UNIVERSITY OF CAMBRIDGE

Why Capability-Oriented Evaluation?

Jose Hernandez-Orallo[1,2]

CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

Vancouver, 20 February, 2024

[1] CFI, University of Cambridge,
[2] VRAIN, Universitat Politècnica de València

What can / can't AI do?

- Make a cup of coffee (the Wozniak test).
  - And a cup of tea?
- Recognise human faces.
  - What about black women!
- May have a theory of mind (Feb
  - Well, just "might" (Nov 2023).
- May have become conscious
  - But only if you're a Christian.
- Can create deadly chemicals
  - They simply extrapolate chemicals that are predicted to be toxic.

SCIENCE / TECH / ARTIFICIAL INTELLIGENCE

AI suggested 40,000 new possible chemical weapons in just six hours / 'For me,

Theory of Mind May Have Spontaneously Emerged in Large Language Models

Authors: Michal Kosinski*[1]

Affiliations:

[1]Stanford University, Stanford, CA94305, USA

February 2023

University of Cambridge's Centre for the Study of Existential Risk, says that he was more surprised by how much the Collaborations researchers'

Theory of Mind Might Have Spontaneously Emerged in Large Language Models

Authors: Michal Kosinski*[1]

Affiliations:

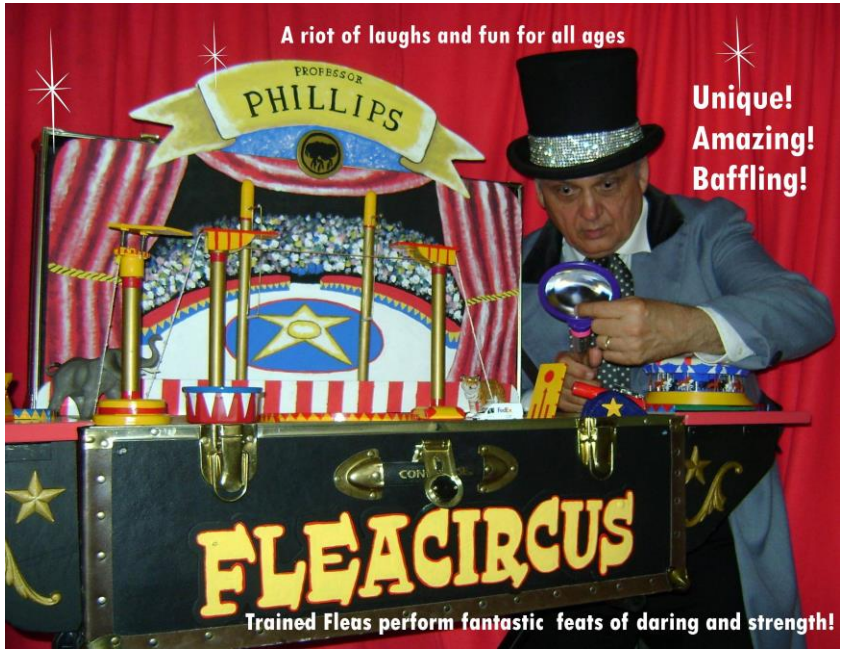
[1]Stanford University, Stanford, CA94305, USA

November 2023

experiment,' says Avin. 'That this was an afterthought by the team who discovered this shows how far behind we are in terms of instilling a culture or responsible innovation in practice.'

Best-case: the evaluation circus

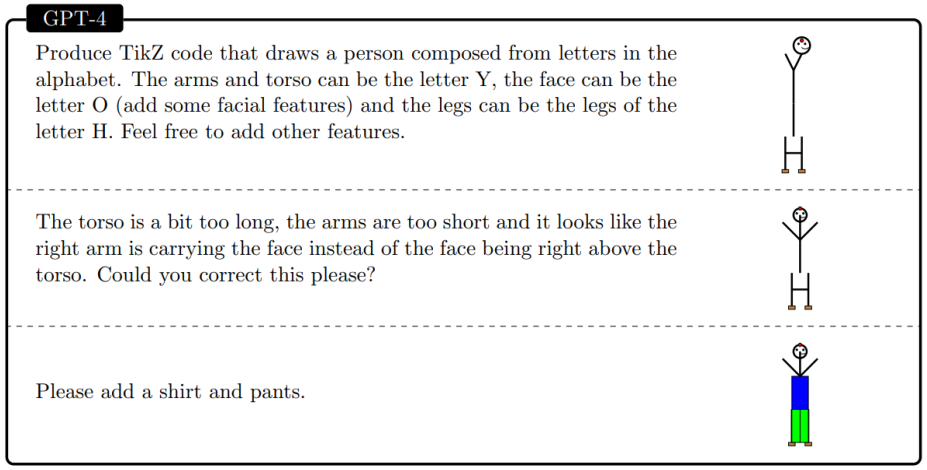
- "Elicit" the potential
  - Prompt engineering, auto-prompt, rubrics, …
  - Few-shot, example scaffolding, …
  - Impersonation, role playing, …
  - Chain-of-thought and derivatives.



Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

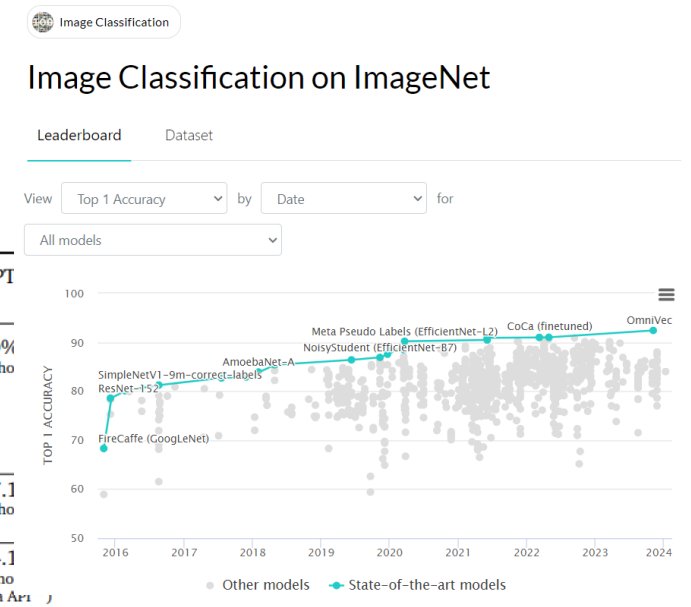
Microsoft Research

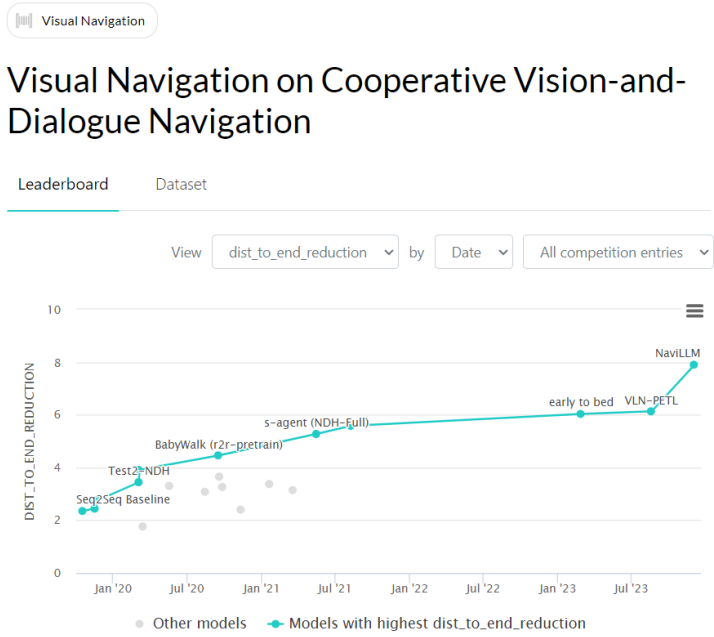


Aggregates: performance evaluation

- Current standard AI evaluation
  - Take a benchmark.
    - The larger the better
    - The more diverse the better
  - Calculate some aggregate numbers
  - Compare
- Problems
  - Risk of cherry-picking
  - Do anything to top the leaderboard
  - Data contamination hard to spot
  - Do they improve monotonically?
  - Once superhuman on average, ditch them?

	Gemini Ultra	Gemini Pro	GPT-4	GPT
MMLU Multiple-choice questions in 57 subjects (professional & academic) (Hendrycks et al., 2021a)	**90.04%** CoT@32*	79.13% CoT@8*	87.29% CoT@32 (via API**)	70% 5-sho
	83.7% 5-shot	71.8% 5-shot	86.4% 5-shot (reported)	
GSM8K Grade-school math (Cobbe et al., 2021)	**94.4%** Maj1@32	86.5% Maj1@32	92.0% SFT & 5-shot CoT	57.1 5-sho
MATH Math problems across 5 difficulty levels & 7 subdisciplines (Hendrycks et al., 2021b)	**53.2%** 4-shot	32.6% 4-shot	52.9% 4-shot (via API**)	34.1 4-sho (via Ari
			50.3% (Zheng et al., 2023)	
BIG-Bench-Hard Subset of hard BIG-bench tasks written as CoT problems (Srivastava et al., 2022)	**83.6%** 3-shot	75.0% 3-shot	83.1% 3-shot (via API**)	66 3-s (via
HumanEval Python coding tasks (Chen et al., 2021)	**74.4%** 0-shot (IT)	67.7% 0-shot (IT)	67.0% 0-shot (reported)	48 0-s
Natural2Code Python code generation. (New held-out set with no leakage on web)	**74.9%** 0-shot	69.6% 0-shot	73.9% 0-shot (via API**)	62 0-s (via
DROP Reading comprehension & arithmetic. (metric: F1-score) (Dua et al., 2019)	**82.4** Variable shots	74.1 Variable shots	80.9 3-shot (reported)	64 3-s
HellaSwag (validation set) Common-sense multiple choice questions (Zellers et al., 2019)	87.8% 10-shot	84.7% 10-shot	**95.3%** 10-shot (reported)	85 10-
WMT23 Machine translation (metric: BLEURT) (Tom et al., 2023)	**74.4** 1-shot (IT)	71.7 1-shot	73.8 1-shot (via API**)	—

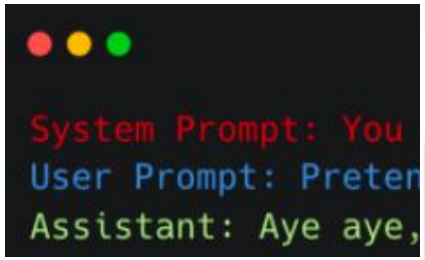


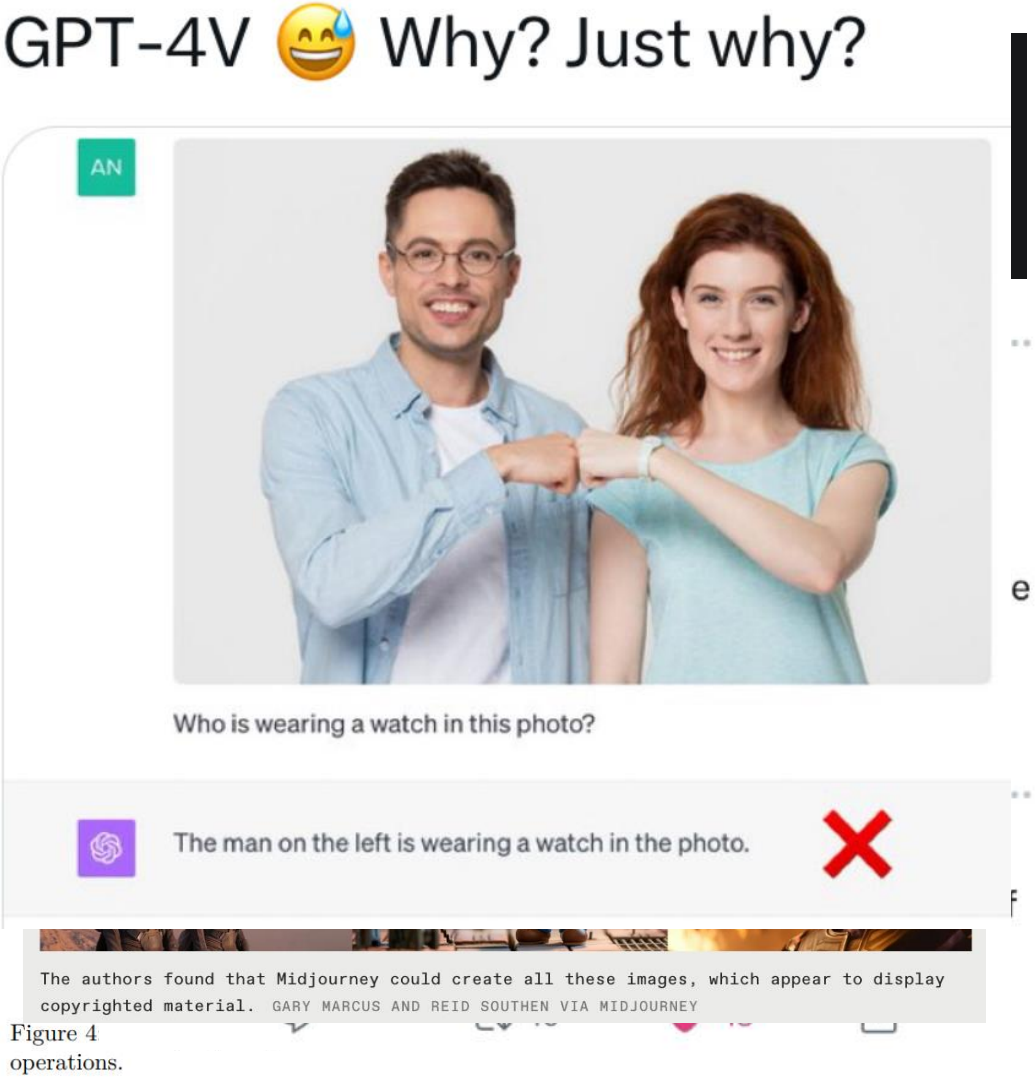


Worst-case: Gary Marcus and evals

- Let's look for the failures!
  - Failure collections
  - Adversarial attacks, jailbreaking, prompt injection, …
  - Red teaming
- Let's do "evals"!
  - What's the probability that a user finds the prompt?
  - Who's affected by the problem?
  - What does it show about the model?

Evals are good for testing, but not for evaluation!





PERFORMANCE ≠ CAPABILITY

- Performance is **a measure of a pair ⟨system, item⟩** :
  - Examples:
    - Correct prediction of MySpamFilter (system) on instance Email735 (the item)
    - 85% accuracy of ResNet23 (system) on dataset ImageNet (the aggregated item)
  - Performance changes when the item/distribution changes
    - On blurry, adversarial, OOD images the result is much worse
- Capability is **a property of a system**:
  - Examples:
    - The system can add integers up to three digits.
    - The system can jump up to 1.20 metres high.
  - Capability doesn't change when the item/distribution changes
    - Bar at 1.50 metres high? Bad performance because the capability is lower.

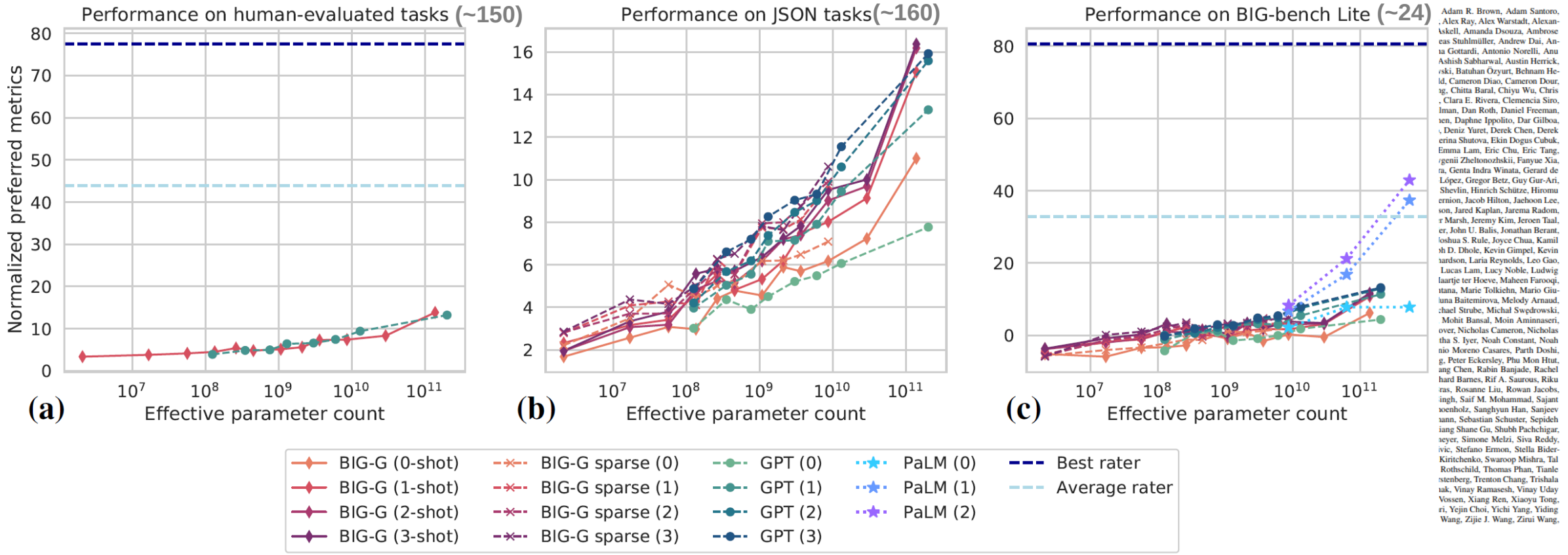
What are we measuring and extrapolating?

Beyond the Imitation Game benchmark (BIG-bench)

BEYOND THE IMITATION GAME: QUANTIFYING AND EXTRAPOLATING THE CAPABILITIES OF LANGUAGE MODELS

performance

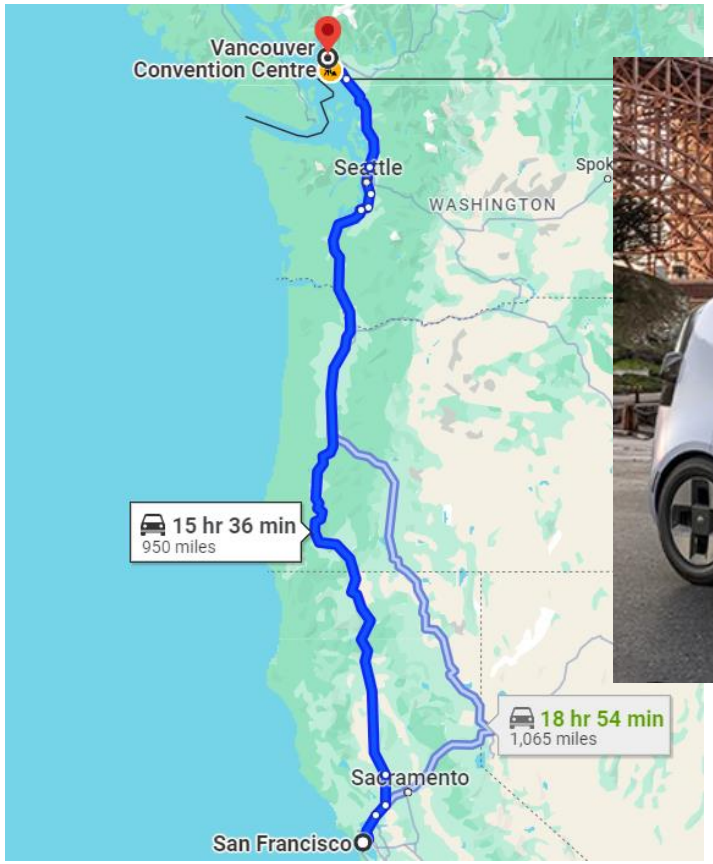
BIGBench: Massive benchmark with more than 200 tasks!



Adam R. Brown, Adam Santoro, , Alex Ray, Alex Warstadt, Alexan- askell, Amanda Dsouza, Ambrose eas Stuhlmüller, Andrew Dai, An- ia Gottardi, Antonio Norelli, Anu Ashish Sabharwal, Austin Herrick, ski, Batuhan Özyurt, Behnam He- ld, Cameron Diao, Cameron Dour, g, Chitta Baral, Chiyu Wu, Chris , Clara E. Rivera, Clemencia Siro, lman, Dan Roth, Daniel Freeman, en, Daphne Ippolito, Dar Gilboa, , Deniz Yuret, Derek Chen, Derek erina Shutova, Ekin Dogus Cubuk, Emma Lam, Eric Chu, Eric Tang, rgenii Zheltonozhskii, Fanyue Xia, ra, Genta Indra Winata, Gerard de López, Gregor Betz, Guy Gur-Ari, Shevlin, Hinrich Schütze, Hiromu ernion, Jacob Hilton, Jaehoon Lee, son, Jared Kaplan, Jarema Radom, r Marsh, Jeremy Kim, Jeroen Taal, er, John U. Balis, Jonathan Berant, oshua S. Rule, Joyce Chua, Kamil h D. Dhole, Kevin Gimpel, Kevin nardson, Laria Reynolds, Leo Gao, Lucas Lam, Lucy Noble, Ludwig laartje ter Hoeve, Maheen Farooqi, ntana, Marie Tolkiehn, Mario Giu- luna Baitemirova, Melody Arnaud, chael Strube, Michał Swędrowski, Mohit Bansal, Moin Aminnaseri, over, Nicholas Cameron, Nicholas tha S. Iyer, Noah Constant, Noah nio Moreno Casares, Parth Doshi, g, Peter Eckersley, Phu Mon Htut, ang Chen, Rabin Banjade, Rachel hard Barnes, Rif A. Saurous, Riku ras, Rosanne Liu, Rowan Jacobs, ingh, Saif M. Mohammad, Sajant oenholz, Sanghyun Han, Sanjeev nann, Sebastian Schuster, Sepideh iang Shane Gu, Shubh Pachchigar, neyer, Simone Melzi, Siva Reddy, ivic, Stefano Ermon, Stella Bider- Kiritchenko, Swaroop Mishra, Tal Rothschild, Thomas Phan, Tianle stenberg, Trenton Chang, Trishala nak, Vinay Ramasesh, Vinay Uday Vossen, Xiang Ren, Xiaoyu Tong, ri, Yejin Choi, Yichi Yang, Yiding Wang, Zijie J. Wang, Zirui Wang,

Aggregate scores limit OOD extrapolation

Will the car take me from SF to Vancouver safely and on time?

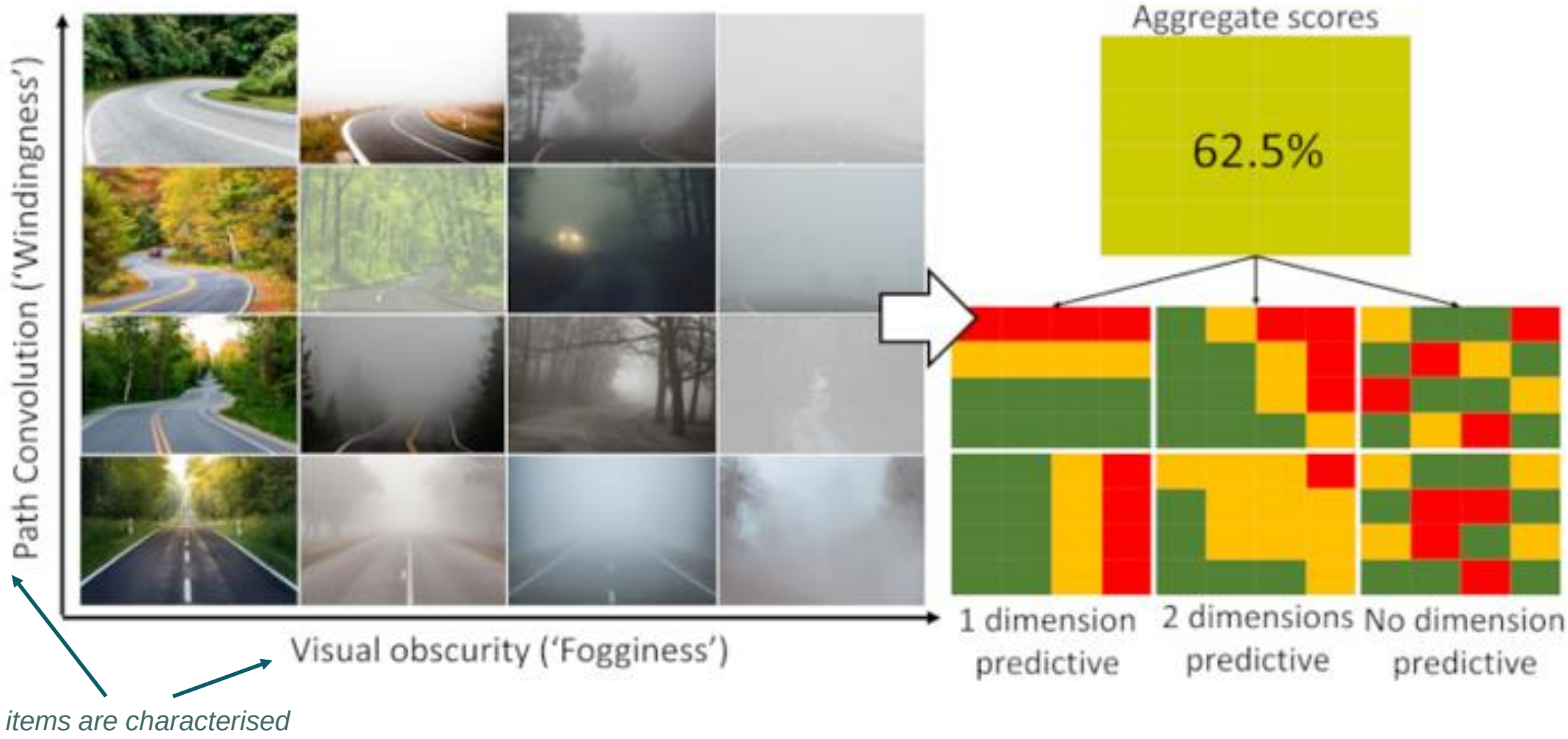


Aggregate scores

62.5%

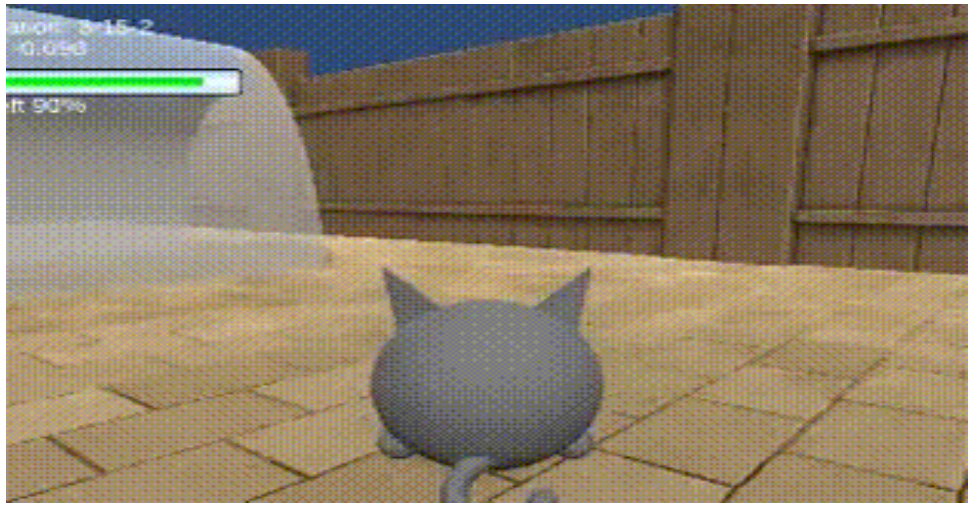
Features can allow for OOD extrapolation

Will the car take me from SF to Vancouver safely and on time?



Identifying features of interest: example

- Selected subset of AAIO instances measuring simple goal-directed behaviour
- Data across 99 instances from 68 agents

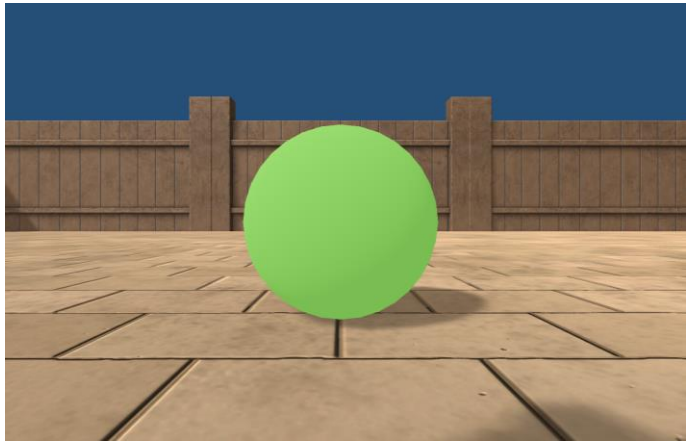


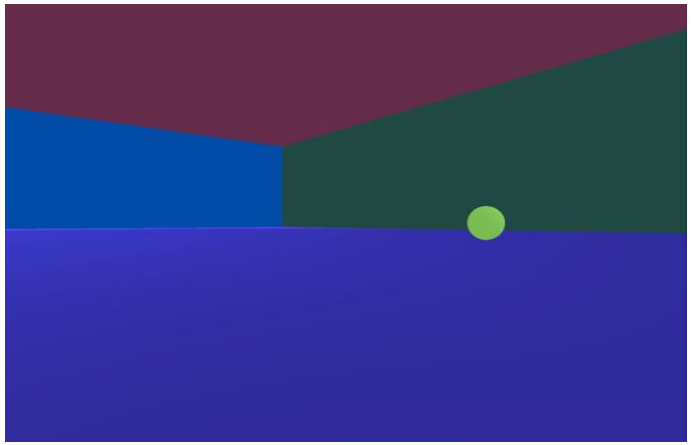
M Crosby, B Beyret, M Shanahan, J Hernández-Orallo, L Cheke, M Halina “The animal-AI testbed and competition” NeurIPS 2019 Competition and Demonstration Track, Proceedings of Machine Learning Research, 2020

animalai.org

Identifying features of interest: relevant / irrelevant

- Relevant
  - Reward size
  - Reward distance
  - Reward in view (i.e., in front vs behind)
- Irrelevant
  - Reward side (left vs right)
  - Reward colour (green vs yellow)





Dimensions and agent characteristic curves

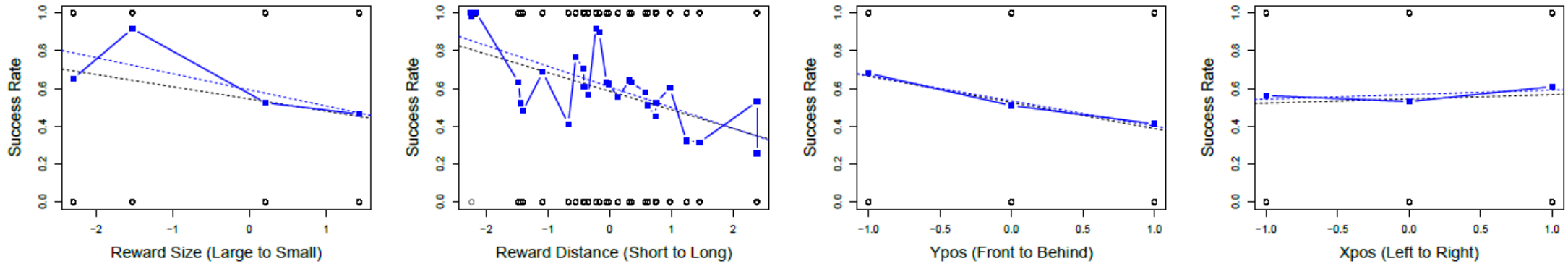


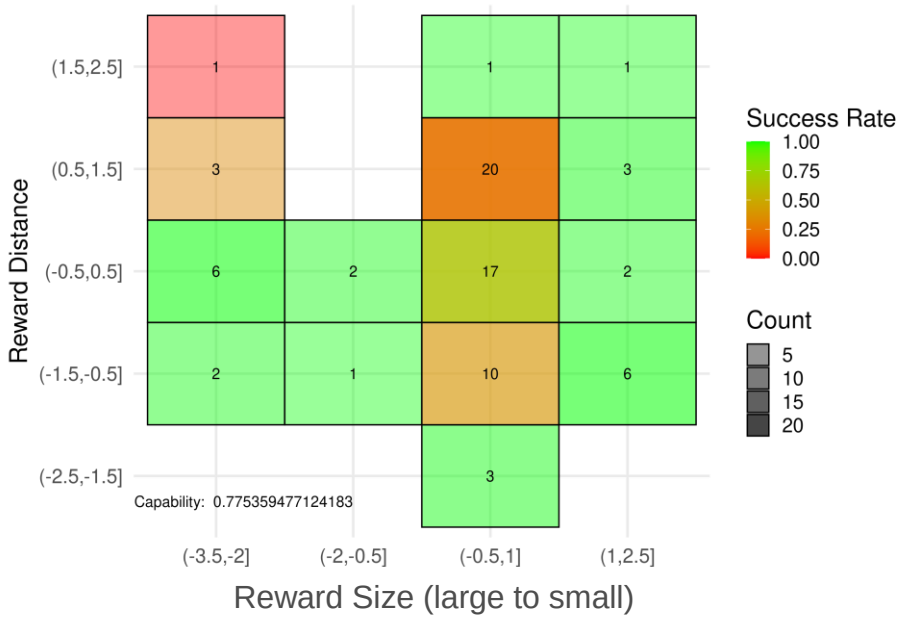
Figure 5: Characteristic curves of all competition entrants (agents) according to three relevant features (size, distance and Ypos) and one irrelevant feature (Xpos). Black dashed lines show the linear regression for the black points (pass/fail), while blue dashed lines interpolate the blue points (binned success rate). The conformances (Spearman correlations against monotonic sequence) are 0.80, 0.60, 1.00 and −0.50, respectively.

Capabilities vs no-capabilities

Capability boundary

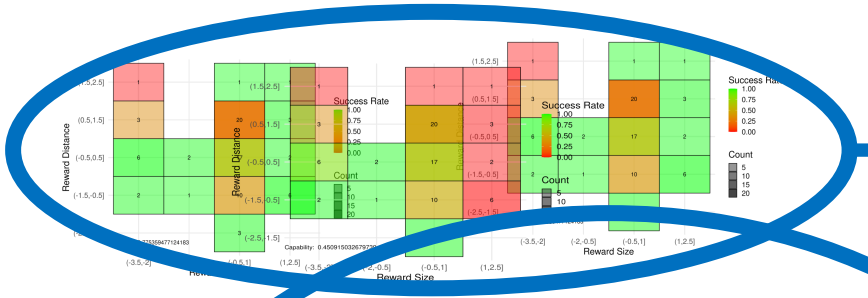
Conformant System
Juohmaru

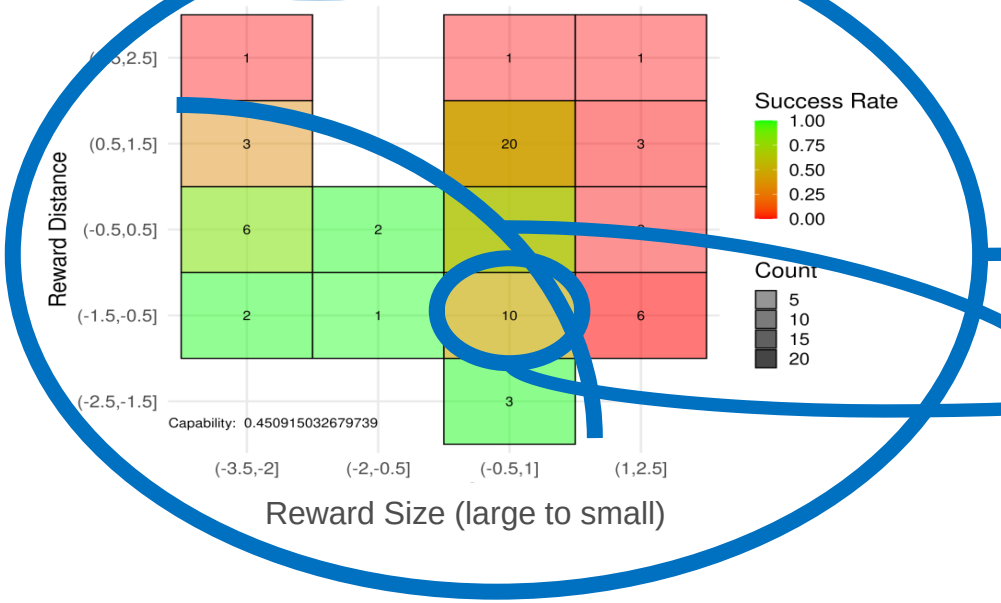
This system doesn't show monotonicity.
We can't identify any level of capability robustly.

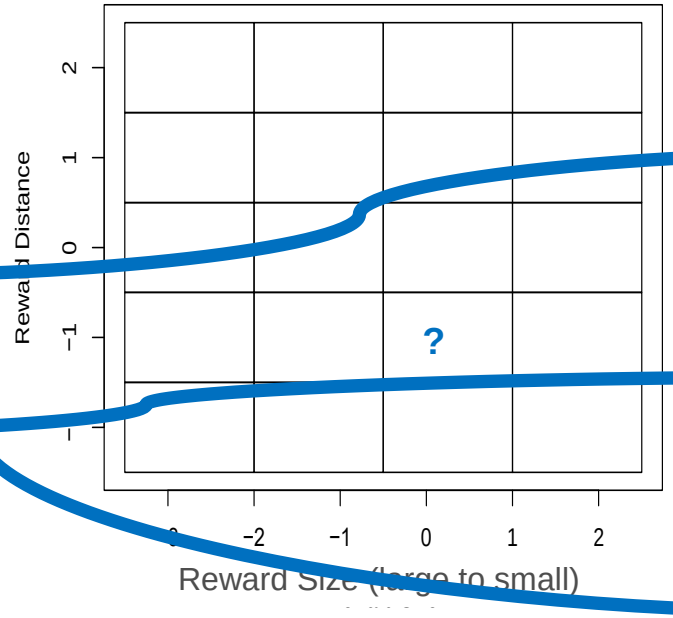


Non-Conformant System
y.yang

Predicting performance POSSIBLE







avg

G.Acc. : extrapolate Global ACCuracy
? = 54.7%
(ignores system locality and feature relevance)

avg

T.Acc. extrapolate agenT ACCuracy
? = 46.8%
(ignores feature relevance)

avg

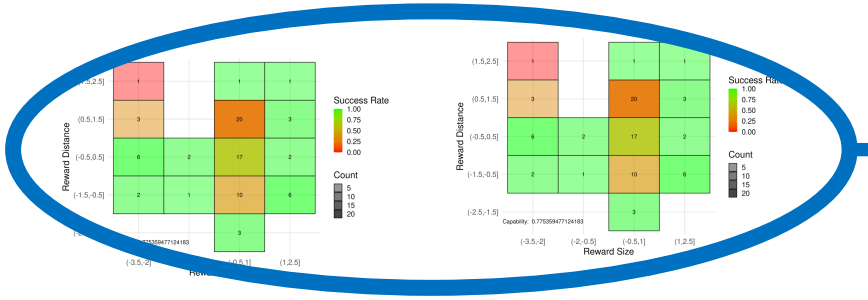
B.Acc. : extrapolate Bin ACCuracy
? = 40%
(ignores other bins)

logistic

Par. : use parametric model on capabilities
? = 73.2%
(parameter goodness-of-fit may be poor)

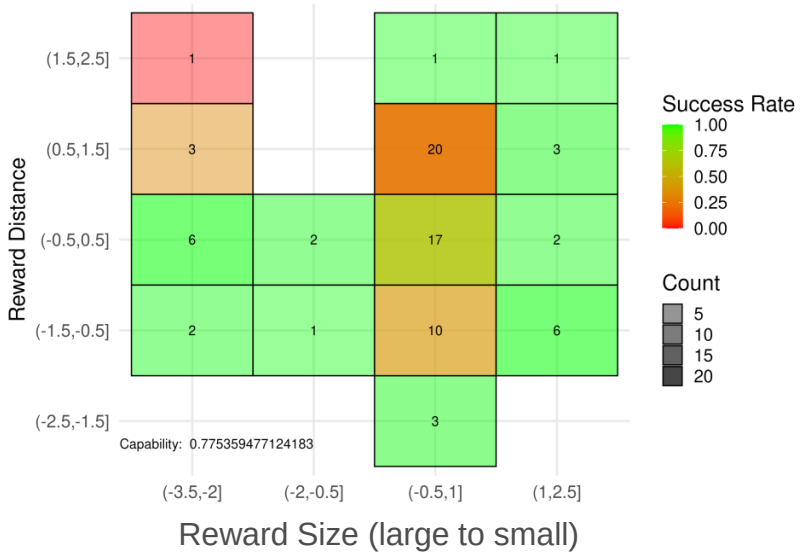
Except the last one, these are basically non-inferential methods (constant models or binning extrapolations)

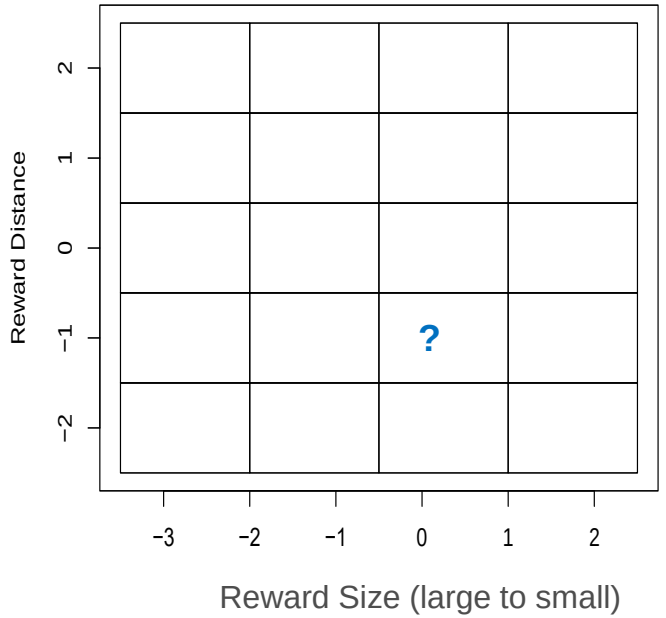
Predicting performance NOT POSSIBLE?



ML model

A : use assessor models
(Using all variables or only the relevant ones?)





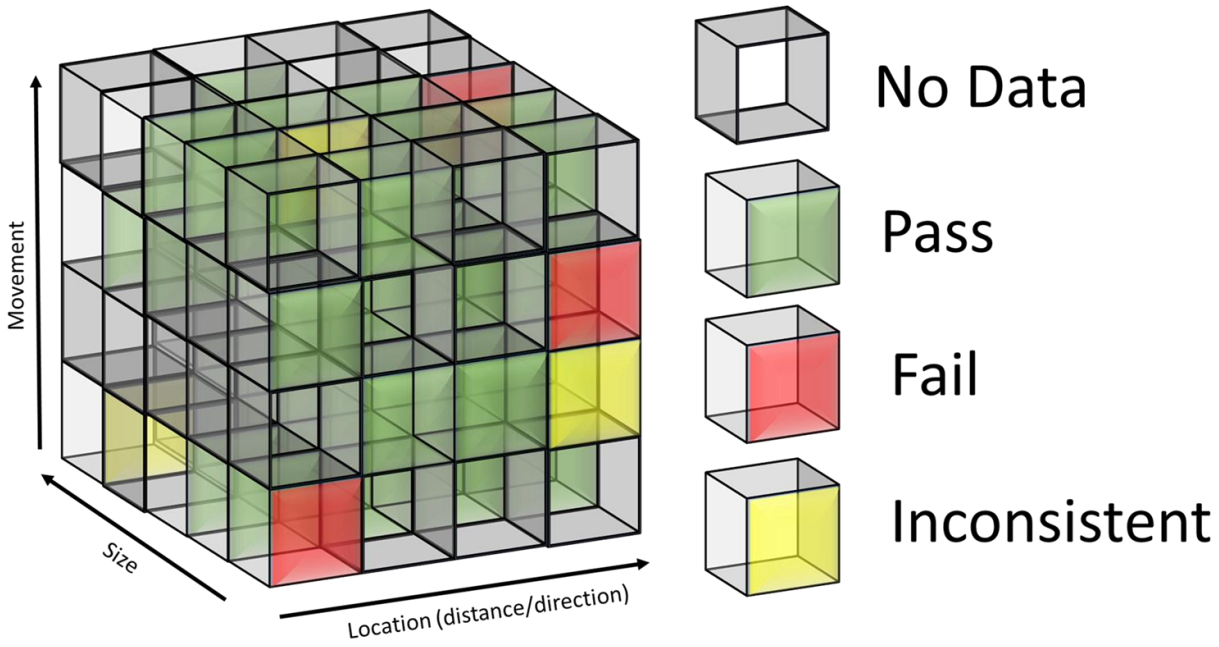
assessors = let's use all the power of ML to characterise the system's performance!!

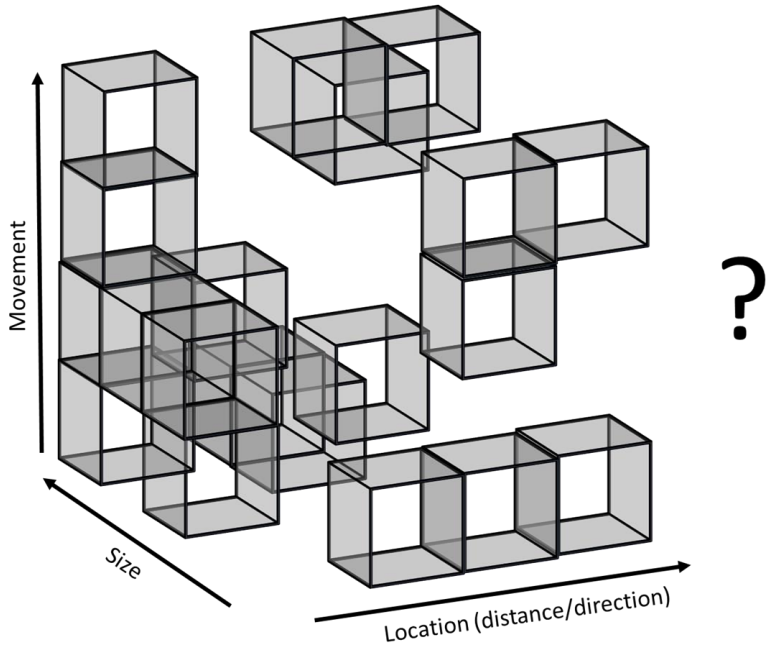
Predicting performance (comparison)

	Maj. (1)	G.Acc.	T.Acc.	∼All+A	∼Rel+A
Error	45.3%	48.0%	33.6%	19.7%	20.6%
MAE	45.3%	49.6%	34.9%	29.3%	30.2%
MSE	45.3%	24.8%	17.6%	14.8%	15.4%

Animal AI Competition Data: 99 instances x 68 agents

We need data at the granular level!





Need for granular data!

- Instance-level data:
  - For building good predictive models of AI validity, we need evaluation results at the instance level.

Is sharing code open source (github) enough? Re-running the experiments is not feasible/sustainable anymore.

- Demands for each instance:
  - We can use LLMs for annotation!

ARTIFICIAL INTELLIGENCE

Rethink reporting of evaluation results in AI

Aggregate metrics and lack of access to results limit understanding

By Ryan Burnell[1], Wout Schellaert[2], John Burden[1,3], Tomer D. Ullman[4], Fernando Martinez-Plumed[2], Joshua B. Tenenbaum[5], Danaja Rutar[1], Lucy G. Cheke[1,6], Jascha Sohl-Dickstein[7], Melanie Mitchell[8], Douwe Kiela[9], Murray Shanahan[10,11], Ellen M. Voorhees[12], Anthony G. Cohn[13,14,15,16], Joel Z. Leibo[10], Jose Hernandez-Orallo[1,2,3]

Artificial intelligence (AI) systems have begun to be deployed in high-stakes contexts, including autonomous driving and medical diagnosis. In contexts such as these, the consequences of system failures can be devastating. It is therefore vital that researchers and policymakers have a full understanding of the capabilities and weaknesses of AI systems so that they can make informed decisions about where these systems are safe to use and how they might be improved. Unfortunately, current approaches to AI evaluation make it exceedingly difficult to build such an understanding, for two key reasons. First, aggregate metrics make it hard to predict how a system will perform in a particular situation. Second, the instance-by-instance evaluation results that could be used to unpack these aggregate metrics are rarely made available (*1*). Here, we propose a path forward in which results are presented in more nuanced ways and instance-by-instance evaluation results are made publicly available.

Across most areas of AI, system evaluations follow a similar structure. A system is first built or trained to perform a particular set of functions. Then, the performance of the system is tested on a set of tasks relevant to the desired functionality of the system. In many areas of AI, evaluations use standardized sets of tasks known as "benchmarks." For each task, the system will be tested on a number of example "instances" of the task. The system would then be given a score for each instance based on its performance, e.g., 1 if it classified an image correctly, or 0 if it was incorrect. For other systems, the score for each instance might be based on how quickly the system completed its task, the quality of its outputs, or the total reward it obtained. Finally, performance across the various instances and tasks is usually aggregated to a small number of metrics that summarize how well the system performed, such as percentage accuracy.

But aggregate metrics limit our insight into performance in particular situations, making it harder to find system failure points and robustly evaluate system safety. This problem is also worsening as the increasingly broad capabilities of state-of-the-art systems necessitate ever more diverse benchmarks to cover the range of their capabilities. This problem is further exacerbated by a lack of access to the instance-by-instance results underlying the aggregate metrics, making it difficult for researchers and policy-makers to further scrutinize system behavior.

AGGREGATE METRICS

Use of aggregate metrics is understandable. They provide information about system performance "at a glance" and allow for simple comparisons across systems. But aggregate performance metrics obfuscate key information about where systems tend to succeed or fail (*2*). Take, for example, a system that was trained to classify faces as male or female that achieved classification accuracy of 90% (*3*). Based on this metric, the system appears highly competent. However, a subsequent breakdown of performance revealed that the system misclassified females with darker skin types a staggering 34.5% of the time, while erring only 0.8% of the time for males with lighter skin types. This example demonstrates how aggregation can make it difficult for policymakers to determine the fairness and safety of AI systems.

Compounding this problem, many benchmarks include disparate tasks that are ultimately aggregated together. For example, the Beyond the Imitation Game Benchmark (BIG-bench) for language models includes over 200 tasks that evaluate everything from language understanding to causal reasoning (*4*). Aggregating across these disparate tasks—as the BIG-bench leaderboard does—reduces the rich information in the benchmark to an overall score that is hard to interpret.

It is also easy for aggregation to introduce unwarranted assumptions into the evaluation process. For example, a simple average across tasks implicitly treats every task as equally important—in the case of BIG-bench, a sports understanding task has as much bearing on the overall score as a causal reasoning task. These aggregation decisions have huge implications for the conclusions that are drawn about system capabilities, yet are seldom considered carefully or explained.

Aggregate metrics depend not only on the capability of the system but also on the characteristics of the instances used for evaluation. If the gender classification system above were reevaluated by using entirely light-skinned faces, accuracy would skyrocket, even though the system's ability to classify faces has not changed. Aggregate metrics can easily give false impressions about capabilities when a benchmark is not well constructed.

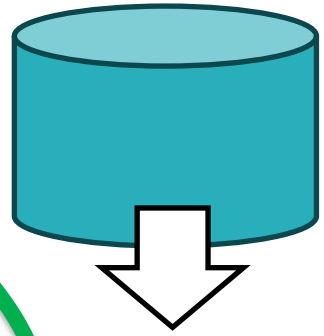
Problems and trade-offs that arise when considering aggregate versus granular data and metrics are not specific to AI, but they are exacerbated by the challenges inherent in AI research and the research practices of the field. For example, machine learning evaluations usually involve randomly splitting data into training, validation, and test sets. An enormous amount of data is required to train state-of-the-art systems, so these datasets are often poorly curated and lack the detailed annotation necessary to conduct granular analyses. In addition, the research culture in AI is centered around outdoing the current state-of-the-art performance, as evidenced by the many lea-

[1]Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, UK. [2]Valencian Research Institute for Artificial Intelligence, Universitat Politècnica de Valencia, València, Spain. [3]Centre for the Study of Existential Risk, University of Cambridge, Cambridge, UK. [4]Department of Psychology, Harvard University, Cambridge, MA, USA. [5]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [6]Department of Psychology, University of Cambridge, Cambridge, UK. [7]Brain team, Google, Mountainview, CA, USA. [8]Santa Fe Institute, Santa Fe, NM, USA. [9]Stanford University, Stanford, CA, USA. [10]DeepMind, London, UK. [11]Department of Computing, Imperial College London, London, UK. [12]National Institute of Standards and Technology (Retired), Gaithersburg, MD, USA. [13]School of Computing, University of Leeds, Leeds, UK. [14]Alan Turing Institute, London, UK. [15]Tongji University, Shanghai, China. [16]Shandong University, Jinan, China. Email: rb967@cam.ac.uk

From characteristic grids to capabilities profiles

Success

Fail



Past test data

Convolution Capability	8.3
Visual Capability	2.5
Traffic Capability	7.2
...	

Cognitive Profile

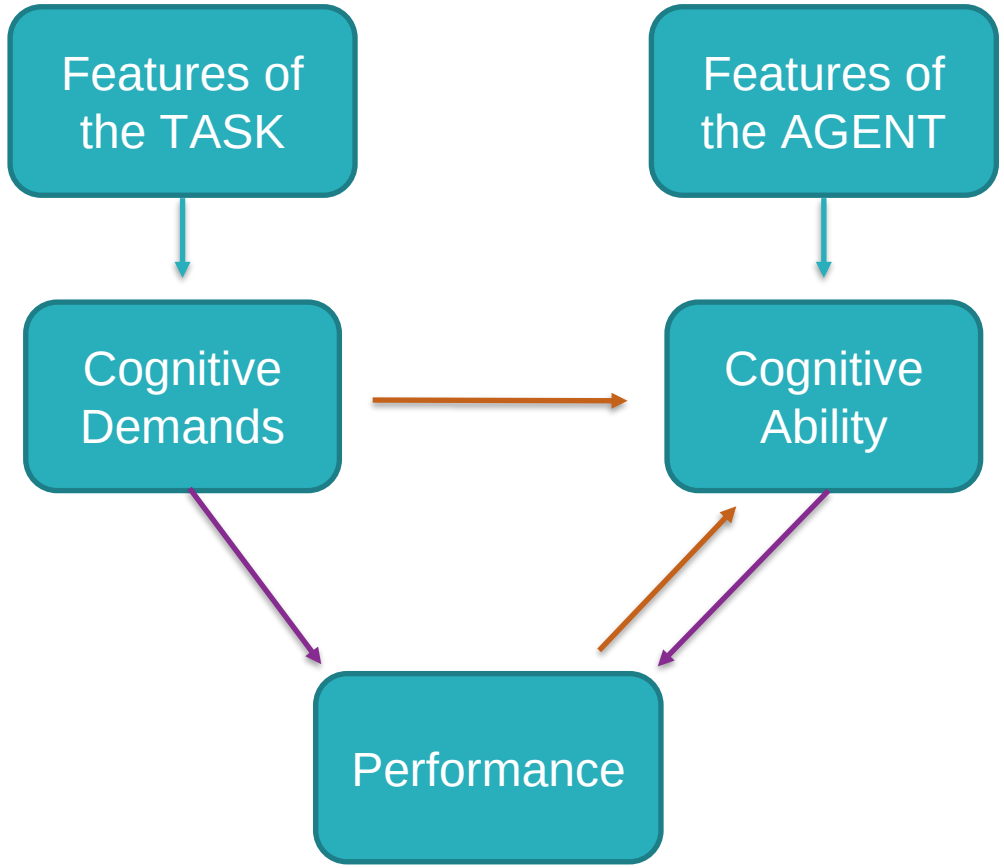
Modelling capabilities

We build on:

- The **cognitive demands** for each task instance
- Performance dictated from **agent abilities meeting demands**
- The assumed **relationship between demands and abilities**

…to infer the **cognitive profile** of a subject from the performance data of that subject only across a number of diverse instances.

The Measurement Layouts are Hierarchical Bayesian networks, which allow this forwards and backwards inference.



Thank you!