Artificial Intelligence

Alex Sverdlov
alex@theparticle.com

1 Introduction

> *"Intelligence is not what you know, it's what you do when you don't know."*
> —Jean Piaget

What we consider AI is evolving, and every generation puts focus on different set of capabilities:

1. AI as computation: a machine calculating trajectories, automating the work of many humans.
2. AI as search: a chess program that searches for the best move would be in the realm of AI.
3. AI as logic: a program that deduces answers from a sequence of if-then statements.
4. AI as probabilities: a program that determines that email is spam.
5. AI as concept learning: a program that learns concepts from labeled data, and later recognizes those concepts.
6. AI is compression: if we can compress text, then we learned things about language.
7. AI as pattern recognition: identify cats in pictures.
8. AI as generating text: a program that generates plausible text.
9. AI as image generation: a program that generates images from text. video generation.
10. AI as assistant: trick the program that generates plausible text to generate text that answers questions.
11. AI as agents: loop on AI assistant that prompts itself with questions/tasks to achieve some goal.
12. AI as ...etc.

AI is intelligence exhibited by non-living things. A more precise definition is problematic (see above), since intelligence itself is poorly defined. Broadly speaking, intelligence is the ability of individuals to effectively navigate their environment and to achieve goals.

We hope *we will recognize intelligence when we see it*, and that said, we expect intelligent beings to act or respond rationally. That means recognizing patterns in their inputs, correctly interpreting those patterns, learning from them, planning, and producing output that maximizes some utility. A mechanism would not be considered intelligent if it failed to learn from experience or acted in ways detrimental to itself.

It appears intelligence is best viewed as an aggregate of attributes (pattern recognition, learning from experience, etc.), some of which may not be present in certain contexts. For example, an agent may respond/behave intelligently, yet have no learning or understanding capability. These capabilities may not be expressed at the granularity of an individual: viruses & bacteria learn! The same "aggregate of attributes" is probably true of conciousness—where no single attribute determines the answer.

2 Mimicry

It appears that mimicry is one of the key capabilities that AI should aim for. Every process generates data. An agent that can make some observations and then properly mimic some useful aspect of those observations is sliding in the right direction towards intelligence.

Why would mimicry matter so much? It implies that the agent can observe and recognize useful patterns, have some internal compact representation of those patterns, do some calculations in order to produce plausible looking outputs.

Recent advances in large language models started by predicting 1 plausible word to follow a prompt. There's a lot of text out there to learn to mimic it.

3 Science

Automation of science is one of AI's goals.

There are essentially three ways to learn: *memorization*, *deduction*, and *induction*. With memorization, the learner simply memorizes (stores in a database or file) all facts and observations. These facts can then be recalled, or matched to input. A rote memorizer cannot cope with inputs that were not previously observed, nor can any predictions be drawn from the data.

With deduction, the learner gets predictive power; starting with knowledge that the learner knows to be *true* (facts, observations, or previously proved items), the learner attempts to derive other truths by applying laws of logic and math. A classic example, given the facts: *"All men are mortal"* and *"Socrates is a man"*, the learner may *deduce* that *"Socrates is mortal"*. One generally ignores that axioms of logic and math are not themselves rigorously proven.

Induction is what we are interested in for intelligent beings; it is deduction with assumptions. The inductive learner makes observations (usually input/output pairs) and assumes (guesses) that they were generated by some *process*, e.g., a normally distributed random variable. Then working backwards, the learner uses these observations to construct a model of this process that mimics the observations in some way, e.g., the learner may construct a stochastic process that has the same *mean* and *variance* as the observations. This model is the 'learned knowledge', and can be used in place of the process to make predictions.

The amorphous concept of *model* can be anything at all: our brain is a model, our computer is a model, etc.

There is a lot of recent talk that "Language Models are Unsupervised Multitask Learners" (lookup this paper). Essentially a learning task is often phrased as estimating conditional distribution $P(output \| input)$. We can incorporate the *task* into the learning task via: $P(output \| input, task)$, for example, the task could be: "translate from english to french." With LLMs, the task is part of the *input* prompt... therefore, LLMs should be able to perform any task.