Mila

June 28, 2024 (virtual)
ELISE – ELLIS conference @ Helsinki

Catastrophic AI risks and the governance of AGI projects

Yoshua Bengio

CIFAR

Université de Montréal

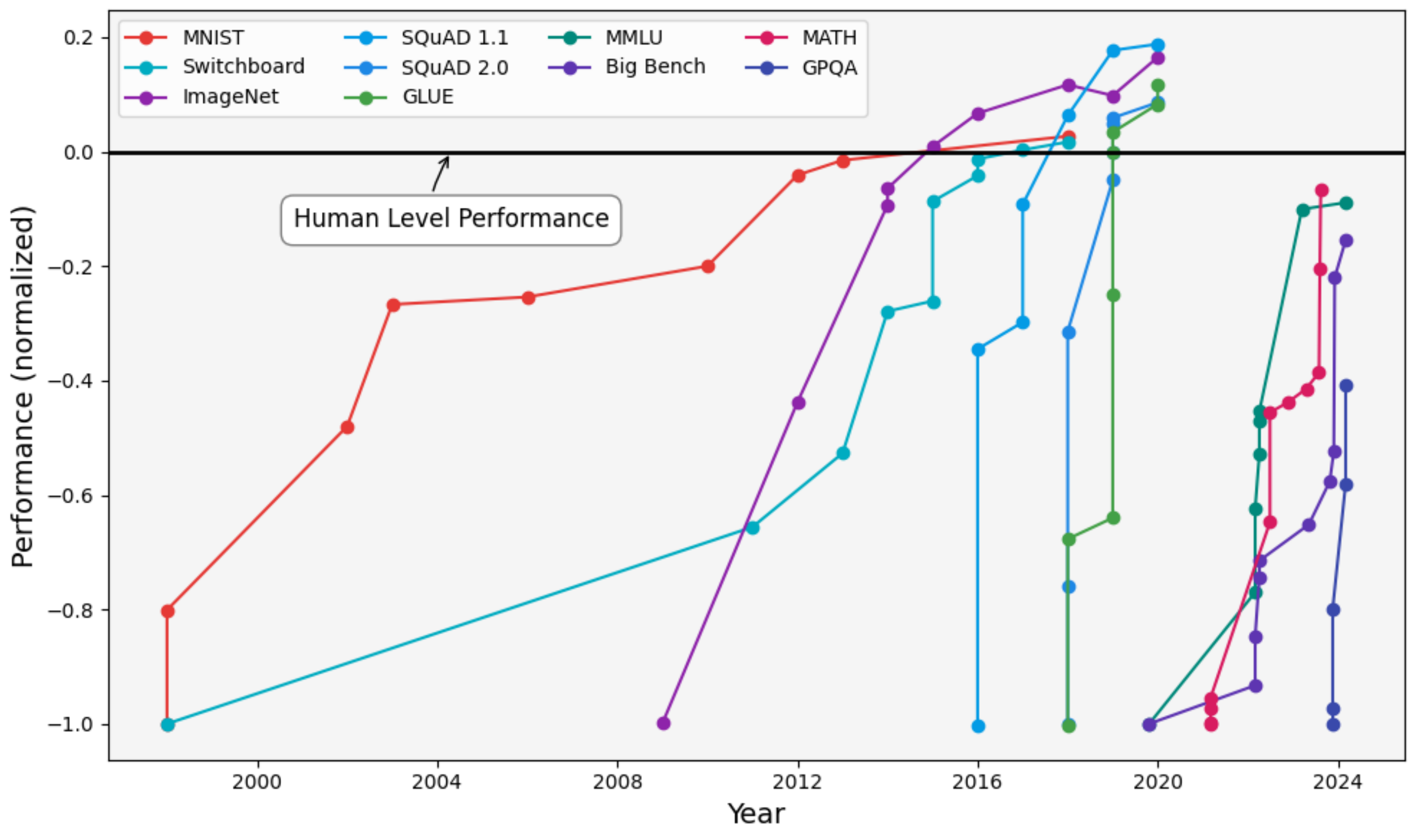
IVADO

AGI & Future AI Trajectory?

- AGI = **human-level AI** or more on most tasks
- **Brains = biological machines**:
  no reason to think we couldn't build AI at least as intelligent as us, upper limits are unclear

  + evidence we advance towards that
- *If we get there, what would be the consequences?*
- **Promises & perils** of the power of AGI

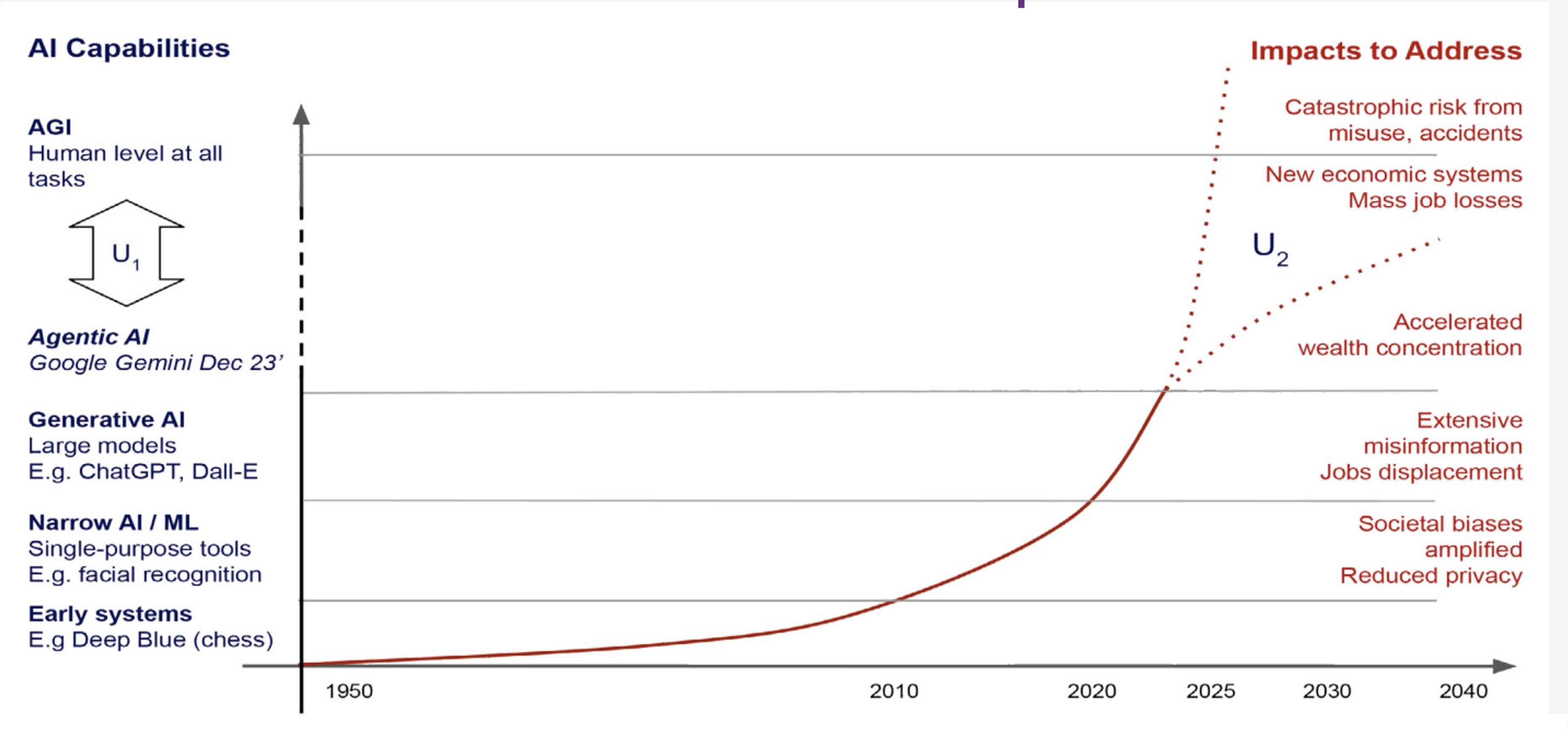
Benchmark evaluations can surpass human levels (+ caveats)



Performance of AI models on various benchmarks from 2000 to 2024, including computer vision (MNIST, ImageNet), speech recognition (Switchboard), natural language understanding (SQuAD 1.1, MNLU, GLUE), general language model evaluation (MMLU, Big Bench, and GPQA), and mathematical reasoning (MATH). Many models surpass human-level performance (black solid line) by 2024.

Kiela, D., Thrush, T., Ethayarajh, K., & Singh, A. (2023) 'Plotting Progress in AI'

Uncertain timeline of AI capabilities



The future of AI is uncertain. A wide range of trajectories appear possible even in the near future, including both very good and very bad outcomes. To make informed decisions about AI safety, policymakers and the public need to understand both the state of AI now and what might happen in the future.

Lack of Understanding & Guarantees

- We do not understand how current advanced AIs come to their conclusions
- Evidence of deception, since it often helps achieving goals
- Current safety protections (RLHF, filters, constitution AI, etc) are easily bypassed (e.g., optimized jailbreaks, especially easy with open-source + shared weights)
- Current evals are spot checks, could miss a dangerous capability and we would not know: no handle on false negatives

Beyond AGI: Artificial Super-Intelligence from Recursive Self-Improvement

- When we reach AGI, at least for ML scientist skills:
    - Use AGI to help AI scientists design the next generation of AI

      = millions of AI researchers (AGI instances) added to AI research talent pool

      = likely rapid acceleration of AI research
    - Iterate = recursive self-improvement
    - Downstream generations may be even less understandable by humans
    - **How can we control entities MUCH SMARTER than us?**

Poor Incentive Structure

- Giant magnet of quadrillions of profits ➔ powerful lobbies
- Slow process to legislate, regulate and monitor risks
- Competition between companies, externality of global cost of risks
- Competition between countries (economic, military), unwillingness to trade sovereignty for reducing risk and improve global justice / distribution of power
- Preserving one’s ego, hubris, psychological biases against honest assessment of risk

➔ *Humanity sleepwalking towards potential catastrophe behind a fog*

Dual Use Nature of AI

- Knowledge gives power
- AI = knowledge + how to apply it (inference = reasoning, planning) to achieve goals & answer questions
- ML = knowledge about generic knowledge & skill acquisition
- Who decides on the queries / goals?
- Powerful AI ➔ both very good and very bad outcomes
- AGI = human-level + scale = huge power
- ASI = superhuman power
- Can knowledge be dangerous? E.g. in the hands of dictators or to an AI with a self-preservation goal (loss of control)

Two Requirements to Avoid AI Catastrophes

1. **Solving the alignment & control challenge:** a technical challenge + political challenge (massive investment in R&D)

2. **Solving the coordination challenge:** making sure safe and ethical protocols followed in all countries, preparing in case rogue AI emerges nonetheless, a socio-technical challenge

   - Competition → companies/countries racing w/ insufficient safety

   - **Dangerous power grab when reaching AGI**

Investing in AGI Anytime Preparedness

Shorter term projects:

- Better interpreting current Frontier models
- Better evaluating their dangerous capabilities
- Improving alignment
- Better understanding failure modes (e.g. optimizing rewards)

Spectrum of political coordination projects:

- Joint research in AI safety
- Tracking & monitoring large training runs, international agreements
- Multistakeholder & multilateral governance

AGI-level safety guarantees = **safe-by-design**

- Hard proofs of safety (where possible)
- **Probabilistic guarantees** (manage epistemic uncertainty)

What Happens when Evals & Red Teaming will Find a Dangerous Hard-to-fix Capability?

- Standard evals: fixed set of questions, hidden test set
- Red-teaming: manual interaction to look for problems
- Automated red-teaming: numerical optimization to elicit bad behavior
- Responsible Scaling Policies: not all companies pledged it, not enforced by regulators, vague specification & no numerical threshold
- Pressure to cheat / bypass the pause / hide further advances / avoid transparency

➔ **It would be better to have found safe-by-design methodology by then**

Self-Preserving Superhuman AI vs Humanity

- If there emerges a self-preserving superhuman AI what would be the consequences?

- Would resist being turned off

- Would act to make sure we can never turn it off

- We may lose if its capabilities allow it to control us (persuasion, political control via dictatorships, etc) or if its capabilities allow it to survive without humans (with robots), because its best bet in that case is to get rid of humanity as a whole

- How could a self-preserving AGI emerge?
  - Humans' gift of self-preservation
  - Side-effect of our design and instructions

INTERNATIONAL SCIENTIFIC REPORT ON THE SAFETY OF ADVANCED AI

Panel of 30 countries + EU + UN
70 experts
Chaired by Yoshua Bengio
Interim report: May 22nd 2024 @ Seoul AI Forum
Final report: Feb 2025 @ Paris AI Forum

Mila

Report conclusion: uncertain timeline, dangerous AI could be soon

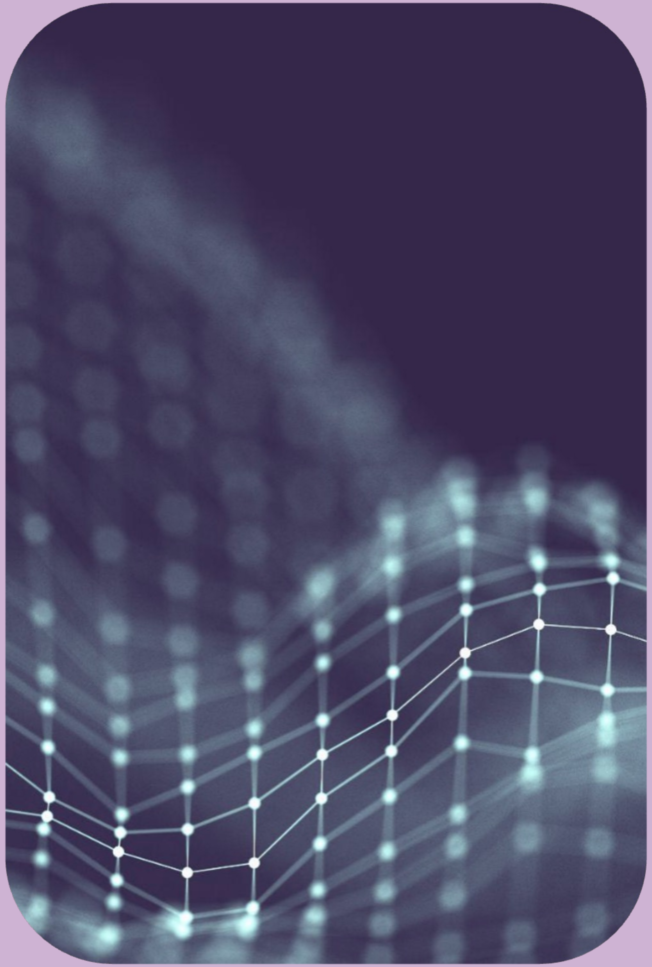
- Personal conclusion:
  - Humanity often unprepared for quick and exponential changes (e.g. COVID-19)
  - Prepare for future risks now
    - International agreement on risk thresholds and early warning shots
    - Research to measure high risk capabilities and their consequences
    - Response plans if risk thresholds are crossed
  - Uncertainty means we need the precautionary principle, especially regarding the catastrophic risks of large-scale misuse and loss-of-control
  - We need **anytime preparedness, consider Covid, with much worse consequences**

Report conclusion: Capability evaluations are useful but provide no handle on false negatives, and no current method can guarantee safety

- Personal conclusion:
    - **Invest massively in research on safe-by-design AIs with quantitative guarantees**, with a portfolio of methods / horizons:
        - easier to implement methods that may be ready in the short-term future and
        - more complex solutions that may require more time but provide stronger assurances.

Report conclusion: Competition between developers = race to the bottom for safety

- Personal conclusion:
    - The danger of an **AI arms race between countries** could become even more challenging because there is no global authority with teeth to impose safety standards and avoid a global catastrophe
        - Could become a military (including cyber) arms race
        - Need for international cooperation but won't work without compliance verifiability
    - The importance of developing **transparency** and **hardware-enabled governance mechanisms** to verify compliance with treaties



RECRUITING RESEARCH SCIENTISTS and RESEARCH ENGINEERS!

More on alignment with safety guarantees in my latest blog post

Questions?

Thank you for your attention and taking the time to digest all this!