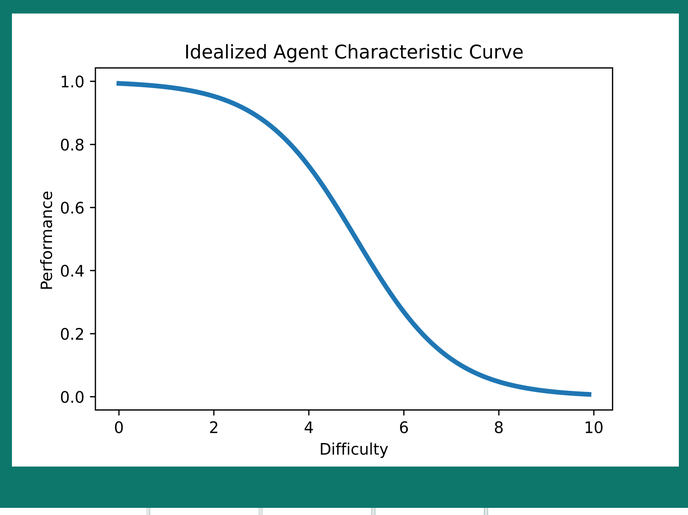
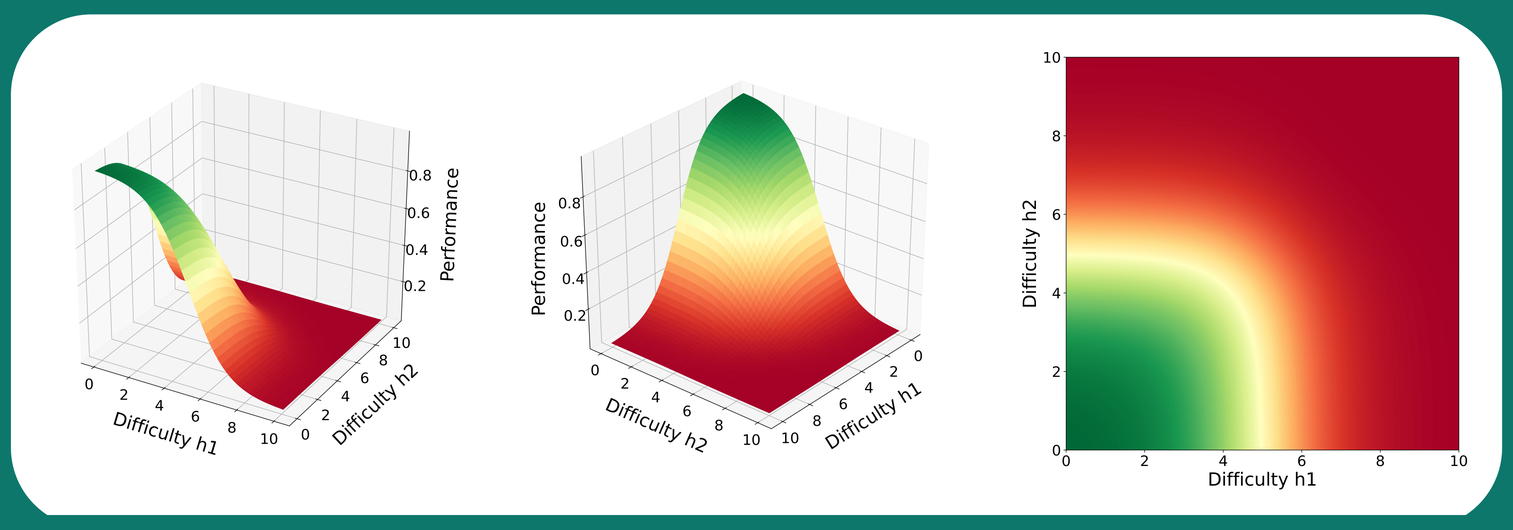
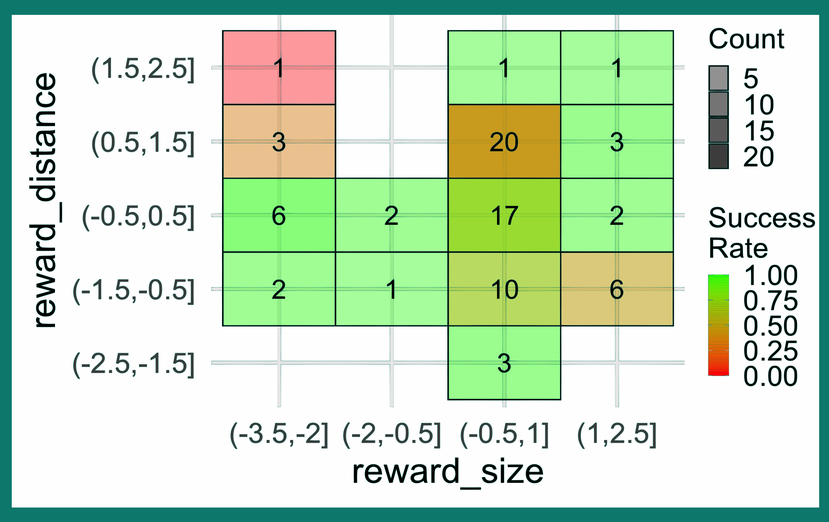
AI Evaluation can be improved with high dimensional Agent Characteristic Grids

John Burden
Centre for the Future of Intelligence

Psychometric inspired approaches have been used in AI Evaluation: Agent Characteristic Curves and Item Response Theory [1] can give a better understanding of a system's capabilities [2]. Our previous work introduced Agent Characteristic Grids for visualising a system's performance in 2 or 3 dimensional problem-space [3].







Standard IRT evalaution requires ordering all instances by difficulty.

Many complex tasks are inherently high dimensionnal and different features contribute to instance difficulty. Creating such an ordering is often infeasible for interesting problems.

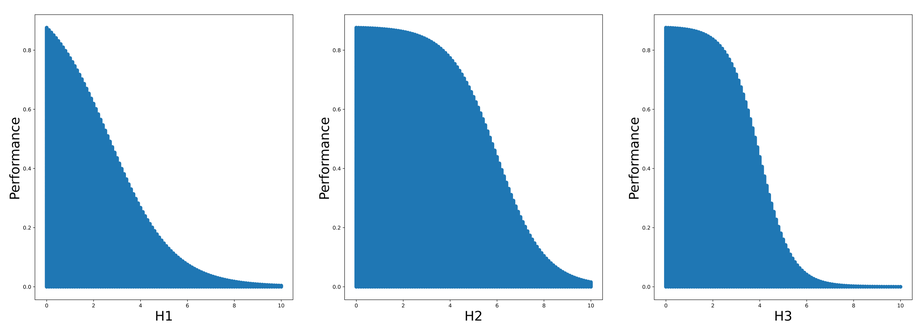
The solution: Embrace high dimensionality!

We can extend IRT into higher dimensions by modelling agent performance as the product of sigmoids.

$$\psi(x) = \prod_{i=0}^{n} \frac{1}{1 + e^{-c_i + \alpha_i x_i}}$$

In high dimensions visualisation can become difficult, but we can project the function to compare performance against a single element of difficulty

E.g., here we take $\alpha = (0.75, \vec{1}, 1.5), c = (-2, \vec{-6}, -6)$



We can further extend existing metrics of system capability, generality, as well as introducing a new concept of intra-dimensional generality.

In IRT, capability is the area under the agent's characteristic curve. In higher dimensions this extends to the hyper-volume bounded by the agent's performance curve and the origin:

$$\Psi = \int_{h_n \in H_n} \dots \int_{h_1 \in H_1} \psi(h_1, \dots, h_n) dh_1 \dots dh_n$$

Generality is an origin-compactness measure of the function. Defined in high dimensions as:

$$\Gamma = \frac{1}{\sqrt[n+1]{(2^n \prod_{i=1}^{n} \mathbb{H}^{H_i} - \Psi)\Psi^n}}$$

Where:

$$M_A^{H/H_i} = \int_{h_n \in H_n} \dots \int_{h_1 \in H_1} h_i \psi(h_1, \dots, h_n) dh_1 \dots dh_n$$

$$\mathbb{H}^{H_i} = \frac{M_A^{H/H_i}}{\Psi}$$

We also introduce the idea of dimension-oriented capability and generality: Exclude the desired dimension in the integrals above, and the result is a collection of functions showing how agent capability and generality change over the desired dimension. Exploration of these dimension-oriented functions and their utility remains as future work.

References
1]S. E. Embretson and S. P. Reise. Item response theory for psychologists. L. Erlbaum, 2000.
[2]Hernández-Orallo, J., Loe, B.S., Cheke, L. et al. General intelligence disentangled via a generality metric for natural and artificial intelligence. Sci Rep 11, 22822 (2021). https://doi.org/10.1038/s41598-021-01997-7
[3]Not a Number: Identifying Instance Features for Capability-Oriented Evaluation. Ryan Burnell, John Burden, Danaja Rutar, Lucy Cheke, Jos´e Hern´andez-Orallo, Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, 2022