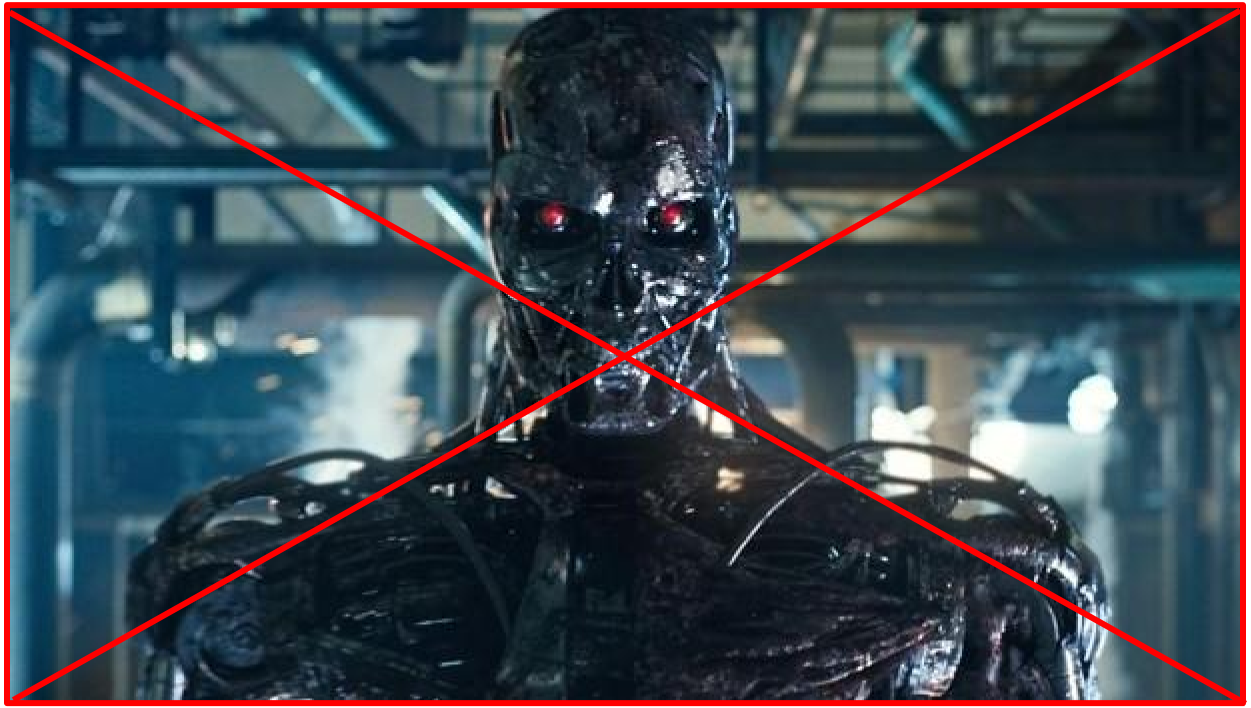
AI risk: why and how?

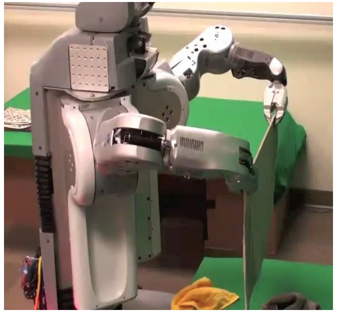
future of life
INSTITUTE

Victoria Krakovna, co-founder

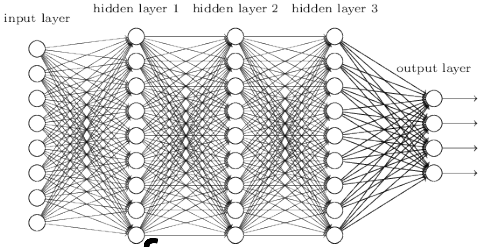
Danger: competence, not malice



Advanced AI considerations



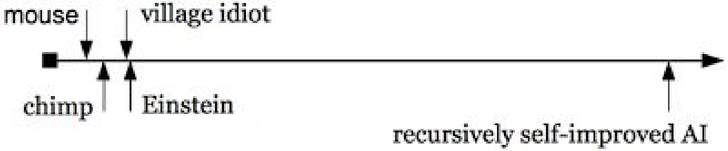
- accelerating progress in AI in the last 10 years
- many experts expect human-level AI **this century**
- **not imminent**, but would be a big deal!
- **best case:** solves disease, poverty, corruption, other coordination problems
- **worst case:** existential risk
- extremely high-variance technology
- we can do research now to increase chances of a positive outcome



AlphaGo

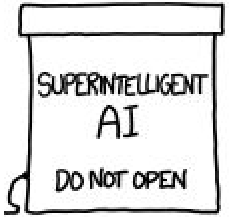
Risk aspects

- **intelligence explosion**

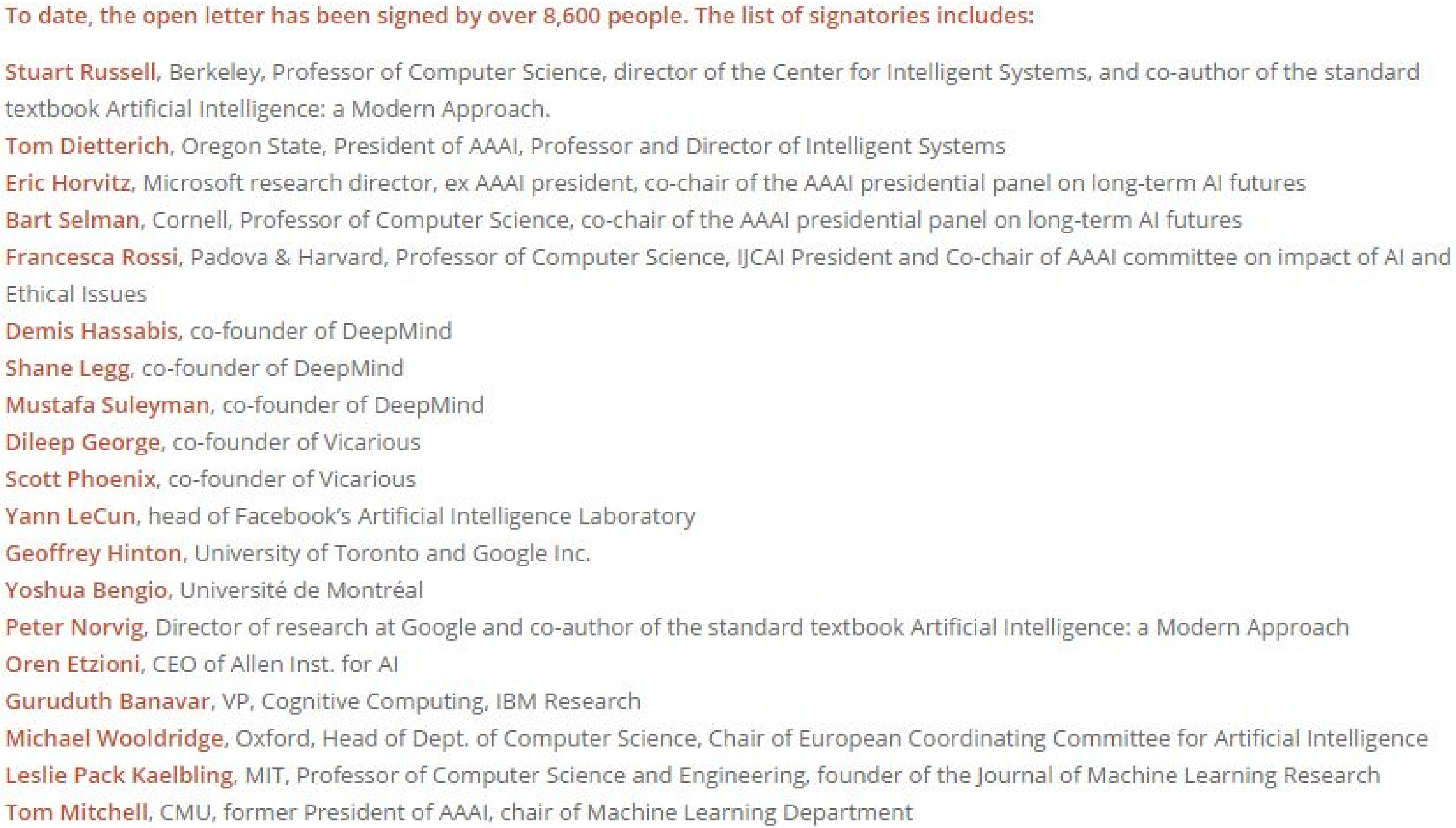


- problematic **instrumental objectives:**
    - resource acquisition, self-preservation, etc
- misspecification / misalignment of **values:**
    - setting omitted variables to extreme values
- **unintended consequences** / shortcuts

- **human incentives**: bad actors, arms race potential, etc
- **containment**



Open letter on robust & beneficial AI



To date, the open letter has been signed by over 8,600 people. The list of signatories includes:

Stuart Russell, Berkeley, Professor of Computer Science, director of the Center for Intelligent Systems, and co-author of the standard textbook Artificial Intelligence: a Modern Approach.
Tom Dietterich, Oregon State, President of AAAI, Professor and Director of Intelligent Systems
Eric Horvitz, Microsoft research director, ex AAAI president, co-chair of the AAAI presidential panel on long-term AI futures
Bart Selman, Cornell, Professor of Computer Science, co-chair of the AAAI presidential panel on long-term AI futures
Francesca Rossi, Padova & Harvard, Professor of Computer Science, IJCAI President and Co-chair of AAAI committee on impact of AI and Ethical Issues
Demis Hassabis, co-founder of DeepMind
Shane Legg, co-founder of DeepMind
Mustafa Suleyman, co-founder of DeepMind
Dileep George, co-founder of Vicarious
Scott Phoenix, co-founder of Vicarious
Yann LeCun, head of Facebook's Artificial Intelligence Laboratory
Geoffrey Hinton, University of Toronto and Google Inc.
Yoshua Bengio, Université de Montréal
Peter Norvig, Director of research at Google and co-author of the standard textbook Artificial Intelligence: a Modern Approach
Oren Etzioni, CEO of Allen Inst. for AI
Guruduth Banavar, VP, Cognitive Computing, IBM Research
Michael Wooldridge, Oxford, Head of Dept. of Computer Science, Chair of European Coordinating Committee for Artificial Intelligence
Leslie Pack Kaelbling, MIT, Professor of Computer Science and Engineering, founder of the Journal of Machine Learning Research
Tom Mitchell, CMU, former President of AAAI, chair of Machine Learning Department

Research priorities

- **verification:** *did I build the system right?*
- **validity:** *did I build the right system?*
  - e.g. value learning, decision theory
- **security:** *how to prevent external or internal manipulation?*
  - e.g. containment, anomaly detection
- **control:** *OK, I built the system wrong, can I fix it?*
  - e.g. interruptibility, corrigibility

FLI's research grants program

- $10M **funding** pledged by Elon Musk
- Announced 3-year international **grants program** to fund AI safety research (Jan 2015)
- Received ~300 **applications** totalling $100M in requested funds (Mar 2015)
- **Awarded** grants to 37 teams: 32 project grants, 1 center grant, 4 outreach grants (Jul 2015)
- Considering a **second round** of grant program in 2017

Funded research projects

- Alex Aiken: Verifying Deep Mathematical Properties of AI Systems
- Thomas Dietterich: Robust and Transparent Artificial Intelligence Via Anomaly Detection and Explanation
- Owain Evans: Inferring Human Values: Learning “Ought”, not “Is”
- Benya Fallenstein: Aligning Superintelligence With Human Interests
- Percy Liang: Predictable AI via Failure Detection and Robustness
- Francesca Rossi: Safety Constraints and Ethical Principles in Collective Decision Making Systems
- Stuart Russell: Value Alignment and Moral Metareasoning
- and **30 other projects**