# A Desiderata for Machine Learning

I was recently having lunch with a friend when he was talking about an article he had seen about how crows are really smart because they can use sticks to hunt insects.

While I get the point about the crows, I think we pretty consistently have double standards for measuring our own intelligence and that of other animals. This extends to our perception of AI.

---

Basically everyone who has an opinion on the matter seems to either deny the possibility of AI replacing us or has accepted our impending obselesence. Below are four reasons why I think our fate may not be sealed just yet.

- Language models, like GPT-3, are only really trained to *pretend* to be intelligent. Unless we work out how to reorient LLM training to prioritise problem solving skills over the current game of statistical word association, general reasoning at a human level (e.g. the ability to play chess well) is going to take an ungodly amount of compute.
- The original idea of "sample the dataset and hope the model generalises to the full distribution" doesn't seem to scale particularly well anymore. I find it sort of ironic that now skilled humans can now be *more* consistent than machines.
- GPT-3 training was performed on ~700B tokens. With dedicated hardware, a child learns to speak with about one millionth of that,[1,2] while consuming 10W of power (compared to ~25,000W for GPT-3[3]) on hardware around the same size as a single GPU.[4]
- Assuming there is a (useful) total order on intelligence,[5] there must be some least upper bound for how much you can fit into a $m^3$. This is probably a lot higher than the density of a human brain, but I would argue that it represents a change in degree rather than in kind, especially since there are still far more human brains than GPUs.[6]

---

I think these differences naturally lead us to the following requirements for future ML models.

- Models need to account for (or minimise) the distribution shift between their training distribution and their target task (see Agnostic FL). More generally, models need to be *useful*, rather than just high accuracy (see *model alignment*).
- Models should be robust to unseen data in the tail of the distribution. Both ML models and our brains seem to consist of a mish-mash of heuristics, while a perfectly robust solution would learn a model that more closely describes the causal nature of the data. How can we bridge this gap? George Hotz argues it may be impossible.
- Model training and inference need to be more efficient wrt data and power. A perfect training routine would result in no entropy loss between the model's training data and the updates made to its initial knowledge representation.[7,8]
- Finally, we need to be able to verify all of the above. Current methods of interpreting models generally struggle to reconcile continuous weight spaces into discrete programs for all but the most simple circuits.[9]

As a bonus requirement, if we ever solve these problems, we need to do somehow share these solutions while preventing their use for malicious purposes such as for voilating user privacy, or misinformation (consider Cambridge Analytica, but with access to LLMs; see also here).

---

In 2014, Elon Musk described AI as 'summoning the demon'. I disagree: at the very least, we are constructing it in a far more laborious manner. Practical implications aside, I believe *artificial* neural networks remain some distance behind *real* neural networks.

*originally written summer 2021, edited 07/06/24*

---

1. The advantage of a brain over GPUs is that it skips a layer of simulation: the neural network is 'implemented' directly in physics, while artificial neural networks are simulated on hardware that is itself implemented in physics. Perhaps it is not even possible to achieve the same metrics with the latter setup. ↩

2. This is not a totally fair comparison since a human child learns (or is taught) with on-policy RL (and actively searches for instances in the tail of the distribution), while an LLM performs most training completely off-policy. ↩

3. Adjusted for the same time period. ↩

4. To be fair, GPT-3 is probably smarter than a six year old, but you get the point. ↩

5. I think this is a pretty interesting question in itself. More generally, what does intelligence mean? Perhaps it is the speed by which an agent can traverse the likely paths of a search space. Alternatively, it could be how effectively knowledge is stored in a general representation. This latter definition is neat because it has a very similar vibe to the interchangeability of instructions and state in programming languages. ↩

6. At the time of writing. ↩

7. From an engineering standpoint, we can also get ahead by finetuning a general model rather than starting from scratch. An interesting problem is how a large model like this can then be translated into a more optimised representation. ↩

8. ML (perhaps computer science in general) seems to lead a double life as both the "physics" (understanding - almost like bottom up neuroscience) and the "engineering" (application) of intelligent systems. ↩

9. Mechanistic interpretability is currently built on hard-coded heuristics, much like the AI methods of last century. I hope we can find more general, emergent methods that mirror the shift in perspective that ML has had for AI s a whole. ↩