



# AAAI Tutorial

## Measurement Layouts for Capability-oriented AI Evaluation

John Burden<sup>1</sup>, Marko Tešić<sup>1</sup>, Konstantinos Voudouris<sup>1</sup>, Lucy Cheke<sup>1</sup>, Jose Hernandez-Orallo<sup>1,2</sup>

# Goals

- ❑ Realise the difference between capability-oriented evaluation vs task-oriented evaluation
- ❑ Identify task demands and how they can predict performance
- ❑ Understand the elements of the measurement layouts and their backwards / forwards inferences
- ❑ Effectively apply the measurement layout framework to estimate capabilities
- ❑ Use these capability profiles to infer performance for new task instances.
- ❑ Develop measurement layouts using PyMC in two scenarios:
  - ❑ agents in navigation tasks (in the Animal AI platform)
  - ❑ large language models
- ❑ Discuss the challenges and advanced topics (hierarchical models, demand annotation, etc.)

# Format and Requirements

- ❑ Format:
  - ❑ Presentations
  - ❑ Hands-on practical activities
  - ❑ Discussions
- ❑ Requirements:
  - ❑ Python: basic knowledge
    - ❑ You can use Google Colab if you don't have python in your computer.
  - ❑ PyMC: no previous knowledge needed
  - ❑ Statistics: the very basics (common discrete and continuous distributions).

# Pointers and Resources

- ❑ Burden, J., Voudouris, K., Burnell, R., Rutar, D., Cheke, L. & Hernandez-Orallo, J. (2023) Inferring Capabilities from Task Performance with Bayesian Triangulation. Arxiv preprint arXiv:2309.11975
- ❑ Burnell, R., Burden, J., Rutar, D., Voudouris, K., Cheke, L., & Hernandez-Orallo, J. (2022) Not a Number: Identifying Instance Features for Capability-Oriented Evaluation, *International Joint Conference on Artificial Intelligence (IJCAI)*.
- ❑ Burnell, R., Schellaert, W., Burden, J., Ullman, T.D., Martinez- Plumed, F., Tenenbaum, J.B., Rutar, D., Cheke, L.C., Sohl-Dickstein, J. Mitchell, M., Kiela, D., Shanahan, M. Voorhees, E.M., Cohn A. G., Leibo, J.Z. & Hernandez-Orallo, J. (2023) “Rethink reporting of evaluation results in AI: Aggregate metrics and lack of access to results limit understanding”, *Science*, Vol 380, Issue 6641, pp. 136-138.

<https://github.com/Kinds-of-Intelligence-CFI/measurement-layout-tutorial/>

# Schedule

- ❑ 14:00 - 14:10 Introduction
- ❑ 14:10 - 14:35 Why Capability-oriented Evaluation?
- ❑ 14:35 - 15:30 Measurement Layout Framework in PyMC
- ❑ 15:30 - 16:00 *Break*
- ❑ 16:00 - 16:55 Designing Good Benchmarks
- ❑ 16:55 - 17:30 Inferring the Capabilities of Large Language Models
- ❑ 17:30 - 18:00 Discussion