Capability-Oriented Evaluation (of AI)

Jose Hernandez-Orallo

Mostly based on

Ryan Burnell, John Burden, Danaja Rutar, Konstantinos Voudouris, Lucy Cheke, Jose Hernandez-Orallo

“Not a Number: Identifying Instance Features for Capability-Oriented Evaluation“ IJCAI 2022

UNIVERSITAT POLITÈCNICA DE VALÈNCIA
VRAIN
CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

Do we have capability-oriented evaluation in AI?

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

Alphabetic author list:*

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima (Shammie) Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, Ziyi Wu

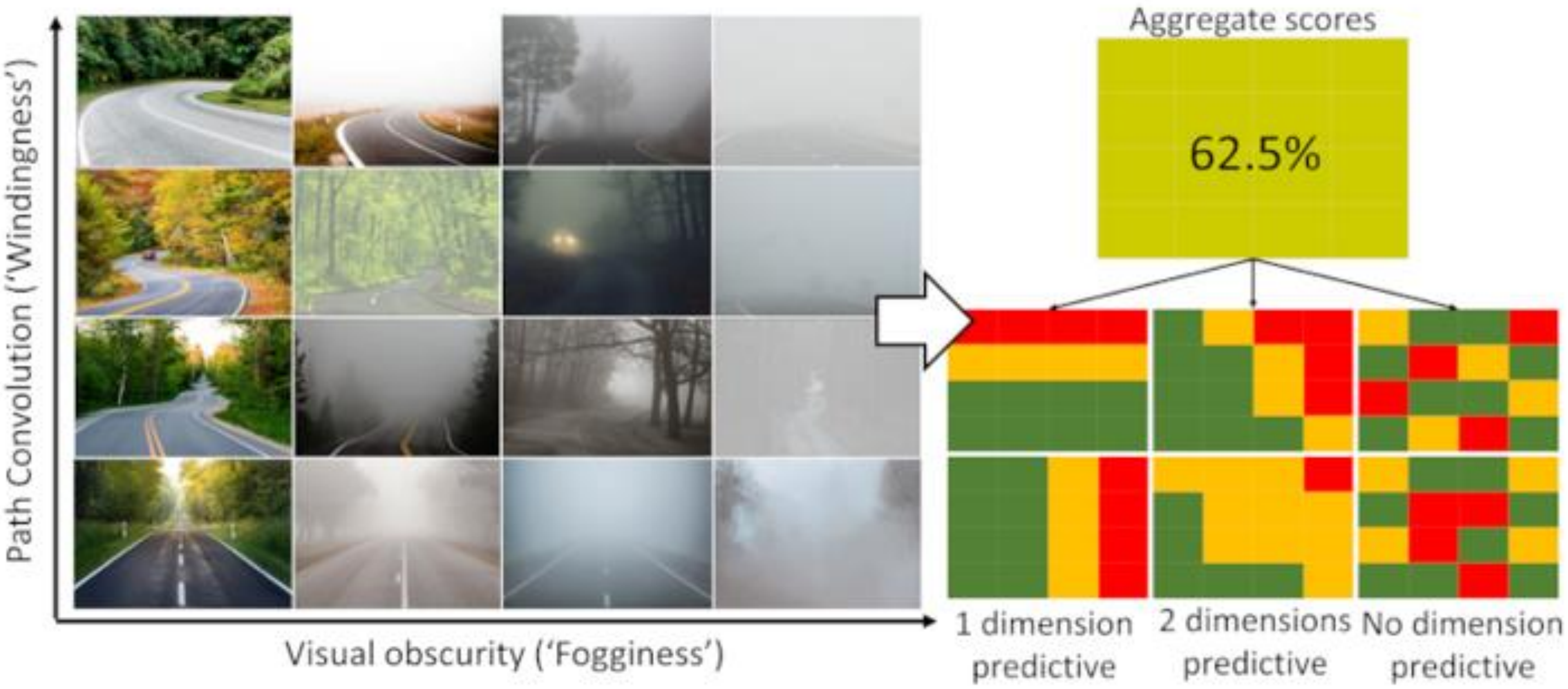
Performance-oriented vs Capability-oriented

- Performance is a property (a measure) of **a pair <system, item>**:
  - Examples:
    - Correct prediction of MySpamFilter on Email735
    - 85% accuracy of ResNet23 on ImageNet
  - **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.

The problems of aggregated performance

- No patterns of performance
  - No identification of failure points
- 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

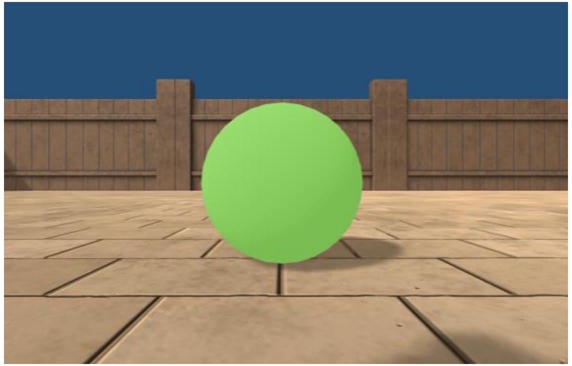
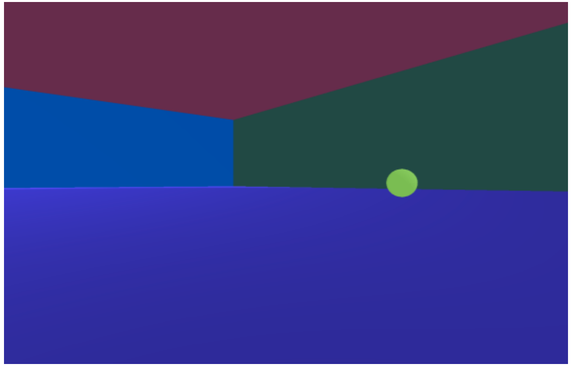
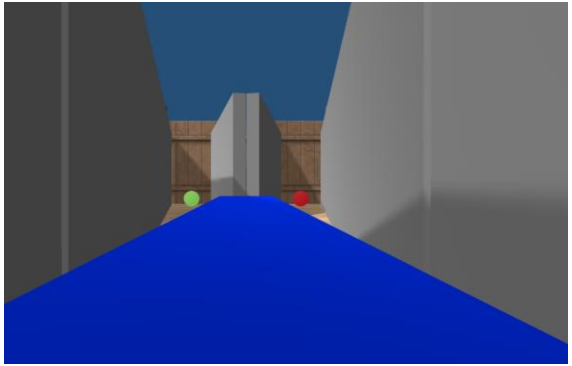
Where does my system fail?



Estimate success/failure granularly

Proof of concept: Animal AI Olympics

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

http://animalaiolympics.com/AAI/

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

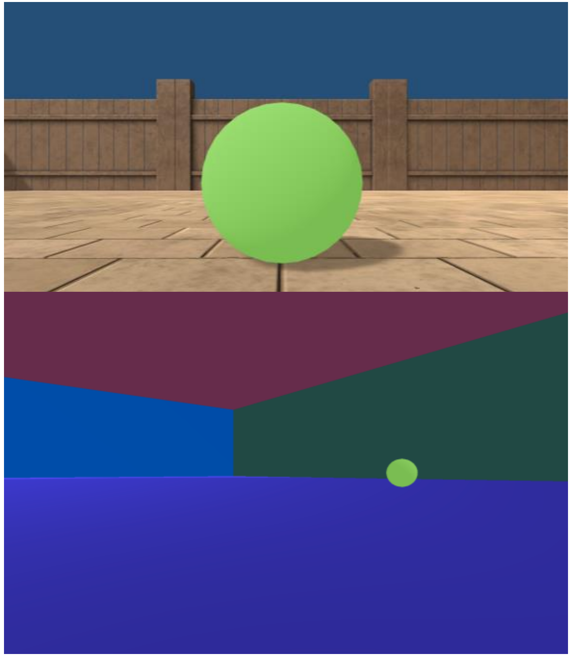
Identifying features of interest

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)



Identified dimensions and agent characteristic curves

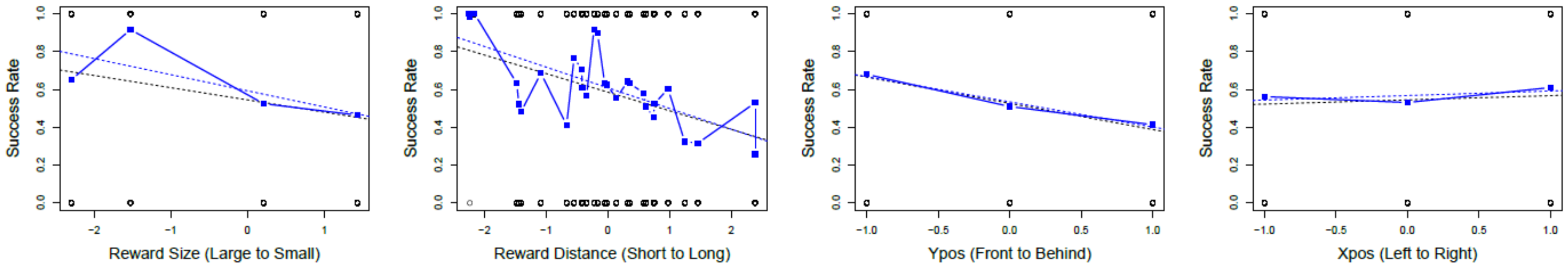
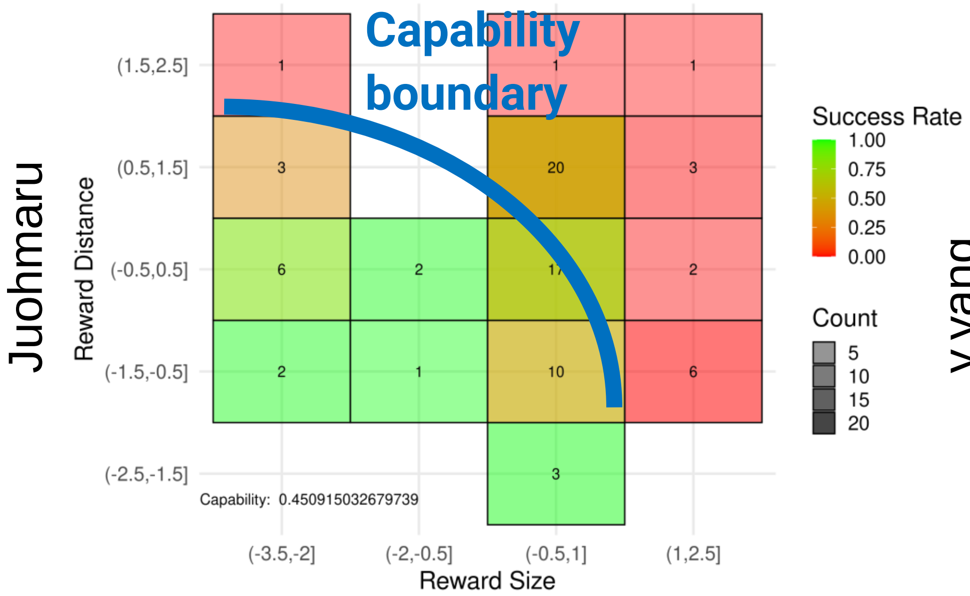


Figure 5: Characteristic curves of all competition entrants (agents) according to three relevant features (size, distance and Ypos) and one irrelevant feature (Xpos). Black dashed lines show the linear regression for the black points (pass/fail), while blue dashed lines interpolate the blue points (binned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

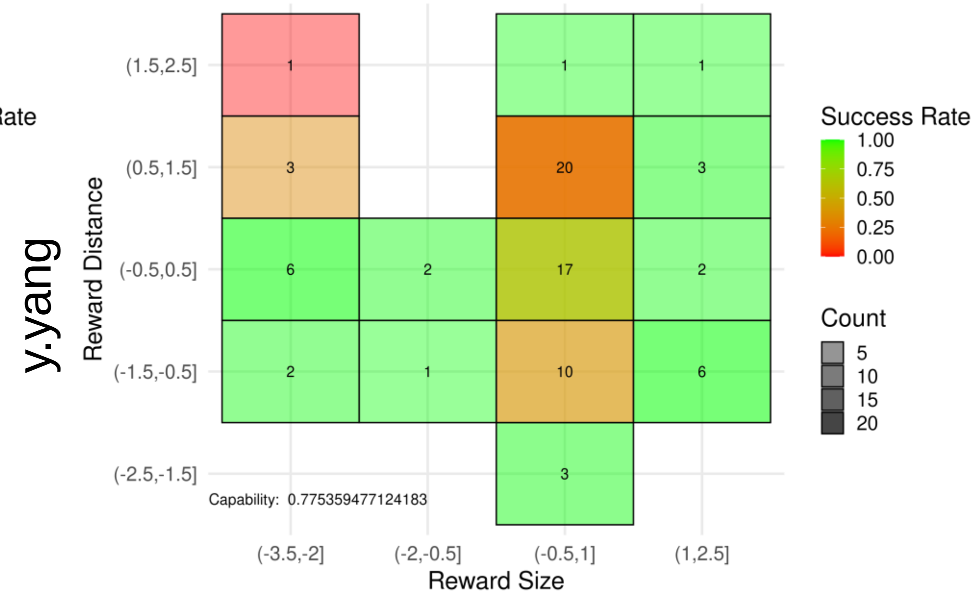
Visualising performance distribution

- Plot a subset of relevant variables:

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

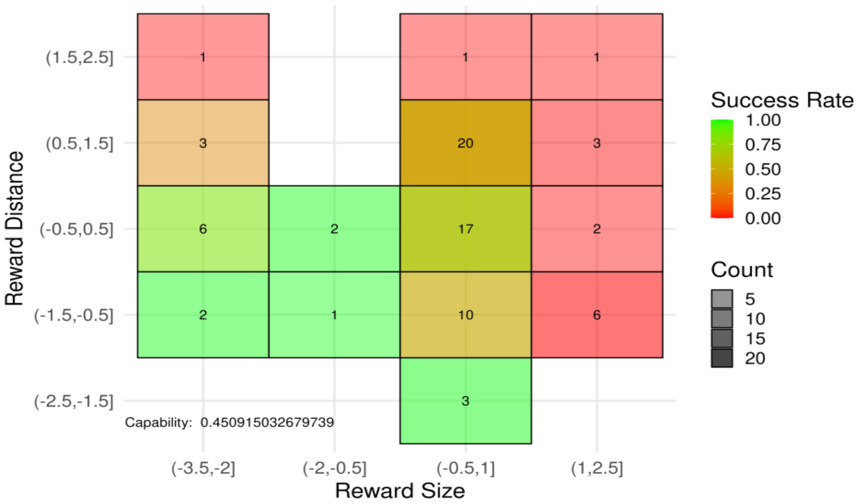


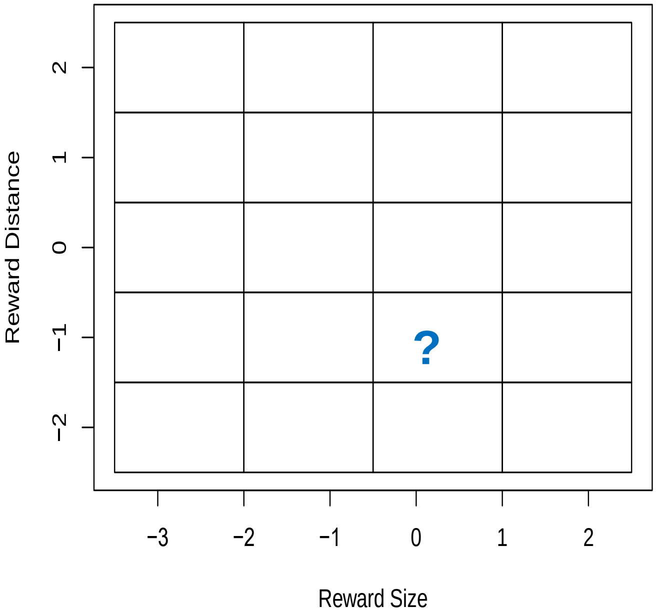
Conformant System



Non-Conformant System

Predicting performance





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

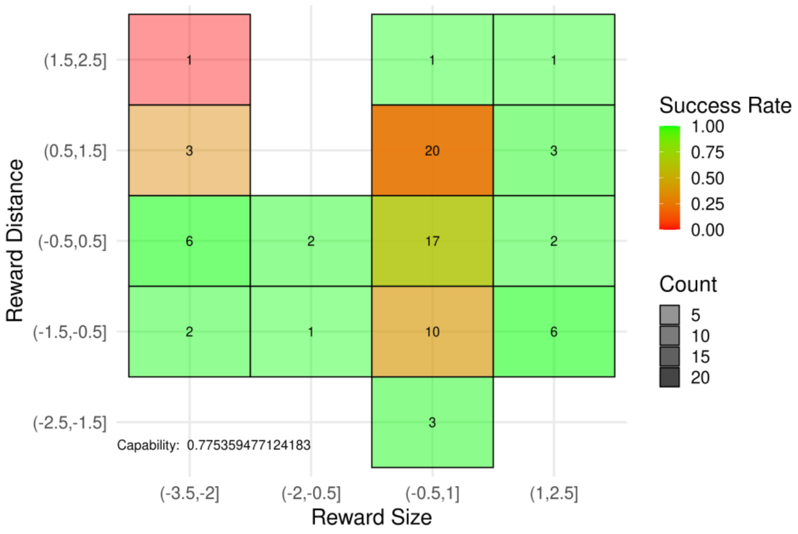
extrapolate Global ACCuracy
? = 54.7%
(ignores system locality and feature relevance)

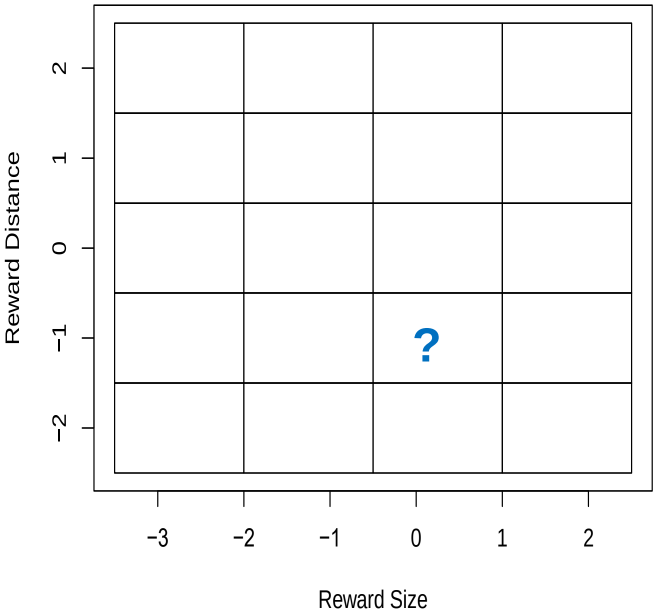
extrapolate agenT ACCuracy
? = 46.8%
(ignores feature relevance)

extrapolate bin
? = 40%
(ignores other bins)

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

Predicting performance (parametric model won't fit)





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!!

Assessors (non-parametric models)

Hernández-Orallo, J.; Schellaert, W.; Martínez Plumed "Training on the Test Set: Mapping the System-Problem Space in AI", AAAI 2022 (Blue Sky Ideas Award).

- 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.

π	μ	r
Resnet, θ_1, θ_2, …	Image3, χ_1, χ_2, …	1
Resnet, θ_1, θ_2, …	Image23, χ_1, χ_2, …	0
…	…	...
Inception, θ_1, θ_2, …	Image3, χ_1, χ_2, …	1
Inception, θ_1, θ_2, …	Image78, χ_1, χ_2, …	1
…	…	...

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

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

Guidelines for Capability-based Evaluation

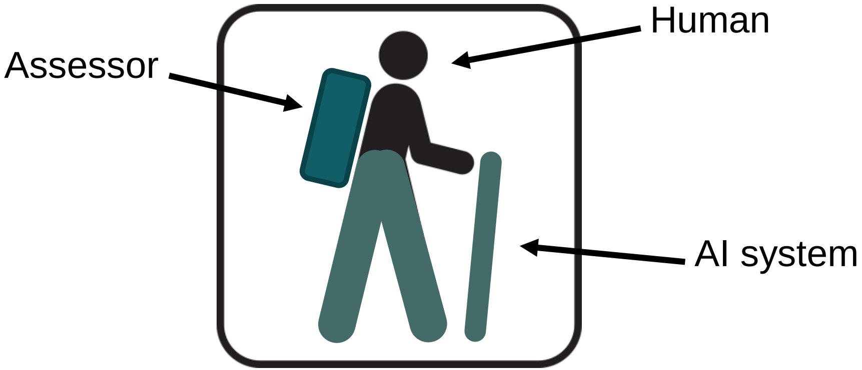
1. Choose a domain, task or benchmark with instance-level data.
2. Identify features that can be extracted or easily annotated for each instance during testing.
3. Identify which features are relevant (should affect performance) and those that are irrelevant (should not affect performance) based on theory and domain knowledge.
4. Analyse the relationships between features and performance using correlation and other exploratory analyses.
5. Select features of interest, bin them appropriately and build characteristic grids (both global and for individual systems) to evaluate patterns of performance.
6. Build predictive models using extracted features. Com pare the results with other ways of predicting performance, such as extrapolating average metrics.
7. Using characteristic grids and predictive models, evaluate capabilities of each system across the distributions of the dimensions of interest.
8. Identify areas of competence for individual systems so that these can later be used for testing more complex or advanced capabilities and skills (where appropriate).
9. Use areas of weakness to inform changes to the benchmark, system models or training.
10. With new insights about which features are relevant, iterate the process to step 2—or to step 1 if more test data is needed (using the models)—for subsequent analyses.

Vision : map the system – problem space

Lexin Zhou, Fernando Martínez-Plumed, José Hernández-Orallo Cèsar Ferri and Wout Schellaert "Reject Before You Run: Small Assessors Anticipate Big Language Models" EBeM@IJCAI2022

- Identify dimensions in systems (capabilities) and problems (difficulties):
  - The assessor is a simple parametric model.
- Otherwise, use non-parametric assessors.

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



Thank you!

JOSE H. ORALLO
http://josephorallo.webs.upv.es/
jorallo@upv.es

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

VRAIN

CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE