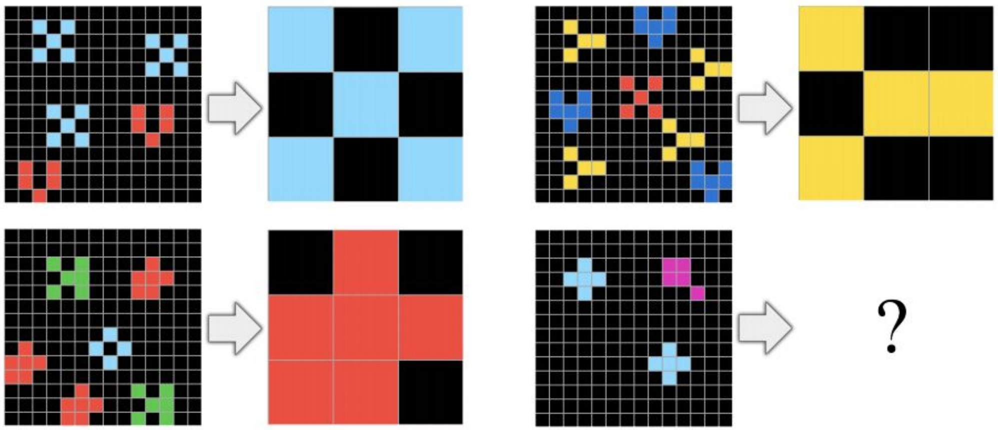


Figure 10: A task where the implicit goal is to count unique objects and select the object that appears the most times (the actual task has more demonstration pairs than these three).

Chollet (2019)

Beyond the Test Error

Beyond the Test Error

- AI performs human tasks, but currently quite differently than humans
- Understand the data generating process
- Is the data representative?
- Add Explainable Machine Learning to your toolbox
- Choose appropriate statistical tools
- Careful formulation of AI use cases

Thank you!

https://upload.wikimedia.org/wikipedia/commons/5/57/Osten_und_Hans.jpg