Stat 9911
Principles of AI: LLMs
Towards Building AI

Edgar Dobriban

Department of Statistics and Data Science, the Wharton School, University of Pennsylvania

January 16, 2025

Goal of AI

- Develop a system that behaves as a competent human in solving problems and performing tasks.

Definitions of AI

- OpenAI: *"Artificial General Intelligence (AGI) is a highly autonomous system that outperforms humans at most economically valuable work."*
- Marvin Minsky: *"AI is the science of making machines capable of performing tasks that would require intelligence if done by humans"* (Minsky, 1988). [specific tasks]
- McCarthy/Hernandez-Orallo: *"AI is the science and engineering of making machines do tasks they have never seen and have not been prepared for beforehand"* (McCarthy, 1987; Hernández-Orallo, 2017). [new tasks]

Definitions of Human Intelligence

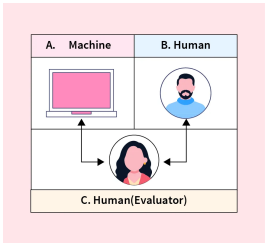
- American Psychological Association: *"Individuals differ from one another in their ability to understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, to overcome obstacles by taking thought"* (Neisser et al., 1996).
- Legg et al. (2007) list 70 definitions of intelligence and conclude: "Intelligence measures an agent's ability to achieve goals in a wide range of environments."

Specific vs General Skills

- For humans, being skilled at a specific task often involves general skills (e.g., no-one is born knowing how to play chess; but to learn chess you need to learn to read, to plan ahead, ...).
- For algorithms, performance on specific tasks does not imply AGI (see e.g., Chollet, 2019, for discussion).
- Chollet (2019): *"The intelligence of a system is a measure of its skill-acquisition efficiency over a scope of tasks, with respect to priors, experience, and generalization difficulty."* [learn to perform new tasks]

Historical Perspective

- The term "artificial intelligence" dates to the 1950s.
- Turing Test: Can an AI be distinguished from a human via text interactions? Proposed by Alan Turing (Turing, 1950).



- If AI knows the distribution of human text, then it can pass (in a statistical sense).
- Knowing a distribution is (roughly) equivalent to optimal compression. A basic claim of information and coding theory (Cover and Joy, 2006; Salomon, 2007).
- This is the basis of the claim that "Intelligence is equivalent to compression"

A More Modern Perspective

- Focus on capabilities, as opposed to building a copy of humans.
- Avoid replicating human flaws (e.g., corruption, bias, societal defects).

Is AGI Possible?

- Are there limitations related to physics, energy, computation, or data?
  - I have not seen a lot of convincing arguments.
- Can a system behave as a competent human in daily tasks?
  - Example: Steve Wozniak's coffee cup test. Entering a previously unknown kitchen and making a cup of coffee (Wozniak and Moon, 2007).
  - Sounds simple, but involves a huge number of tasks: find coffee (instant? or need to roast/grind beans?), find coffee machine, find cup, move things appropriately, ...,

Approaches to Building AGI

- A: Build an exact brain copy (requires engineering 86 billion neurons).
  - Far out of reach; e.g., current SoTA is a model of the fly brain: 140k neurons, 50 million synapses (Dorkenwald et al., 2024) (vid).

- B: Build a simplified system (feasible):
  - Use neural network architectures.
  - Borrow loose inspirations from biology.

Key Requirements for AI

- Process human-like information (visual, textual, sound, video).
- Reason and plan.
- Take actions in an appropriate environment.
- Sub-goal: Process text. Text contains a lot of information, and can also serve as an interface to non-nextual components (e.g., verbal instructions to robot).
- Current SOTA: Large Language Models.

References

F. Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

T. Cover and A. T. Joy. *Elements of information theory*. Wiley-Interscience, 2006.

S. Dorkenwald, A. Matsliah, A. R. Sterling, P. Schlegel, S.-C. Yu, C. E. McKellar, A. Lin, M. Costa, K. Eichler, Y. Yin, et al. Neuronal wiring diagram of an adult brain. *Nature*, 634(8032):124–138, 2024.

J. Hernández-Orallo. Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. *Artificial Intelligence Review*, 48:397–447, 2017.

S. Legg, M. Hutter, et al. A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and applications*, 157:17, 2007.

J. McCarthy. Generality in artificial intelligence. *Communications of the ACM*, 30(12):1030–1035, 1987.

M. Minsky. *Society of mind*. Simon and Schuster, 1988.

U. Neisser, G. Boodoo, T. J. Bouchard Jr, A. W. Boykin, N. Brody, S. J. Ceci, D. F. Halpern, J. C. Loehlin, R. Perloff, R. J. Sternberg, et al. Intelligence: knowns and unknowns. *American psychologist*, 51(2):77, 1996.

D. Salomon. *Data Compression: The Complete Reference*. Springer London, 2007.

A. M. Turing. Computing machinery and intelligence. *Mind*, 59:433–460, 1950.

S. Wozniak and P. Moon. Three minutes with steve wozniak. *PC World*, 2007.