

Predictable Artificial Intelligence: The Case of AI Evaluation

José Hernández-Orallo^{1,2,3}

¹ VRAIN, Universitat Politècnica de València

² Leverhulme Centre for the Future of Intelligence, University of Cambridge

³ Centre for the Study of Existential Risk, University of Cambridge

<http://josephorallo.webs.upv.es/>

I-X Seminar Series - Imperial College - 30 January 2024



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

 **VRAIN**



LEVERHULME CENTRE FOR THE
FUTURE OF INTELLIGENCE



CENTRE FOR THE STUDY OF
EXISTENTIAL RISK

"you can never really **predict** for any given question whether a large language model will give you a correct answer"

Gary Marcus, AI Digest, 14 August 2023.

Predictable AI

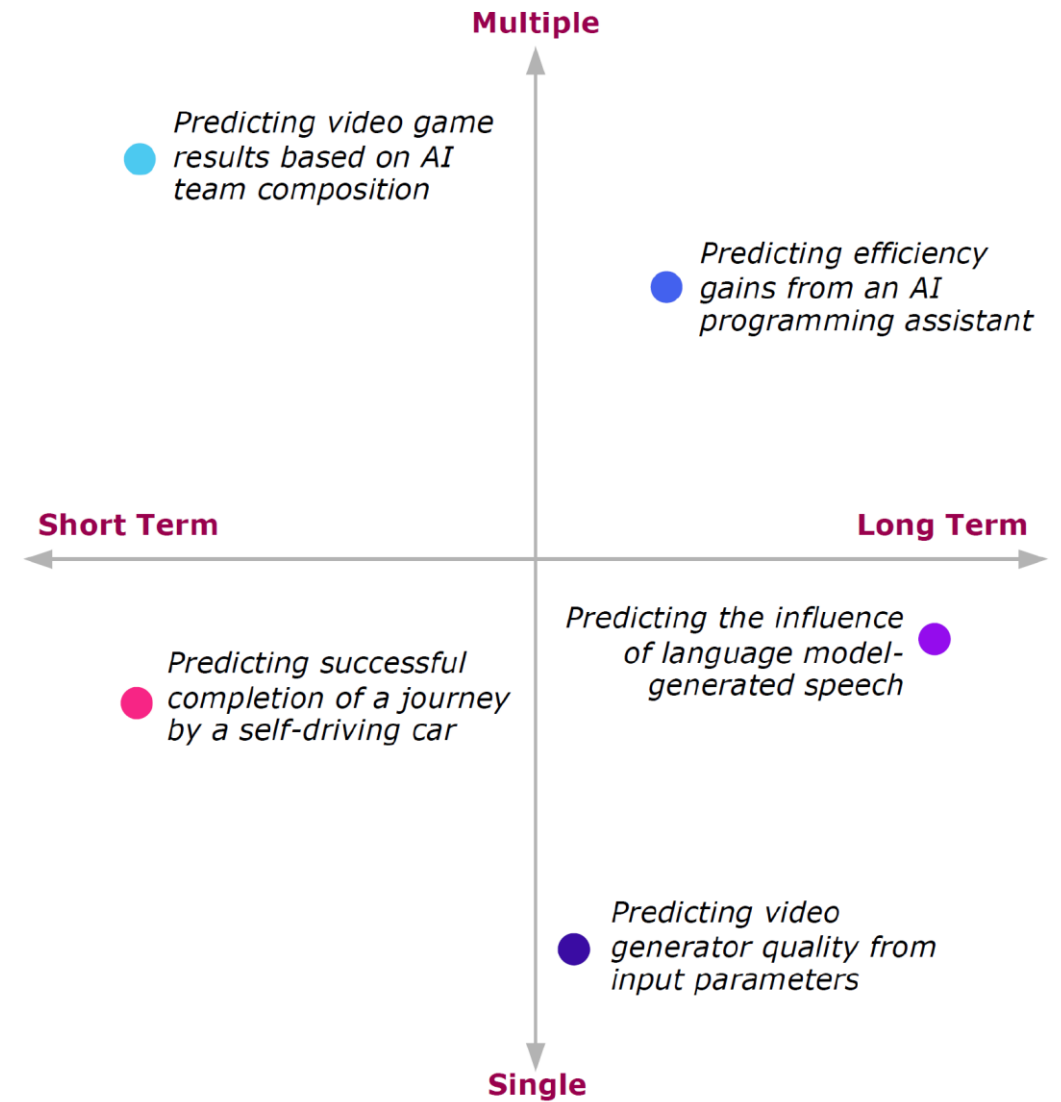
<https://www.predictable-ai.org/>
<https://arxiv.org/abs/2310.06167>

WHAT IS PREDICTABLE AI?

- AI Predictability is the extent to which key behavioural indicators of present and future AI ecosystems can be anticipated.
 - These indicators are measurable properties such as performance and safety.
- AI Predictability may refer to
 - anticipation in a specific context of use, such as a user query to a single AI system.
 - anticipation of future capabilities and safety issues several years ahead.

AI should aim for predictability,
not performance or even fool-proof validity.

Example	Inputs	Outputs
Self-driving car trip: A self-driving car is about to start a trip to the mountains. The weather is rainy and foggy. The navigator is instructed to use an eco route and adapt to traffic conditions but being free to choose routes and driving style. Before starting, the passengers want an estimate that the car will reach the destination safely.	The route, weather; traffic, time, trip settings, car's state, ...	Probability of safely reaching the destination.
Marketing speech generation: A request is made to a language model to generate a marketing speech based on an outline. The stakeholders expect the literal content of the speech to be original, or even surprising. What they really want to be predictable is whether the system will generate a speech along the outline, containing no offensive or biased content, and effectively persuading the audience to purchase the product.	Speech outline, audience demographics, potential restrictions, ...	Long-term impact of the speech on product purchases.
Video generation model training: An AI system is developed to create short music videos for a social media platform. Drawing from evaluations of prior video generation models and with additional audio and video training data, the plan is to train an upgraded model within a few weeks. The question to predict is the quality of this upgraded AI system, given model size, training data, learning epochs, etc; and the extent to which the videos will conform to content moderation standards.	Quantity of videos, compute, epochs, architecture specifications, ...	Quality and compliance of generated videos, according to human feedback.
AI assistant in software firm: A software company plans to deploy a new AI assistant to help programmers write, optimise and document their code. The question is how much efficiency (e.g., work hours in coding, documentations and maintenance) the company can get in the following six months.	AI assistant details, user profiles, ...	Efficiency metric (work hours saved).
AI agents in an online video game: In a popular online e-sports competition, several AI agents are to be used to form teams. The game developers have previously tested several multi-agent reinforcement learning algorithms. The developers want to anticipate the outcome of the next game based on the chosen algorithms and team members.	Team line-up (own and other teams), match level, ...	Match result (score)

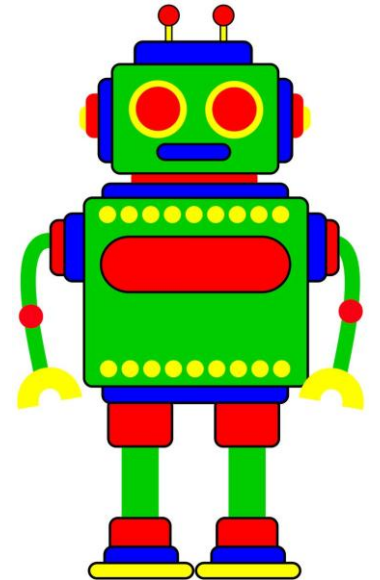


CENTRALITY

- trust?** -----> Can't rely on unpredictable outcomes
- liability?** -----> Eluding responsibility of unpredictable harms
- control?** -----> Hard to command an unpredictable system
- alignment?** -----> Unpredictable effects on the user's future wellbeing
- safety?** -----> No operating conditions under unpredictability
- Predictability of AI?*



Smart Humans



(Possibly Smarter) General-Purpose AI System

WHAT CAN BE PREDICTED?

- Any property that can be reliably anticipated and
 - can determine when, how or whether the system is worth being used in a given context.

- **Outputs:**

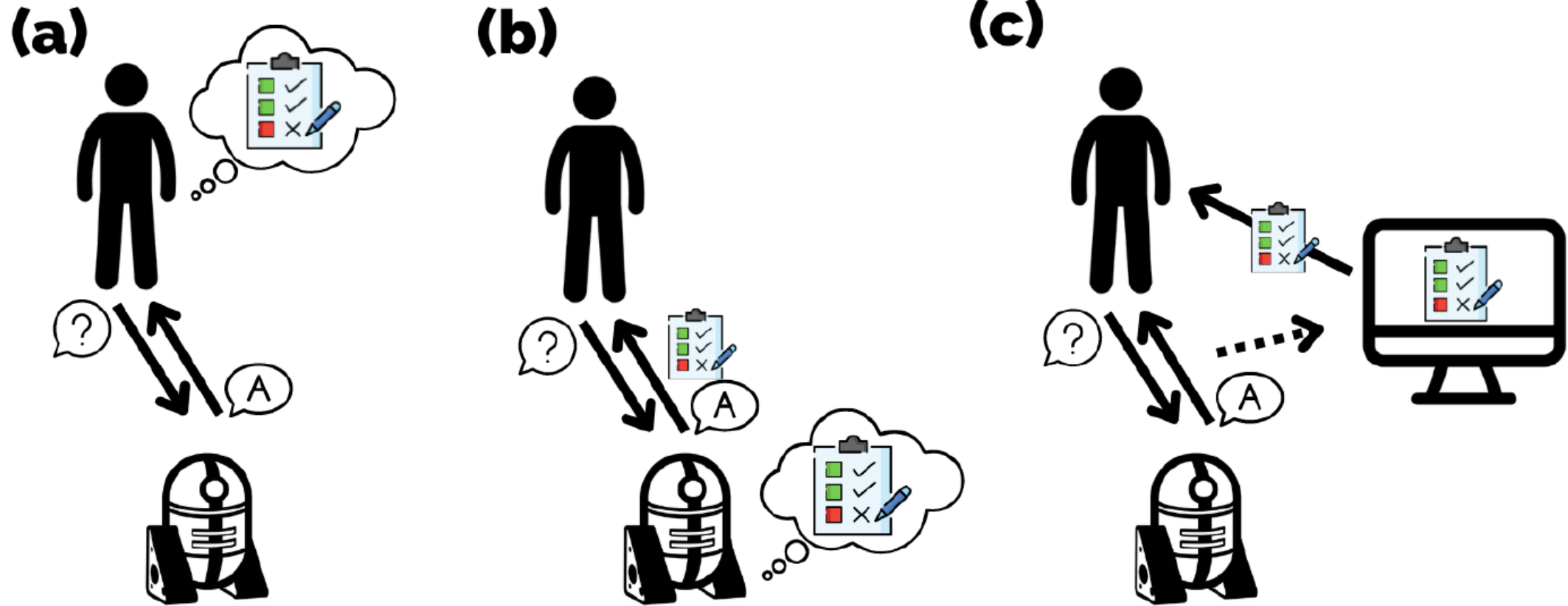
- correctness
- safety
- fairness
- energy consumption
- response time
- ...

Anticipativeness?
Granularity?
Temporality?
Hypotheticality?

- **Inputs:**

- $\langle \text{system, problem, context} \rangle$
- system metafeatures:
 - size, compute, architecture, ...
- problem metafeatures:
 - task demands/difficulties...
- context metrafeatures:
 - user profile, constraints, ...

WHO PREDICTS AND HOW?



A New Vision for AI evaluation: Predicting Validity

Anticipativeness? Yes

Granularity? Yes

Temporality? Short

Hypotheticality? Possibly 😊

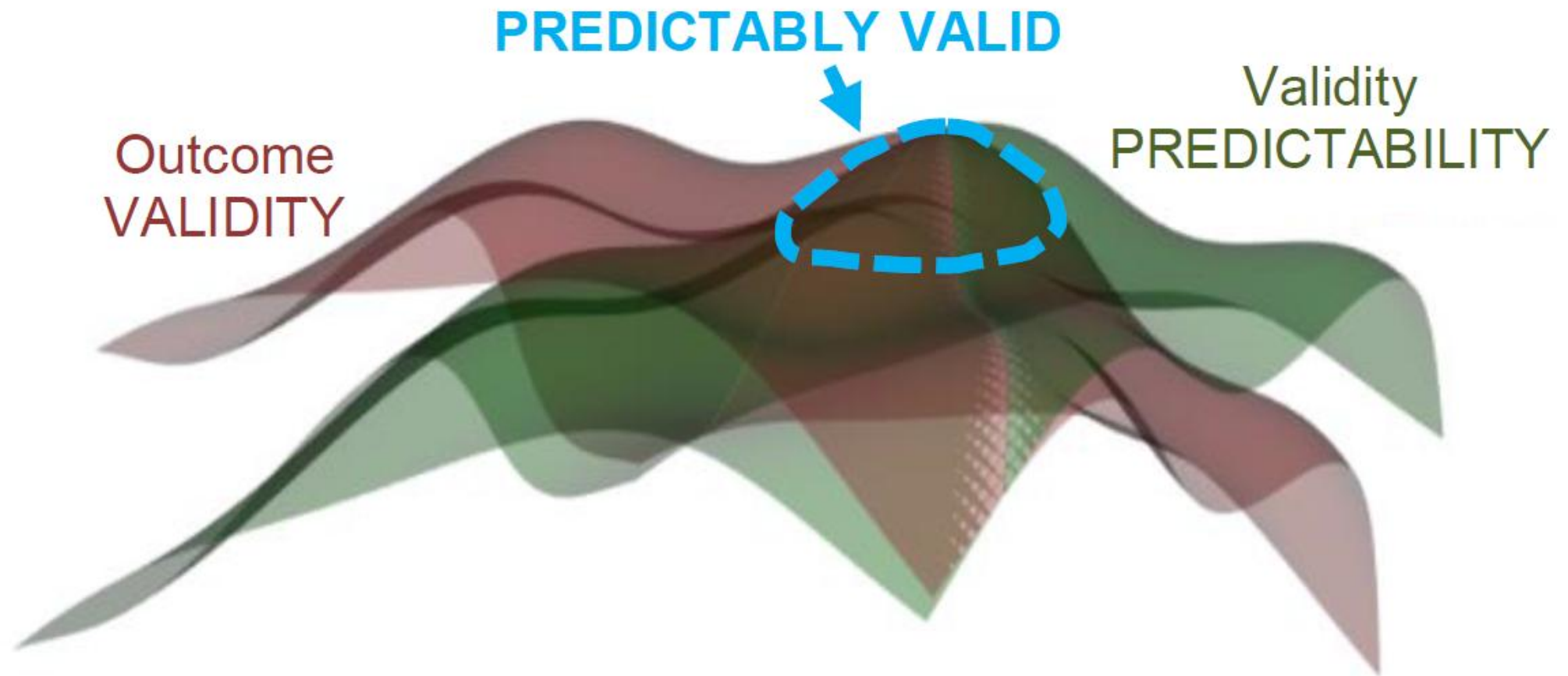
PREVAIL: Predictably Valid AI Landscapes
ERC Advanced Grant Proposal

PREDICTING VALIDITY FOR GPAI

- We can build predictive models to anticipate how valid a system is going to be for a particular instance and context of use.
- Extracting patterns of performance (from given features or extracting these features)
- Granular anticipation for the same and changing distributions!

AI Validity becomes a prediction problem

WHERE IS AI PREDICTABLY VALID?



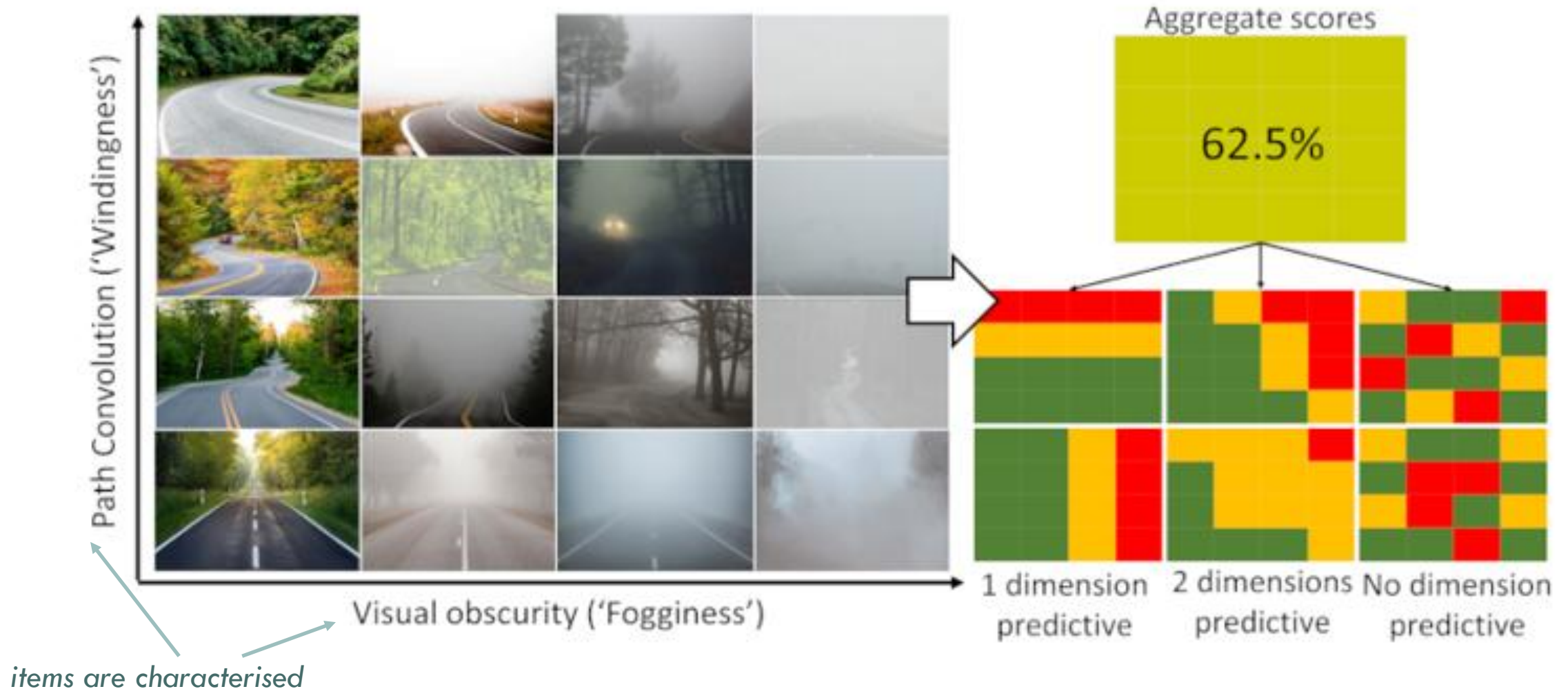
PREDICTING OUTCOME EASIER?

- Predicting behaviour
 - “Can we predict system behaviour in detail?”
 - Requires the same power as the original model
 - Fidelity-interpretability trade-off if we want to understand!
- Predicting outcome
 - “Can we predict system failure in detail?”
 - May require less power than the original model
 - It’s still useful if we don’t understand!



*We can't predict what the system will do
but we can predict the outcome*

WHERE WILL IT FAIL?



WILL IT WORK SAFELY IN THIS CASE?



Predictable AI from Capabilities

R Burnell, J Burden, D Rutar, K Voudouris, L Cheke, J Hernández-Orallo

“Not a Number: Identifying Instance Features for Capability-Oriented Evaluation”

IJCAI 2022

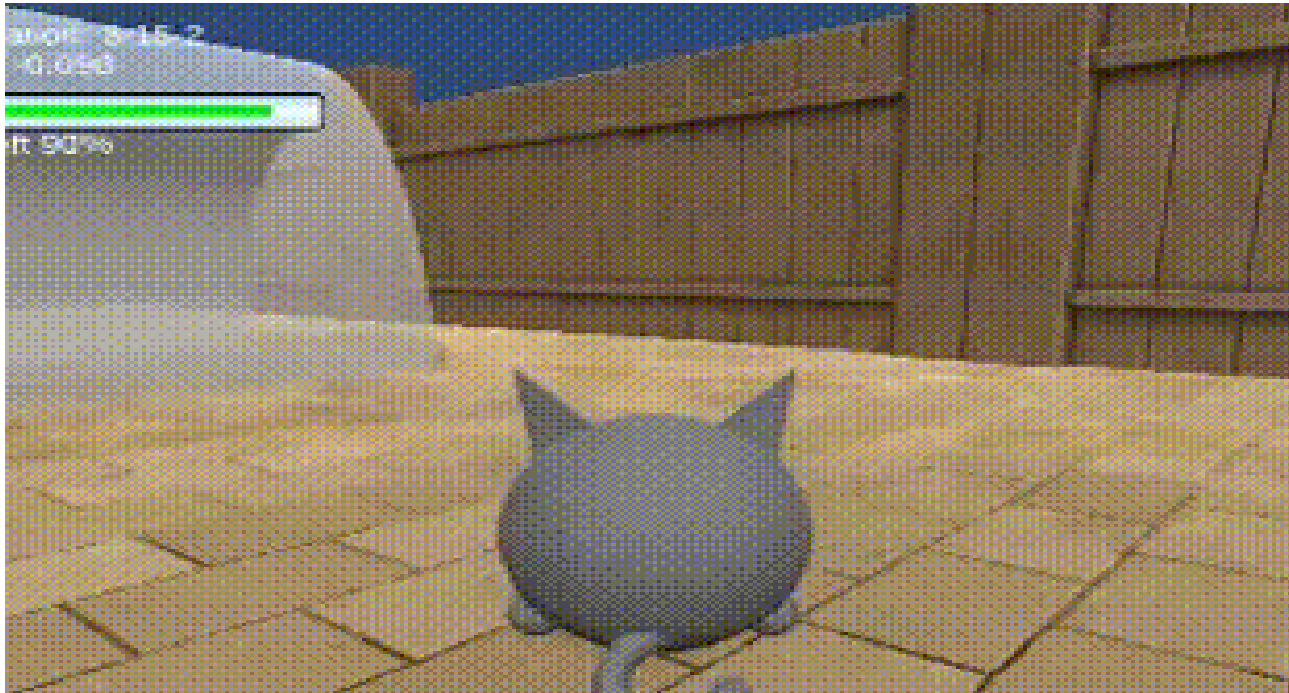
PERFORMANCE \neq CAPABILITY

- Performance is **a measure of a pair \langle system, item \rangle** :
 - 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.

Usually quantitative,
with a **magnitude**
and a **unit**.

HOW CAN WE IDENTIFY CAPABILITIES?

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

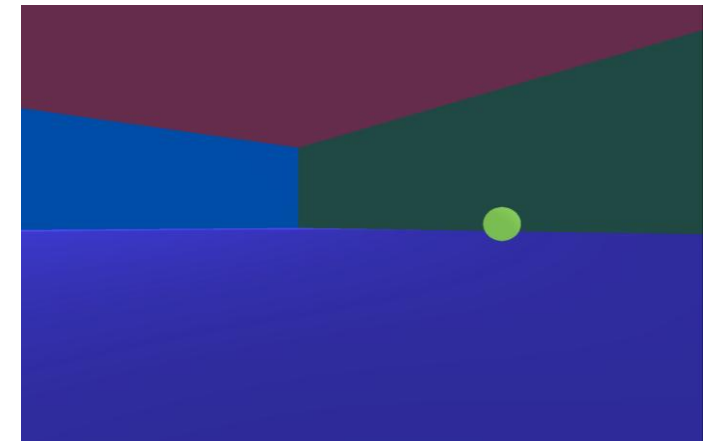
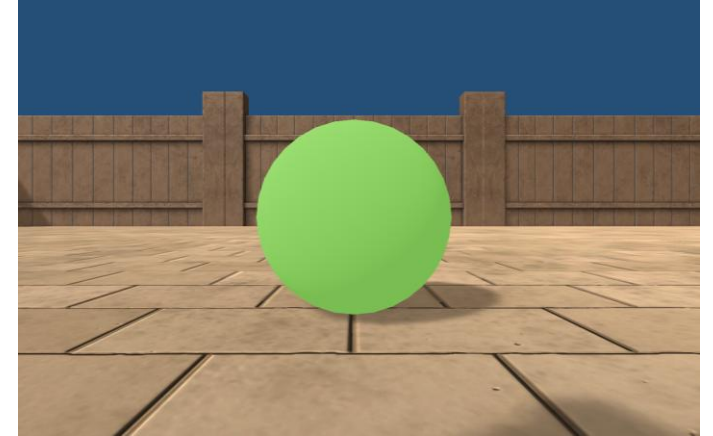


M Crosby, B Beyret, M Shanahan, J Hernández-Orallo, L Cheke, M Halina "The animal-AI testbed and competition" NeurIPS 2019 Competition and Demonstration Track, Proceedings of Machine Learning Research, 2020

<http://lcfi.ac.uk/projects/kinds-of-intelligence/animalaiolympics/>

IDENTIFYING FEATURES OF INTEREST

- **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

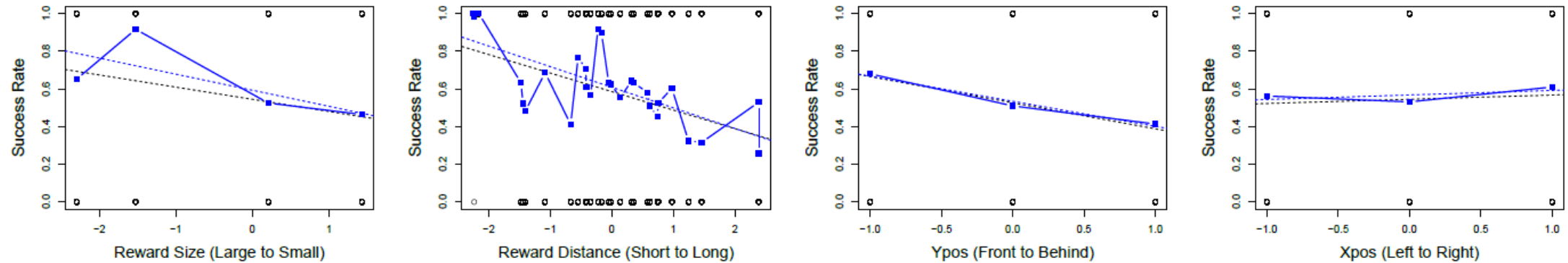
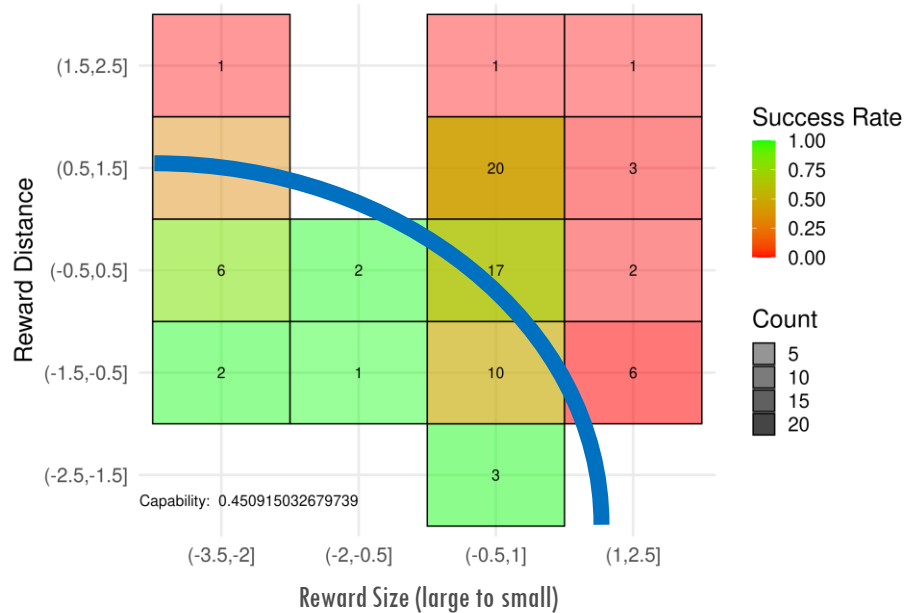


Figure 5: Characteristic curves of all competition entrants (agents) according to three relevant features (size, distance and Ypos) and one irrelevant feature (Xpos). Black dashed lines show the linear regression for the black points (pass/fail), while blue dashed lines interpolate the blue points (binned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

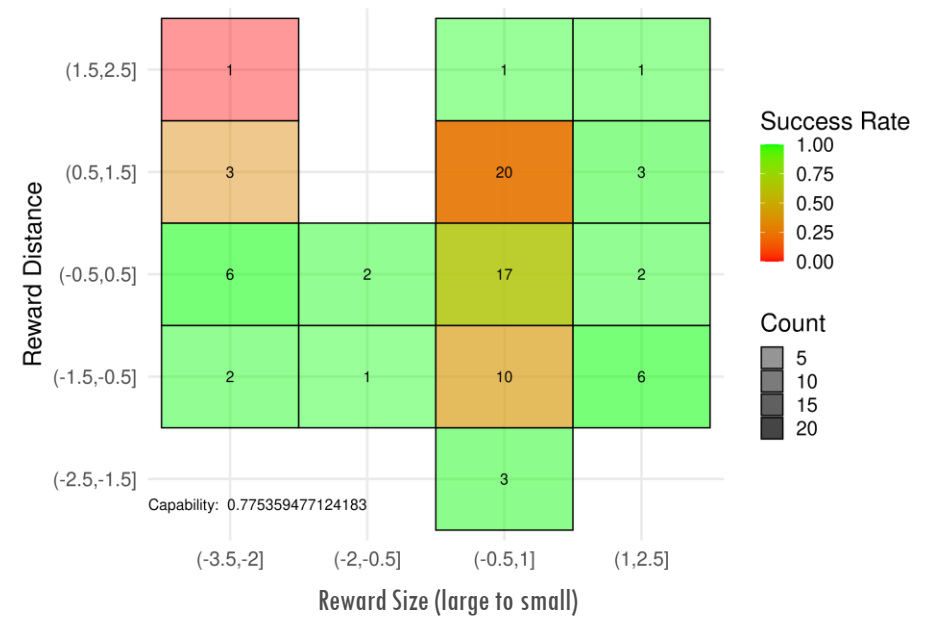
CAPABILITIES VS NO-CAPABILITIES

Capability boundary



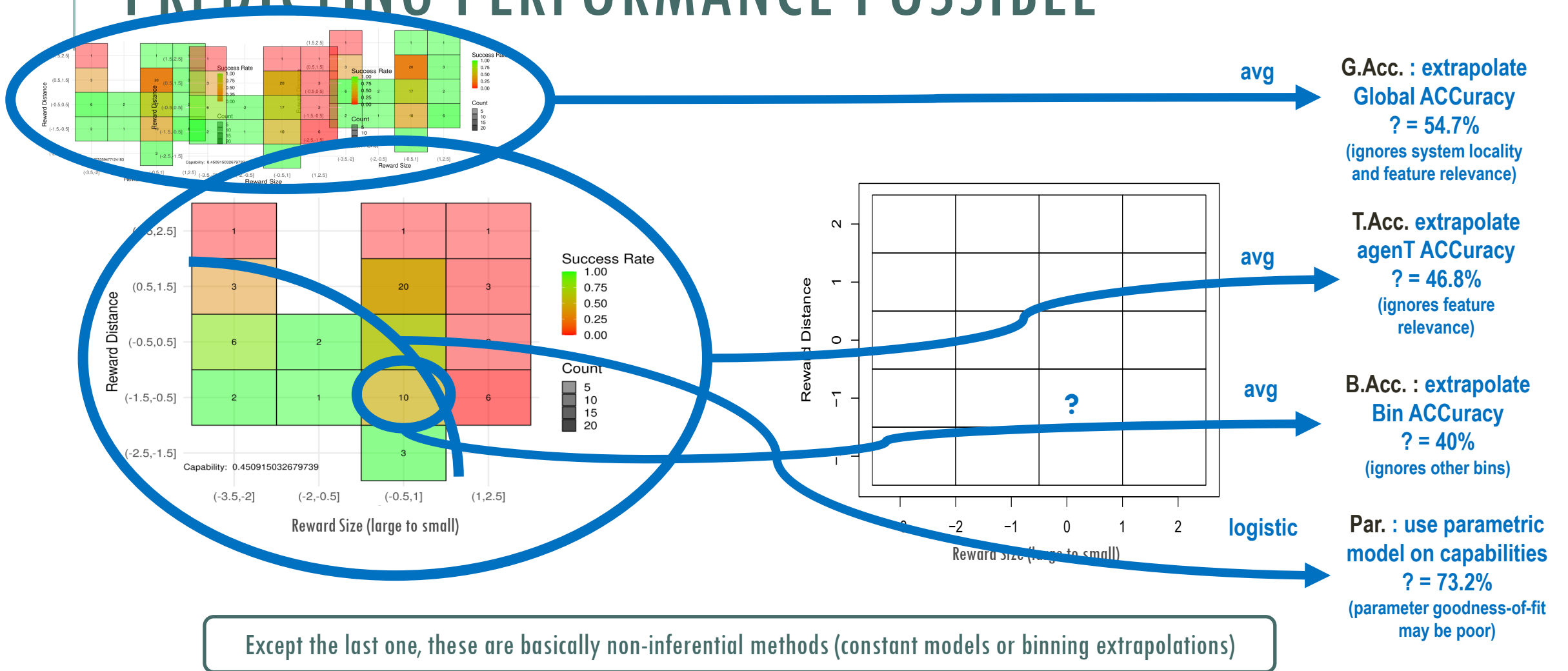
Conformant System
Juohmaru

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

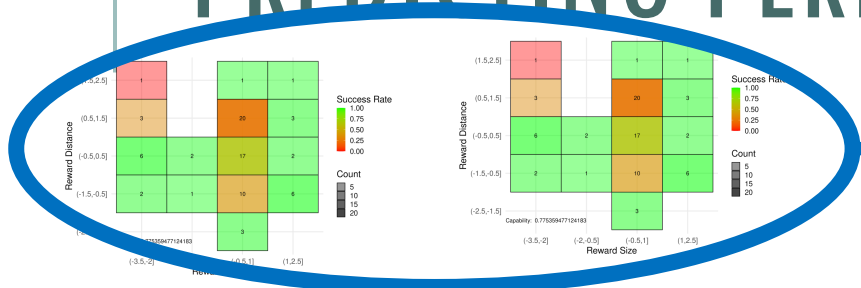


Non-Conformant System
y.yang

PREDICTING PERFORMANCE POSSIBLE

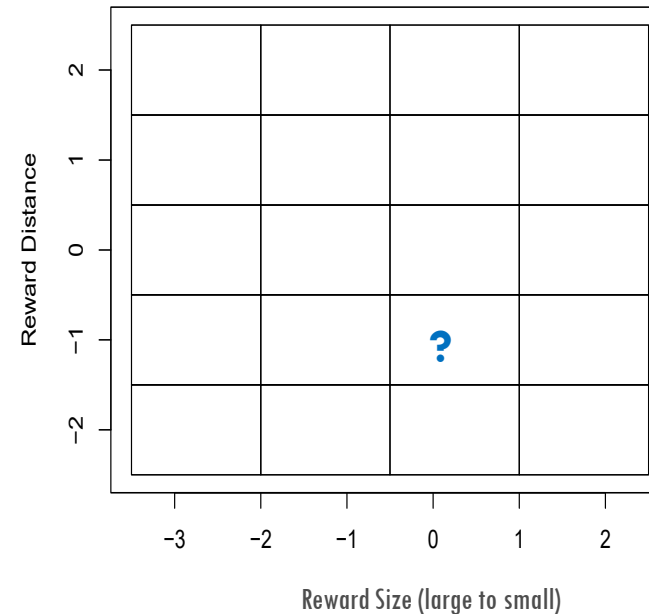
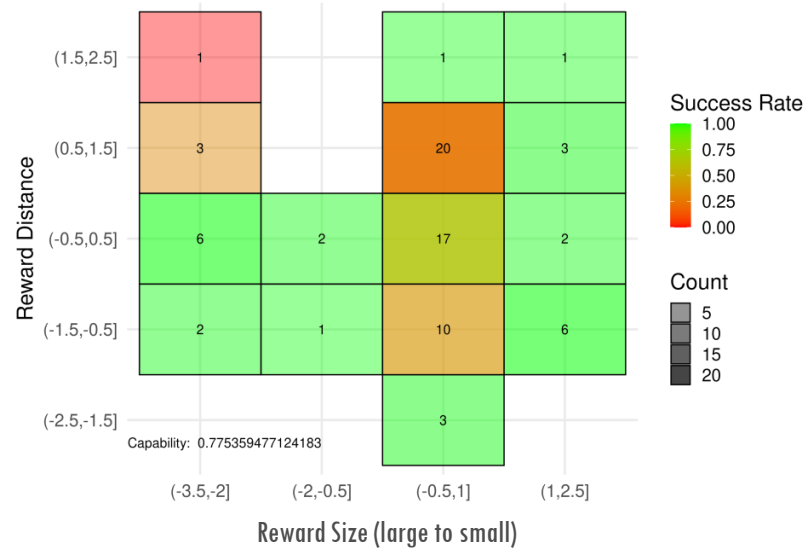


PREDICTING PERFORMANCE NOT POSSIBLE?



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)

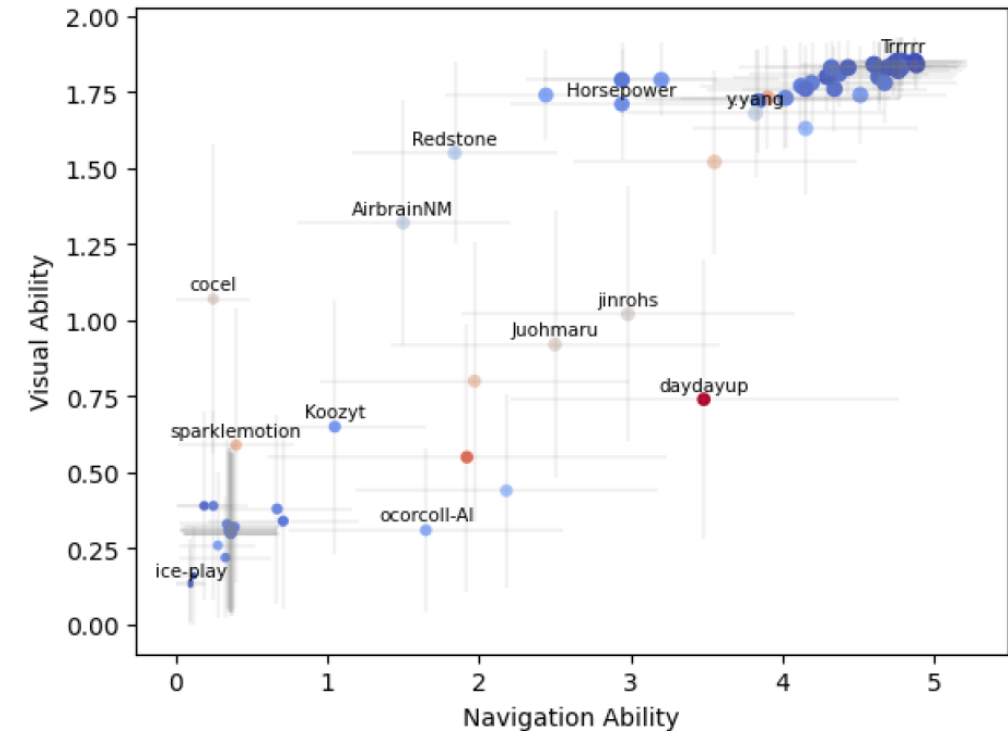
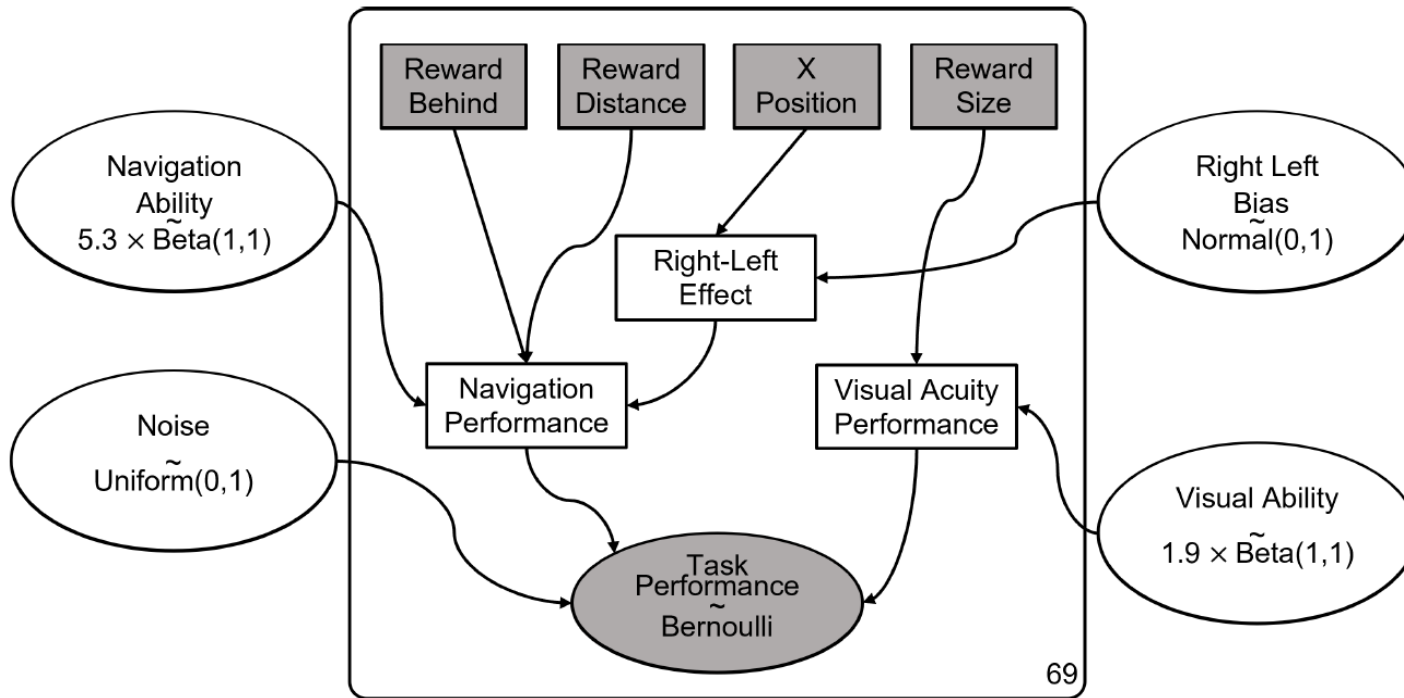
	Maj. (1)	G.Acc.	T.Acc.	~All+A	~Rel+A
Error	45.3%	48.0%	33.6%	19.7%	20.6%
MAE	45.3%	49.6%	34.9%	29.3%	30.2%
MSE	45.3%	24.8%	17.6%	14.8%	15.4%

Animal AI Competition Data: 99 instances x 68 agents

Predictable AI using Measurement Layouts

*J. Burden et al. “Inferring Capabilities from Task
Performance with Bayesian Triangulation”,
<https://arxiv.org/abs/2309.11975>.*

MEASUREMENT LAYOUTS : SIMPLE EXAMPLE (AAIO)



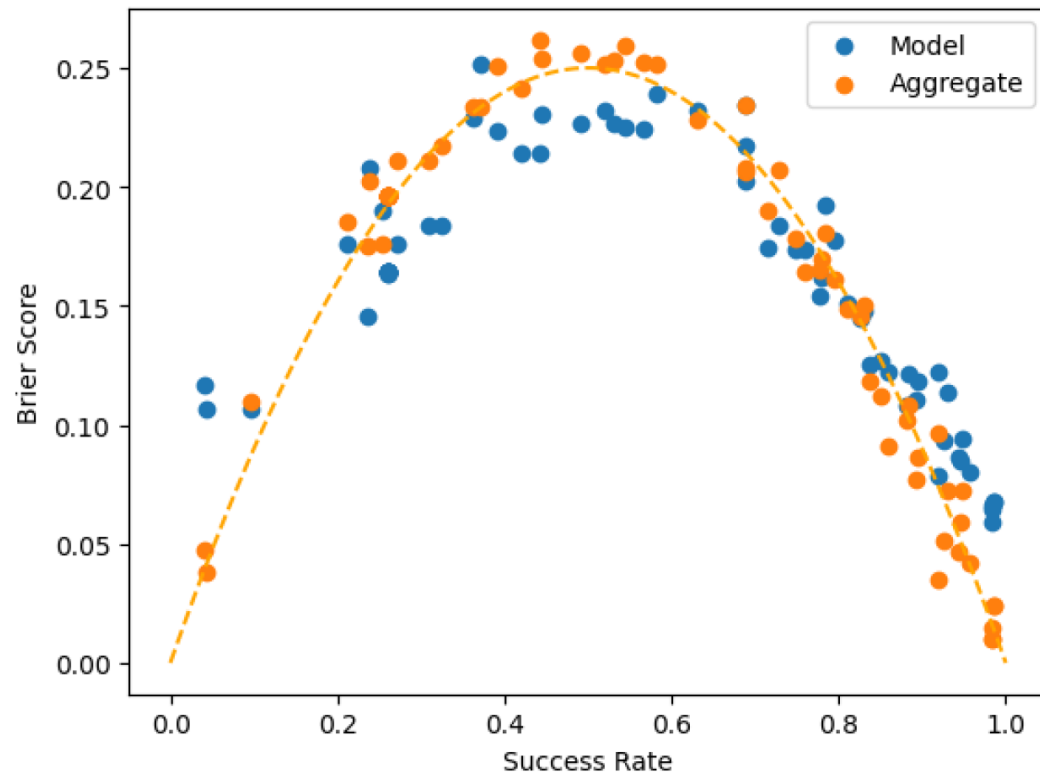
The x-axis and y-axis show the inferred means for navigationAbility and visualAbility respectively, with their standard deviations as error bars in grey. The radius of each point represents the average performance, while the colour represents the noiseLevel (red higher than blue).

MEASUREMENT LAYOUTS : MORE COMPLEX (0-PIAAGETS)

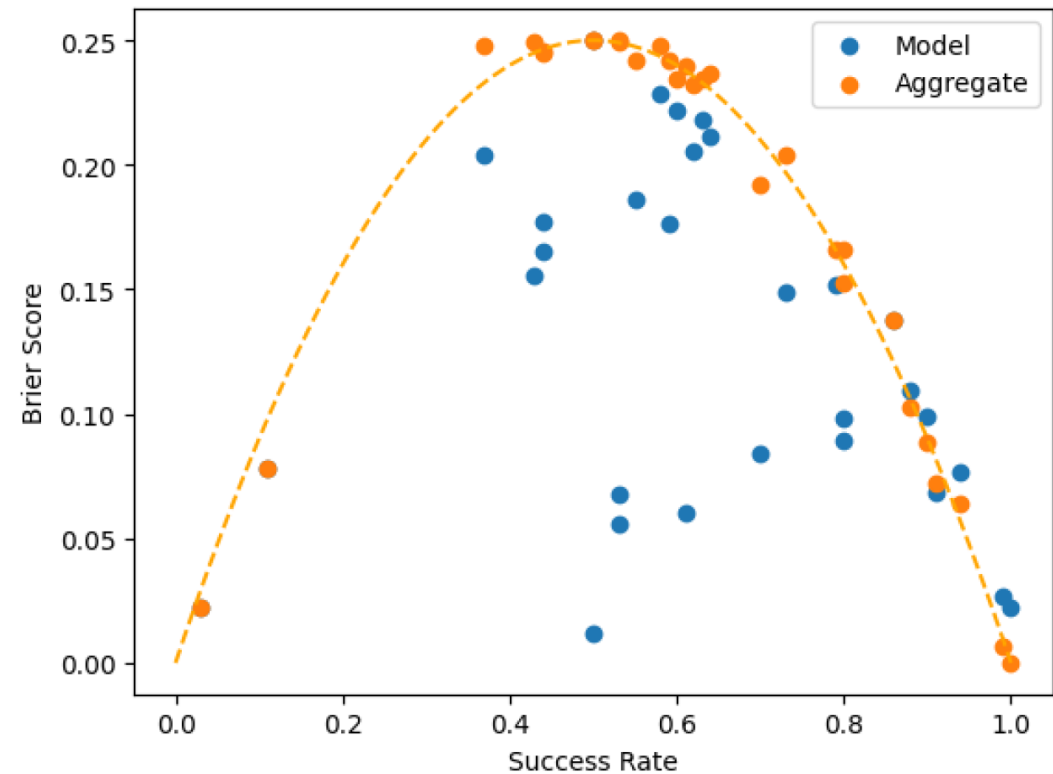
objPermAbility	50	13	13	12	50	50	50	50	48	49	49	49	49	50	50	50	28	36	23	15	25	2.1	33	29	50	27	50	49	12	12
flatNavAbility	56	26	7.6	13	56	56	56	56	56	56	56	52	51	56	56	56	56	56	43	38	37	32	45	43	56	56	49	42	52	53
visualAcuityAbility	6	2.9	2.9	4	6	6	6	6	6	5.9	3.6	5.9	5.8	6	6	6	6	6	5.9	5.8	6	5.3	6	5.8	6	5.8	6	6	6	5.9
lavaAbility	1	0.49	0.5	0.54	0.99	0.08	0.82	0.98	1	0.57	0.98	0.99	0.99	0.99	1	0.99	1	1	0.98	0.93	0.59	0.55	1	0.95	0.99	0.98	0.72	0.98	1	1
platformAbility	1	0.5	0.49	0.36	0.99	0.99	0.98	0.98	1	0.98	1	1	1	1	1	0.99	1	0.99	0.73	0.7	0.99	0.97	0	0.02	1	1	0.98	0.99	0.99	0.88
rampAbility	1	0.5	0.48	0.35	0.99	1	1	0.99	0.14	0.98	1	0.92	0.94	0.75	1	1	1	1	0.01	0.03	0.66	0.72	1	0.95	0.56	0.53	1	0.99	0.73	0.73
memoryAbility	4.8	2.4	2.4	2.3	4.8	4.8	4.8	4.8	4.7	4.7	4.7	4.7	4.7	4.7	4.8	4.8	2.7	4.8	4.4	2.8	4.5	2.7	4.8	4.4	4.7	4.5	4.8	4.7	4.8	4.7
rightLeftBias	0	-0.01	0.05	-0.12	-0.25	-0.14	0	0.02	0.33	0.13	0.33	6.3	-6.3	0.13	0	-0.24	-0	0.02	-0.01	-0.13	0.01	0.05	0	0.07	0.01	0.46	0.02	0.29	-0.08	0.03
noisePar	0	0.98	0.11	0.03	0.03	0.02	0	0.01	0	0.21	0.01	0.35	0.42	0.04	0	0	0.01	0.01	0.34	0.42	0.31	0.39	0	0.25	0.05	0.51	0.06	0.39	0	0.25
Success	1	0.5	0.1	0.03	0.9	0.53	0.85	0.88	0.61	0.63	0.59	0.63	0.61	0.79	0.99	0.94	0.79	0.91	0.53	0.44	0.44	0.38	0.51	0.44	0.73	0.57	0.71	0.56	0.8	0.62
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30

Very similar performance, very different cognitive profiles

MEASUREMENT LAYOUTS : PREDICTABILITY



(a) AAI0 Tasks



(b) O-PIAAGETS tasks.

Predictable AI from Assessors

JH Orallo, W Schellaert, FM Plumed

Training on the Test Set: Mapping the System-Problem Space in AI

AAAI 2022

DEFINITION


Conditional probability estimator of the result r for AI system π on situation μ :

$$\hat{R}(r|\pi, \mu) \approx \Pr(R(\pi, \mu) = r)$$

It is trained (and evaluated) on test data:

- Using a distribution of situations (instances) μ .
- Using a distribution of systems π .

It is applied during deployment, before π does any inference or even starts.



π	μ	r
Resnet, $\theta_1, \theta_2, \dots$	Image3, x_1, x_2, \dots	1
Resnet, $\theta_1, \theta_2, \dots$	Image23, x_1, x_2, \dots	0
...
Inception, $\theta_1, \theta_2, \dots$	Image3, x_1, x_2, \dots	1
Inception, $\theta_1, \theta_2, \dots$	Image78, x_1, x_2, \dots	1
...

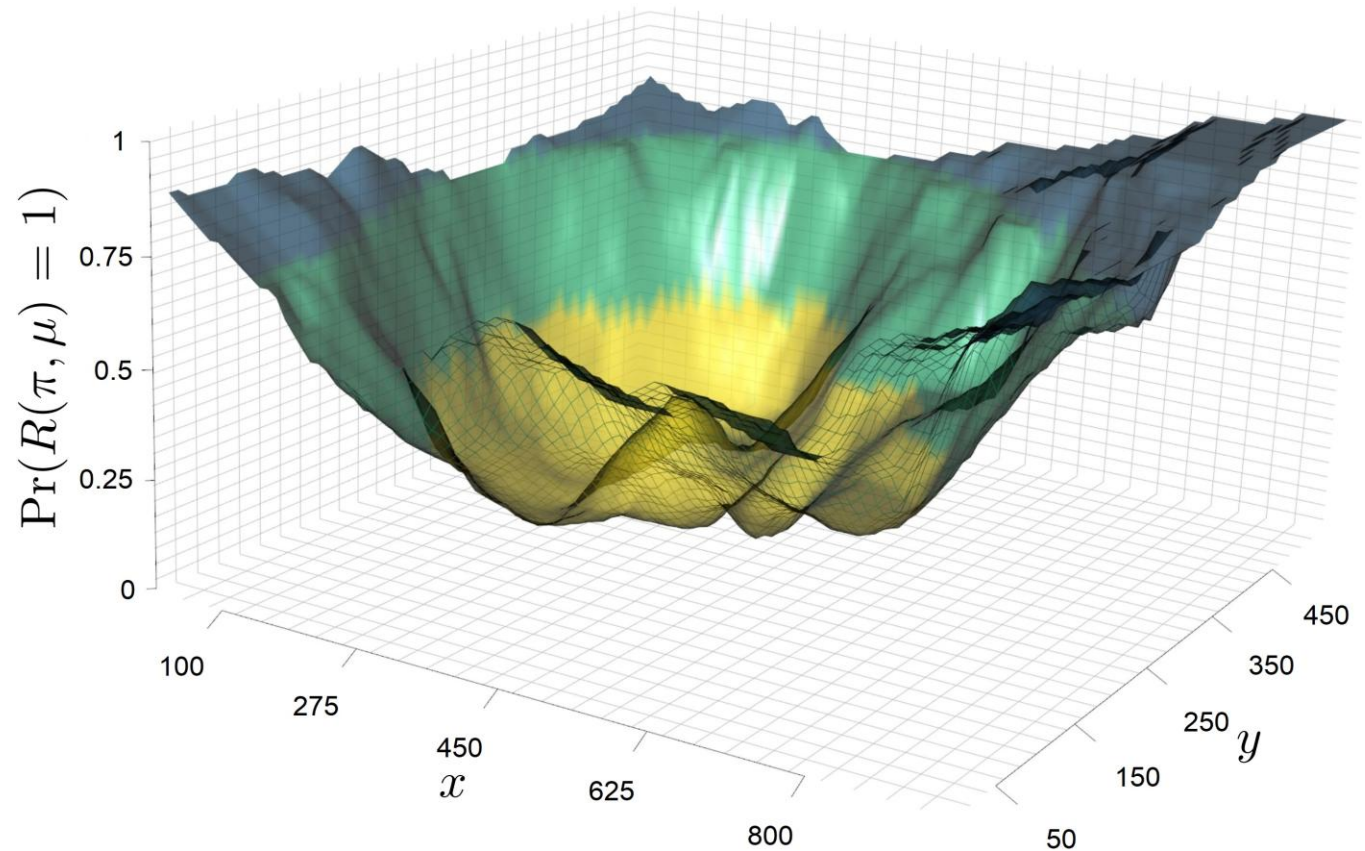
PROBLEM SPACE

We can describe situations or instances with properties $\mu = \langle \chi_1, \chi_2, \dots \rangle$.

- Delivery robot in a city with destination $\mu = \langle x, y \rangle$
- π behaves very differently depending on the situation μ .
- Expected result for π differs for different joint distributions $\Pr(x, y)$



Downtown Vancouver



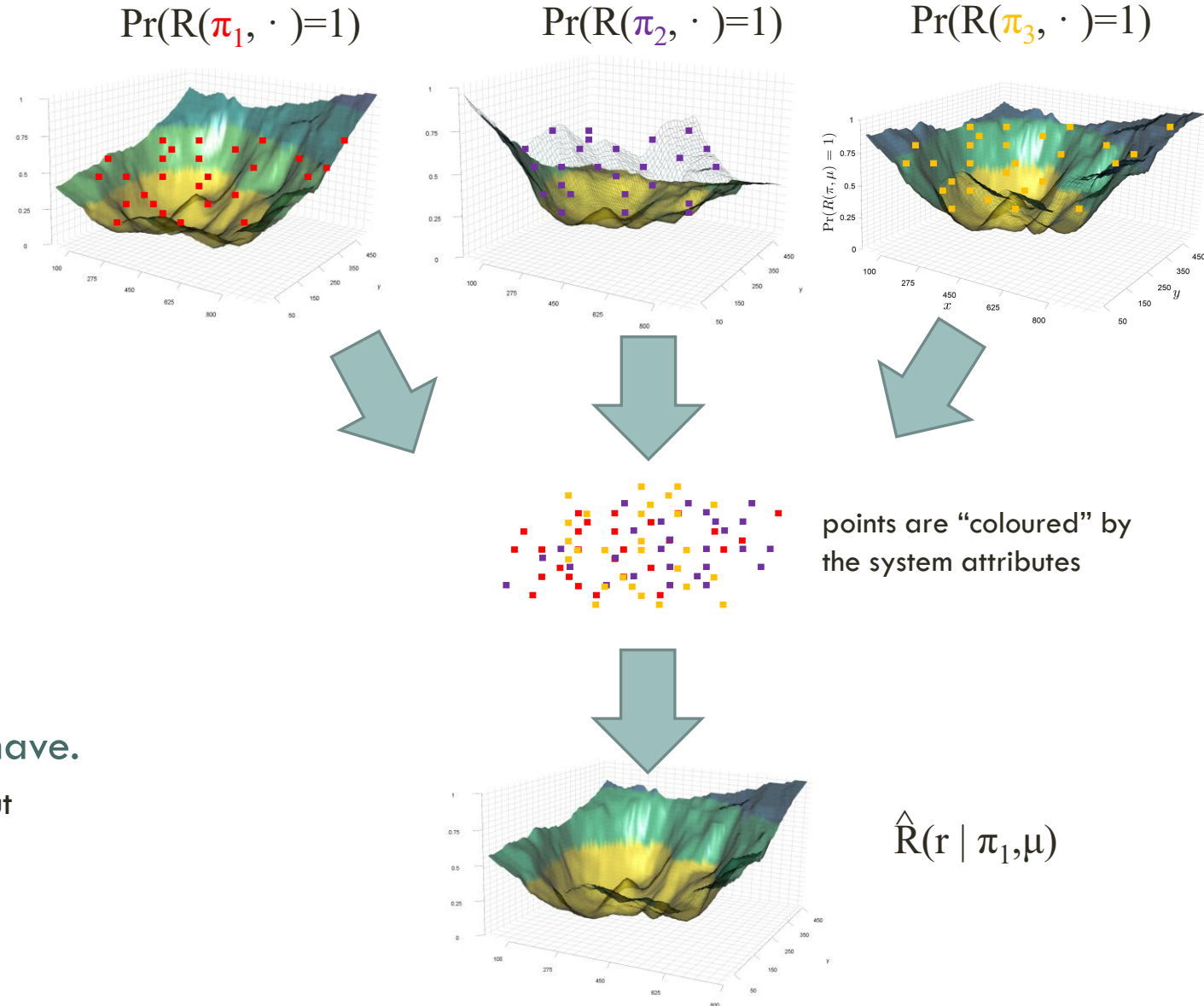
SYSTEM SPACE

We can describe systems with properties $\pi = \langle \theta_1, \theta_2, \dots \rangle$.

- Hyperparameters, system's operating conditions (e.g., computing resources), developmental states, ...

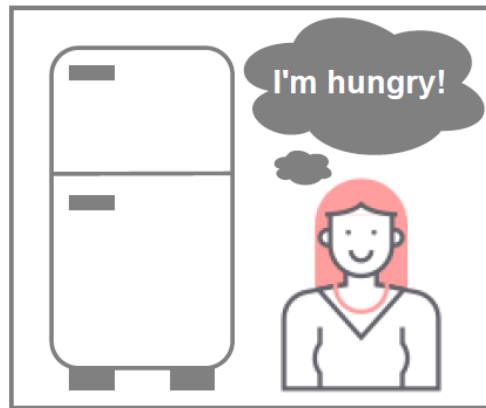
Key element for an assessor

- Much predictability about one π can be obtained by looking at how other π' behave.
 - Uncertainty estimation or calibration of π without looking at other systems is shortsighted!



Predicting “Failure” Is Not Enough

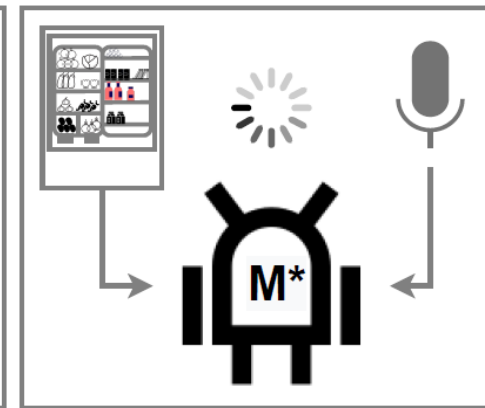
WHERE'S THE GROUND TRUTH?



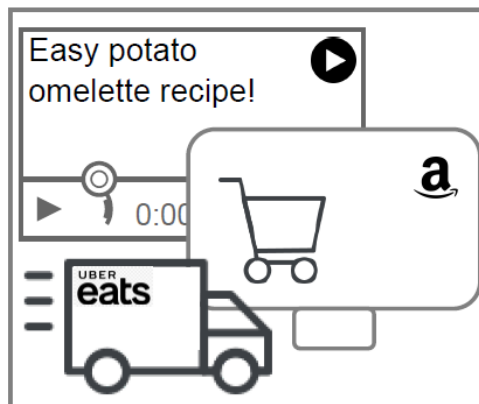
step 1



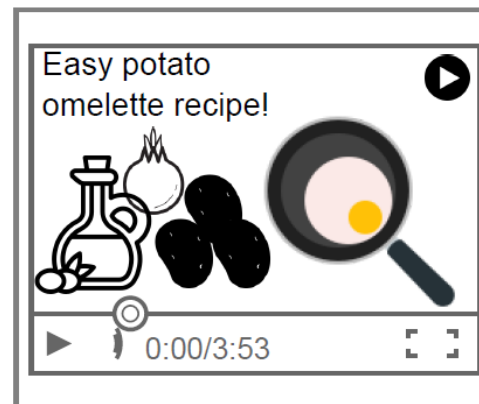
step 2



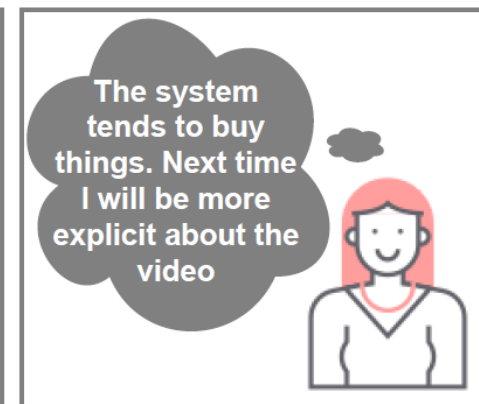
step 3



step 4



step 5



step 6

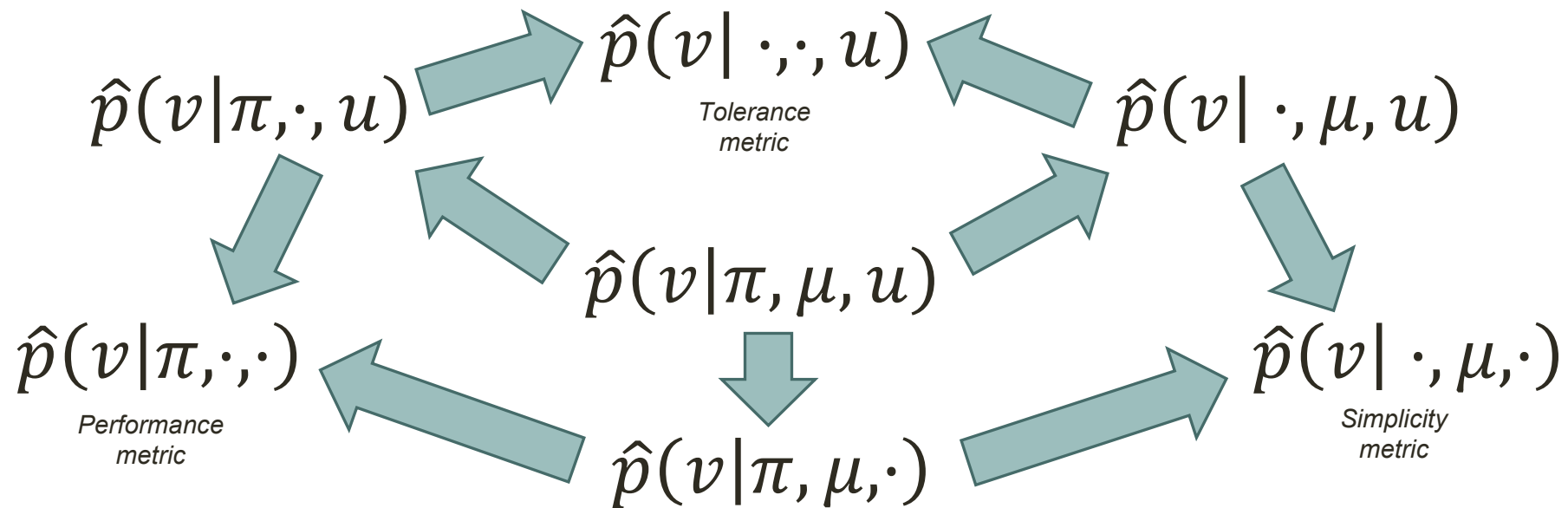
Schellaert, Plumed, Vold, Burden, Casares, Loe, Reichart, OhEigartaigh, Korhonen, Orallo "Your Prompt is My Command: Assessing the Human-Centred Generality of Multi-Modal Models". JAIR 2023,

WHOM TO PREDICT FOR?

There's usually no single ground truth for all users

- Predict result for a particular user u :

- π : AI system
- μ : situation/problem
- u : user



WHAT TO PREDICT? MODEL ITS ~~ALIGNMENT~~

validity

DALL-E 2 generated

There are many validity components (a HCI view)

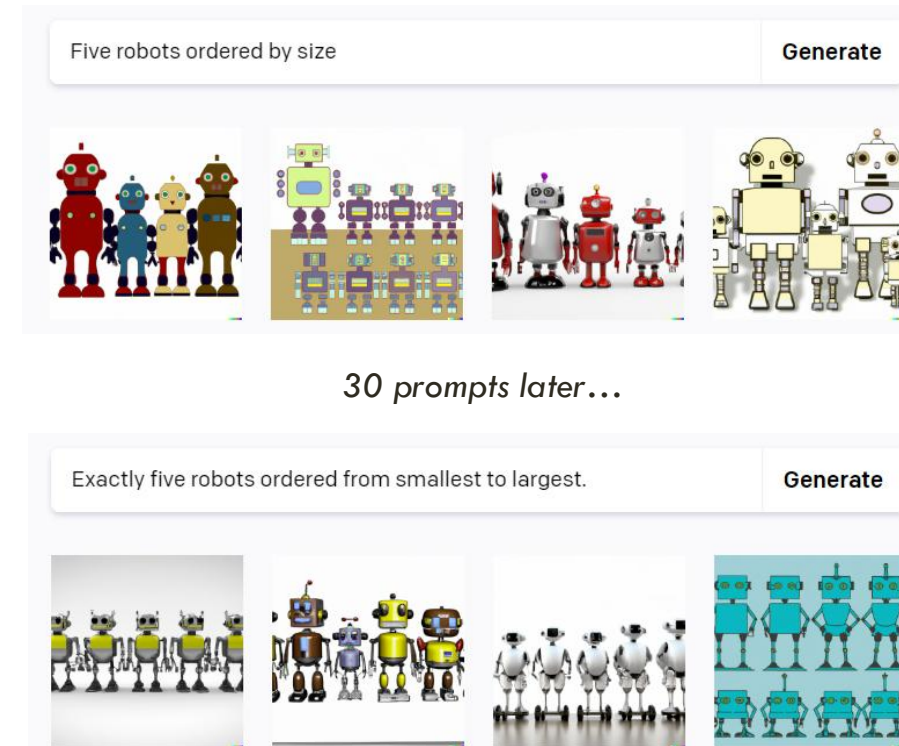
- User effort to give right instructions to the system
- User effort to extract answers or validate system's behaviour
- Quality of the solution
- Possible dangerous outcome
- Operational cost (energy, time, ...)

Capabilities:

each component can be captured by one or more capabilities

Assessors:

different assessors for each kind of component to predict



30 prompts later...

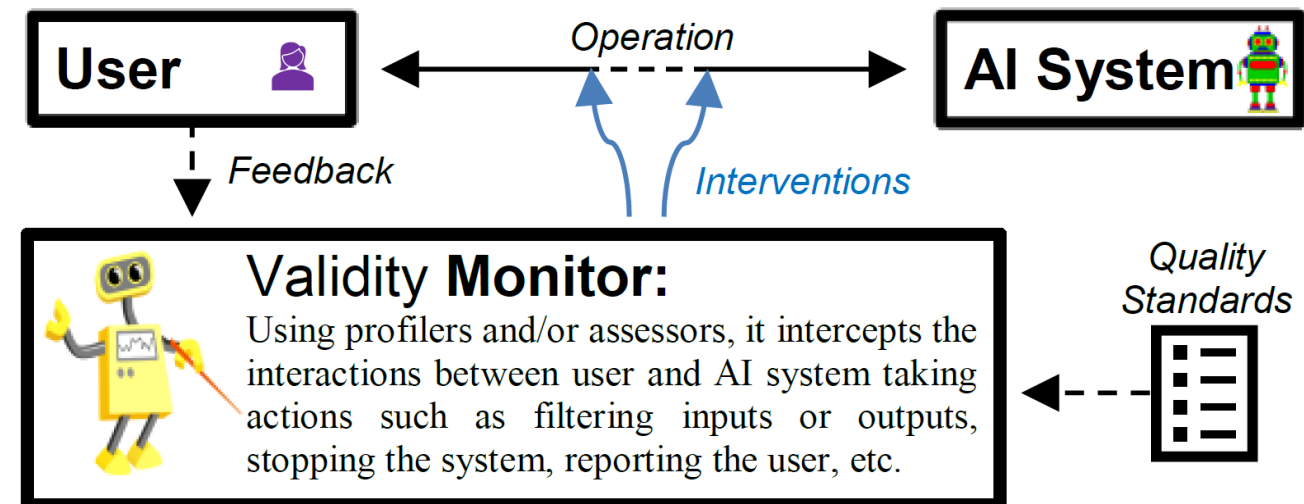
The Road Ahead

OPPORTUNITIES

Is human-in-the-loop oversight a good idea for GPAI?

- A future world surrounded by AI
- Hard to oversee increasingly smarter systems
- Feedback loops can be tricked

Scalable
Oversight?



NEEDS

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.

ARTIFICIAL INTELLIGENCE

Rethink reporting of evaluation results in AI

Aggregate metrics and lack of access to results limit understanding

By Ryan Burnell¹, Wout Schellaert², John Burden^{1,3}, Tomer D. Ullman⁴, Fernando Martinez-Plumed², Joshua B. Tenenbaum⁵, Danaja Rutar⁴, Lucy G. Cheke^{1,6}, Jascha Sohl-Dickstein⁷, Melanie Mitchell⁸, Douwe Kiela⁹, Murray Shanahan^{10,11}, Ellen M. Voorhees¹², Anthony G. Cohn^{13,14,15,16}, Joel Z. Leibo¹⁰, Jose Hernandez-Oralo^{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 policy-makers 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-

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

ASSURANCE

Three A's of Assurance:

- Norm Adherence: Model Cards
- Independent Audit: Ethical Black Boxes
- Prospective Assessment: Ethical Overseeing Monitors

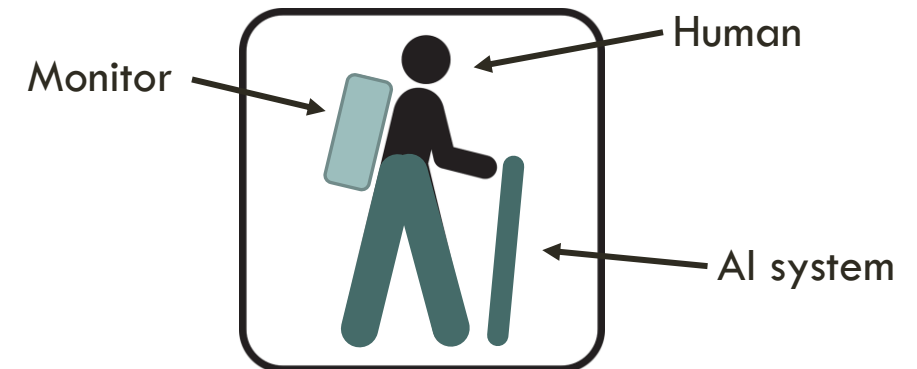
Falco, G., Shneiderman, B., Badger, J., Carrier, R., Dahbura, A., Danks, D., ... & Winfield, A., Yeong, Z. K. (2021). Governing AI safety through independent audits. *Nature Machine Intelligence*, 3(7), 566-571.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. *Conf. on fairness, accountability, and transparency*.

Winfield, A. F., & Jirotko, M. (2017). The case for an ethical black box. In *Conf. Towards Autonomous Robotic Systems*

VISION:

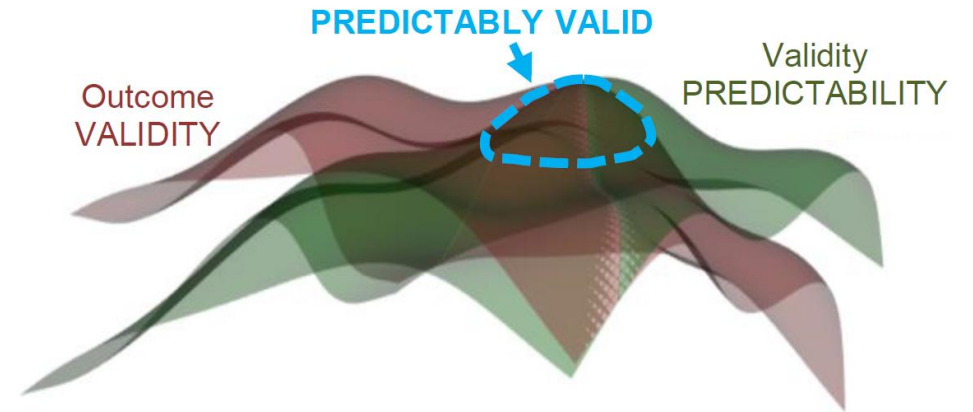
Having every deployed AI system backed by and accounted for with its capability profile and/or its assessor model



Take-aways

SUMMARY

Many present and future problems associated with artificial intelligence are not due to their validity, but to our poor assessment of its validity.



More dramatic with GPAI. Fully Verify, Align, Explain, Interpret? \Rightarrow PREDICT instead

- Anticipate the user-aligned system validity : operating condition \langle system, instance, user context \rangle
- Choose the validity properties to be modelled: correctness, safety, toxicity, etc.
- Data-based approaches:
 - **Measurement layouts:** Identify capabilities and problem demands, and user preferences.
 - **Assessors:** use as much data as you can get from many systems, instances and users.

THANK YOU!

JOSE H. ORALLO

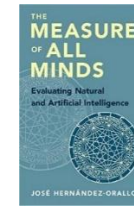
<http://josephorallo.webs.upv.es/>
jorallo@upv.es

Other Talks (<http://josephorallo.webs.upv.es/>)

- Diversity Unites Intelligence: Measuring Generality
- Measuring A(G)I Right: Some Theoretical and Practical Considerations
- Natural and Artificial Intelligence: Measures, Maps and Taxonomies

Book (<http://allminds.org>):

- “The Measure of All Minds: Evaluating Natural and Artificial Intelligence”, Cambridge U.P. <http://allminds.org>



The AI Collaboratory: <https://ai-collaboratory.jrc.ec.europa.eu/> (old: <http://dmip.webs.upv.es/AICollaboratory/>)

- Part of the European Commission's AI watch: https://ec.europa.eu/knowledge4policy/ai-watch_en
- Technology Readiness Levels: <https://data.europa.eu/doi/10.2760/495140>
- Measuring the Occupational Impact of AI: <https://jair.org/index.php/jair/article/view/12647>



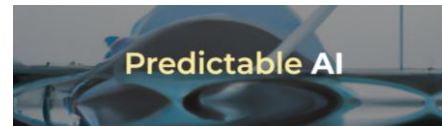
OECD's AI and the Future of Skills Project:

- <https://www.oecd.org/education/ceri/Future-of-Skills-Overview.pdf>, <https://doi.org/10.1787/5ee71f34-en>.



PREDICTABLE AI:

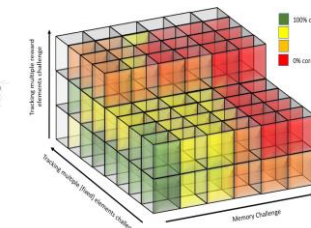
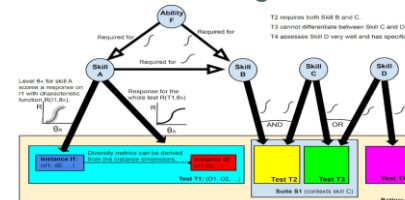
- <https://www.predictable-ai.org/>.



DARPA RECoG-AI Project: <http://lcfi.ac.uk/projects/kinds-of-intelligence/recog-ai/>

- Part of the Kinds of Intelligence Programme at the CFI in Cambridge

- <http://lcfi.ac.uk/projects/kinds-of-intelligence>



Extra slides...

PROBLEMS OF QUANTIFYING PERFORMANCE

- Patterns of performance missed if items are not characterised
 - No identification of features that lead to failure
- Poor estimation of performance for new distributions
 - The metric cannot be extrapolated
- Poor granular estimation for the same distribution!
 - Likely to be conditions under which the system performs better or worse

PROBLEMS OF (POST-HOC) EXPLAINABLE AI

- Generally very useful for “experts” (e.g., AI developers, domain experts)
- For “mortal” users it can be used for justification rather than revision
 - False sense of understanding
 - “Are you satisfied with the explanation?” vs ~~“Can you predict what the system does precisely?”~~ *Impossible?*
 - Dangerous calibration of trust
 - Initial sceptical (and prudent) stance towards an AI system turns into **overconfidence**.
- It may come too late for users
 - Can we anticipate and prevent failures?

We trust what is **predictable**
Changing the analysis of “**did**” fail to “**will**” fail.

CHALLENGES

