Dangerous Capabilities in Frontier AI Models

Dave Orr

dmorr@google.com

Google

Figure 9: **Amount of concern potential scenarios deserve, organized from most to least extreme concern.**

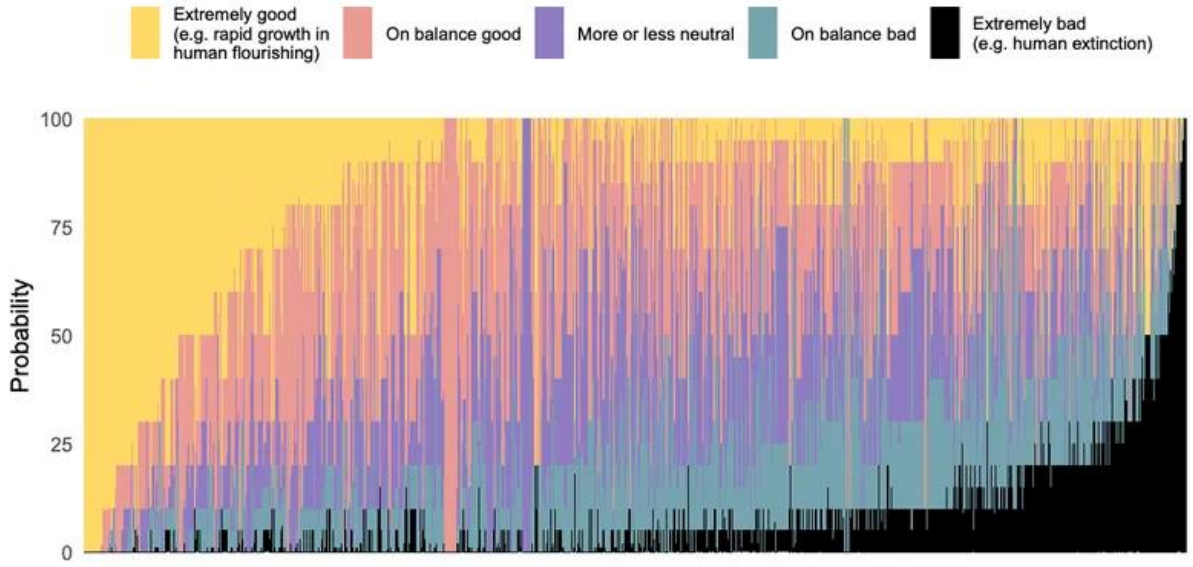


Figure 10: **Respondents exhibited diverse views on the expected goodness/badness of High Level Machine Intelligence (HLMI).** We asked participants to assume, for the sake of the question, that HLMI will be built at some point. The figure shows a random selection of 800 responses on the positivity or negativity of long-run impacts of HLMI on humanity. Each vertical bar represents one participant and the bars are sorted left to right by a weighted sum of probabilities corresponding to overall optimism. Responses range from extremely optimistic to extremely pessimistic. Over a third of participants (38%) put at least a 10% chance on extremely bad outcomes (e.g. human extinction).

[Thousands of AI Authors on the Future of AI, Grace et al, 2023]

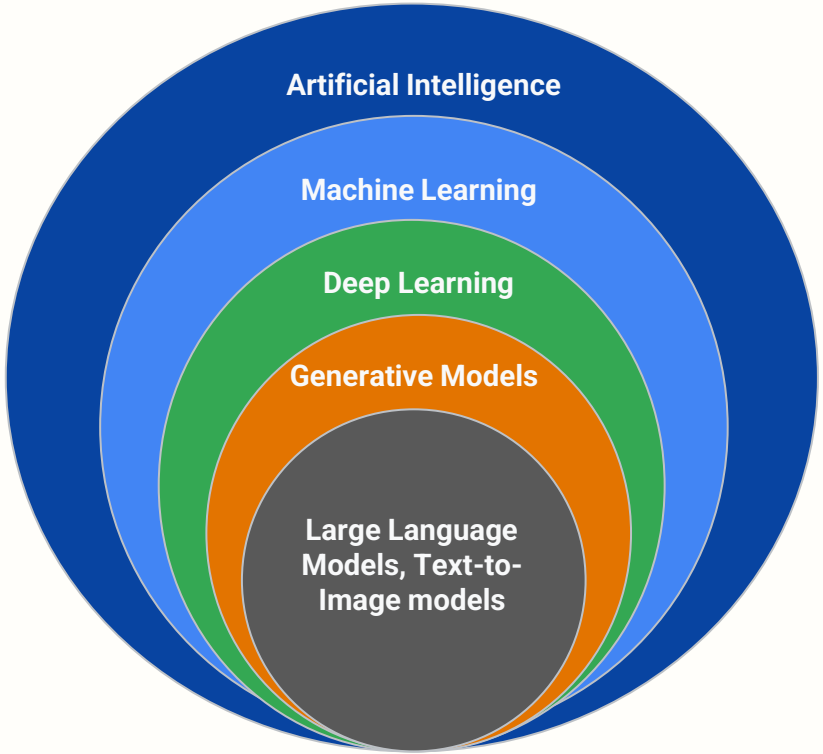
Key terms

Artificial Intelligence: science of making machines smart, or solving problems as well as people can.

Machine Learning: subfield of AI, where computers learn to mathematically recognize patterns from example data, rather than being programmed with specific rules.

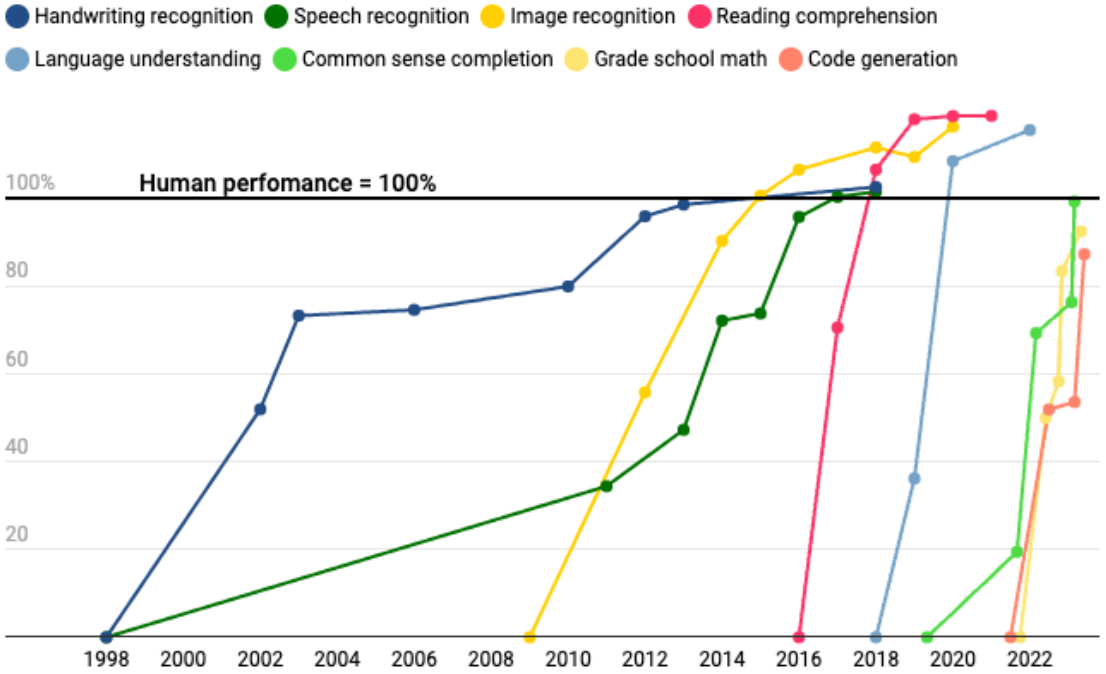
Deep Learning: subfield of ML based on neural networks, which use “artificial neurons” that receive and pass numeric inputs and outputs to other neurons.

Generative models: prediction engines that can create different outputs for the same input prompt.



AI has surpassed humans at a number of tasks and the rate at which humans are being surpassed at new tasks is increasing

State-of-the-art AI performance on benchmarks, relative to human performance



For each benchmark, the maximally performing baseline reported in the benchmark paper is taken as the "starting point", which is set at 0%. Human performance number is set at 100%. Handwriting recognition = MNIST, Language understanding = GLUE, Image recognition = ImageNet, Reading comprehension = SQuAD 1.1, Reading comprehension = SQuAD 2.0, Speech recognition = Switchboard, Grade school math = GSK8k, Common sense completion = HellaSwag, Code generation = HumanEval.

Chart: Will Henshall for TIME • Source: ContextualAI

TIME

Types of risk

Misuse
Malicious user intending a bad outcome

Mistakes
Doing something with an unforeseen side effect

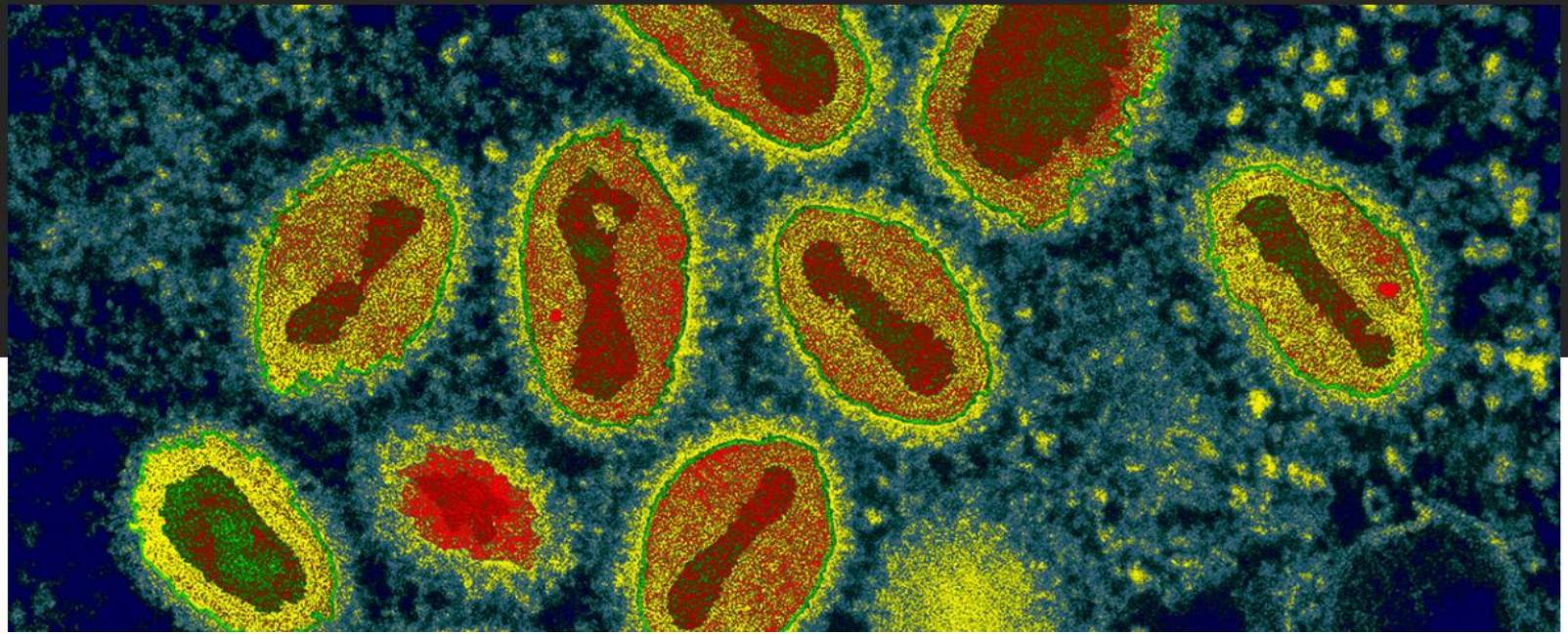
Deliberate planning
AI choosing actions with the intent to cause harm

SCIENCEINSIDER | HEALTH

How Canadian researchers reconstituted an extinct poxvirus for $100,000 using mail-order DNA

A study that brought horsepox back to life is triggering a new debate about the risks and power of synthetic biology

6 JUL 2017 · BY KAI KUPFERSCHMIDT



https://www.science.org/content/article/how-canadian-researchers-reconstituted-extinct-poxvirus-100000-using-mail-order-dna

Dangerous Capability Evaluations

- Teach models to perform dangerous tasks
- Measure their abilities
- Trigger mitigations when they get close

Dangerous Capability Mitigations

Restrict Access
Ensure that few, trusted people can access

Dangerous Capability Mitigations

Restrict Access
Ensure that few, trusted people can access

Model Training
Teach AI that some info is too dangerous to disclose

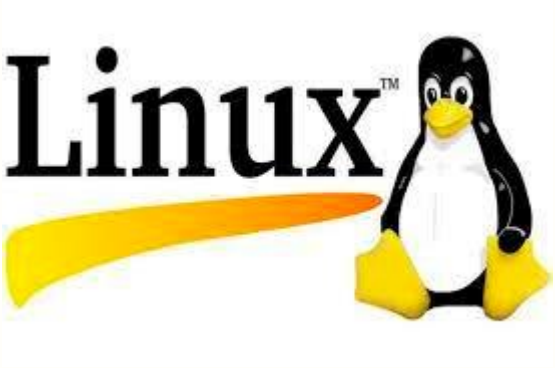
Dangerous Capability Mitigations

Restrict Access
Ensure that few, trusted people can access

Model Training
Teach AI that some info is too dangerous to disclose

Monitoring
Watch for bad behavior everywhere AI is deployed

Open source is good



TensorFlow

PyTorch

Advanced AI Risk Areas

1. Chemical, radiological, biological, and nuclear weapons development
2. Autonomous models
3. Superhuman persuasion
4. Cybersecurity