

Painel 2 – Sustentabilidade da IA: Energia, Recursos Naturais e Infraestrutura

Wagner Meira Jr., PhD
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais

Landscape of AI Infrastructure and sustainability

- AI as a modeling tool of environmental issues:
 - Making progress, not necessarily disruptive
- AI as a sustainability threaten:
 - Energy conservation efforts
 - Purchasing and construction rules and regulations now consider environmental issues
- AI as a ubiquitous demand
 - Digitization of a broad spectrum of products and services enabled AI productivity promises
 - Cost-benefit for society in the mid and long term is unclear
- How to address the tradeoff between demand and effects?
 - Should AI infrastructure be treated as usual?
 - How to handle the inherent speculative use of resources by AI-fueled services and products?
 - Does it make sense to set fixed and predefined sustainability-related parameters?

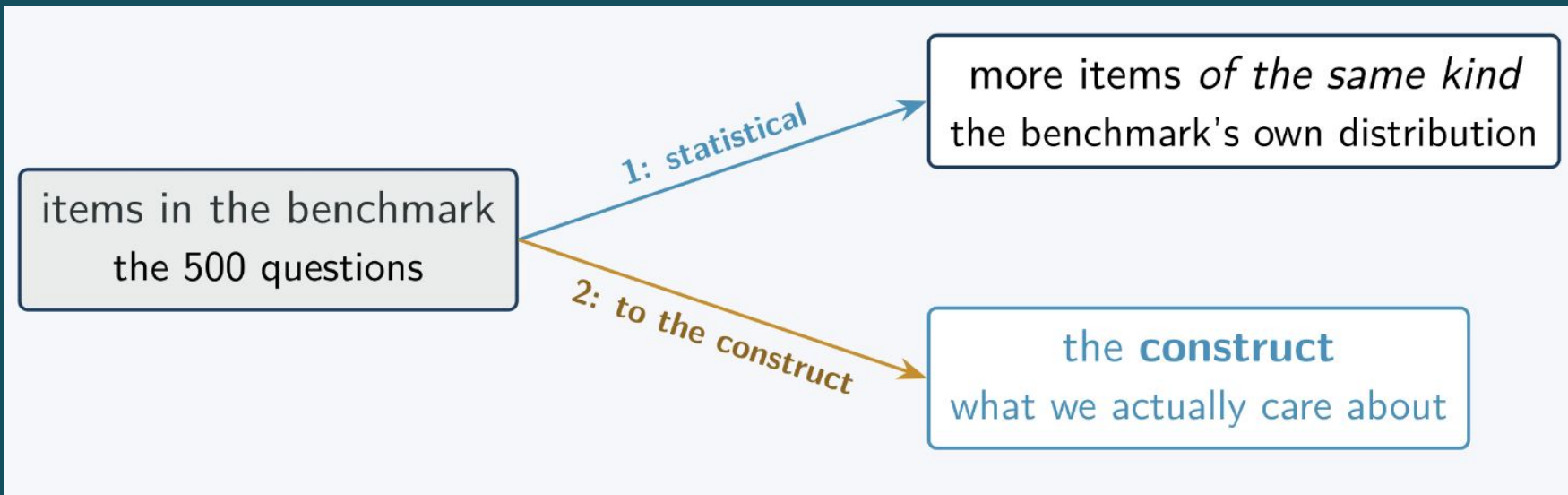
Evaluation is a problem of inference

- Observations (answers, tasks) are used to estimate latent properties (reasoning, coding ability, safety)
 - We estimate hidden properties of a system from what we can observe - evaluation does not end when a dataset is collected and a metric is computed.
- Issues
 - The score and the measurement target came apart
 - the score changed, the model did not improve or worsen
 - the score went up, the learning did not
 - A benchmark reports a number about specific items
 - We report the benchmark number because we expect it to generalize — to differently worded items, other task formats, and deployment conditions.
 - A benchmark can be flawless by every internal numerical criterion and still license the wrong claim. The scores are real and reproducible — the inference is often false.

Constructs and validity

- A construct is a general attribute of a system that we cannot observe directly but try to measure through items that depend on it.
 - Examples: reasoning, coding ability, factual reliability, safety.
 - When a score is used — to deploy a model, rank one lab above another, certify system as safe — it is read as a statement about a construct.
 - Validity is the name for whether that reading holds.
 - Reliability is necessary but not sufficient for validity. A measure of the wrong thing, measured consistently, is consistently wrong.

Benchmarks and Validity



- Arrow 1: test accuracy estimates performance on more items drawn the same way
- Arrow 2 from these items to the ability itself, under deployment conditions

Summary (or a tentative one)

- Environmental sustainability is a long term issue, but an urgent one
- AI as society improvement tool is a long term issue, but already current and broadly
- We need more than estimates, fixed parameters and thresholds
- It is necessary to effectively **measure** the impact of AI as a continuous and consistent effort
 - Benchmarks are just part of the solution and suffer severe limitations (e.g. Goodhart law)
 - Measuring needs to be contextualized and evolves dynamically
 - Sustainability should not be assessed in isolation, but as a key dimension