Challenges in AI Safety and Security

Miles Brundage

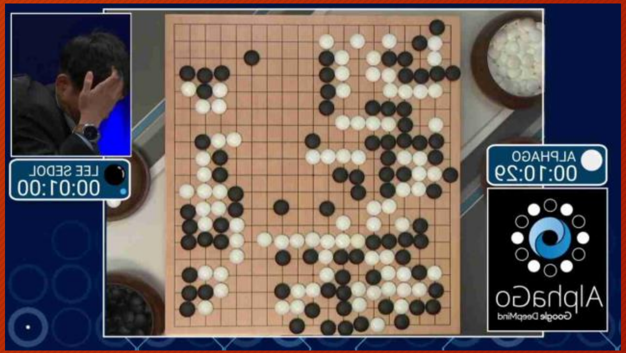
Overview

- Emerging AI capabilities
- AI safety
    - Domain-specific/immediate
    - General/near-term
    - General/long-term
    - Solutions
- AI security
    - Degradation
    - Weaponization
    - Solutions
- The role of policy

“AI”

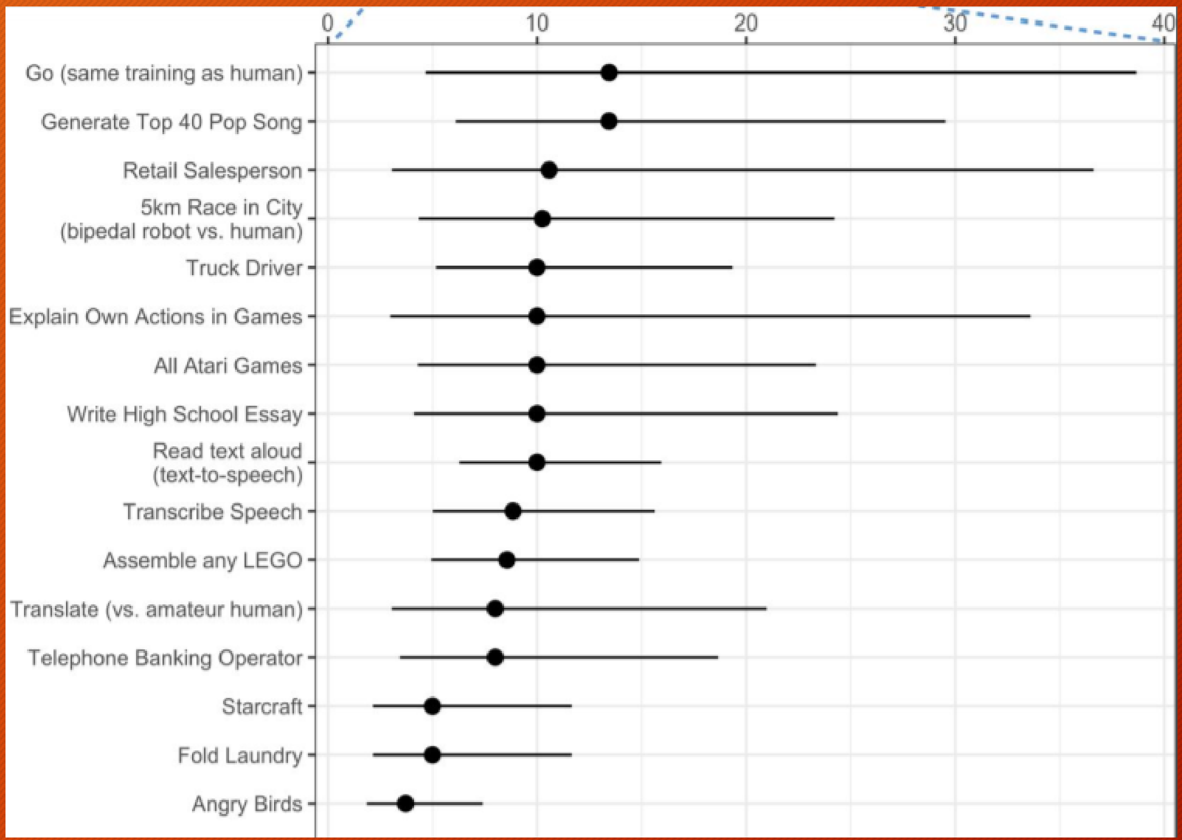
- Digital systems that express appropriate behavior in response to changes and opportunities in their environment
- Spectrum from low-performance, narrow AI to superintelligence (systems that outperform individual humans across almost all relevant cognitive domains)

Emerging AI Capabilities



Emerging AI Capabilities (cont'd)

- Uncertainty about the future rate of progress



Grace et al., 2017

AI Safety

- A taxonomy:
  - Domain-specific/immediate
  - General/near-term
  - General/long-term

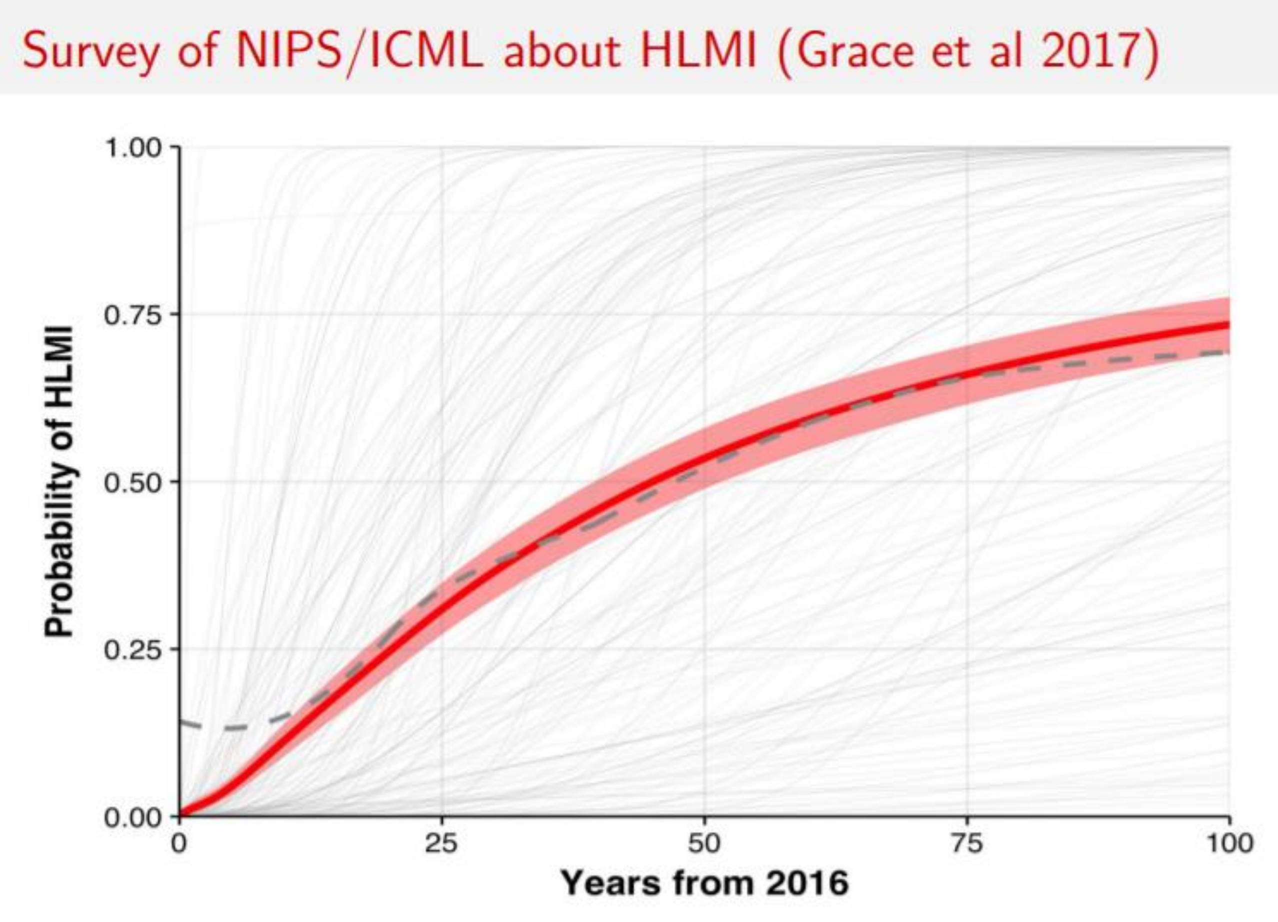
Domain-specific/immediate

- Examples
  - Driverless cars
  - Medical applications of AI
  - Drone delivery
  - Semi-autonomous weapons systems
  - Etc.
- General features of these safety risks
  - Unexpected/undesired behavior
  - Outside the range of what humans would do (hard to intuitively model the risks)
  - Domain-specific triggers (e.g. pedestrians; drones hitting electrical wires; etc.)
  - Verification/validation are hard - large state/action spaces

General/near-term

- https://www.youtube.com/embed/tlOIHko8ySg
- Categories (Amodei and Olah et al., 2016):
    - “having the wrong objective function ("avoiding side effects" and "avoiding reward hacking"), an objective function that is too expensive to evaluate frequently ("scalable supervision"), or undesirable behavior during the learning process ("safe exploration" and "distributional shift")”

General/long-term - how long term?



General/long-term – how high (capability)?

- Lower bound on physically possible long-term capability, assuming no “magic” involved in human thought:
    - Group of the smartest humans who have ever lived, coordinating closely in a virtual environment, a million times faster than natural humans
- Why a lower bound? Perfect coordination and qualitative cognitive advantages
- Lots of uncertainty about the ultimate long-term capabilities of AI and how quickly they may arise

General/long-term (text below from Bostrom, 2016 slides; see also Bostrom, 2014)

- The orthogonality thesis
  - Intelligence and final goals are orthogonal: more or less any level of intelligence could in principle be combined with more or less any final goal.
- The instrumental convergence thesis
  - Several instrumental values can be identified which are convergent in the sense that their attainment would increase the chances of the agent's goal being realized for a wide range of final goals and a wide range of situations, implying that these instrumental values are likely to be pursued by a broad spectrum of situated intelligent agents.
  - self-preservation, goal content integrity, cognitive enhancement, technological perfection, resource acquisition

General/long-term solutions

- Two broad areas of solutions: capability control and motivation control
- Long-term, to realize the advantages of AI, we need to solve motivation control
    - (and capability control is harder than it looks for some plausible systems)

Safety Solutions

- Not using AI where it doesn’t make sense
- Specific responses to specific failure modes
- Human oversight
- “Datasheets for datasets” (Gebru et al., 2018)
- Interpretability
- Verification
- More research needed

AI Security

- Degradation
- Weaponization

Degradation

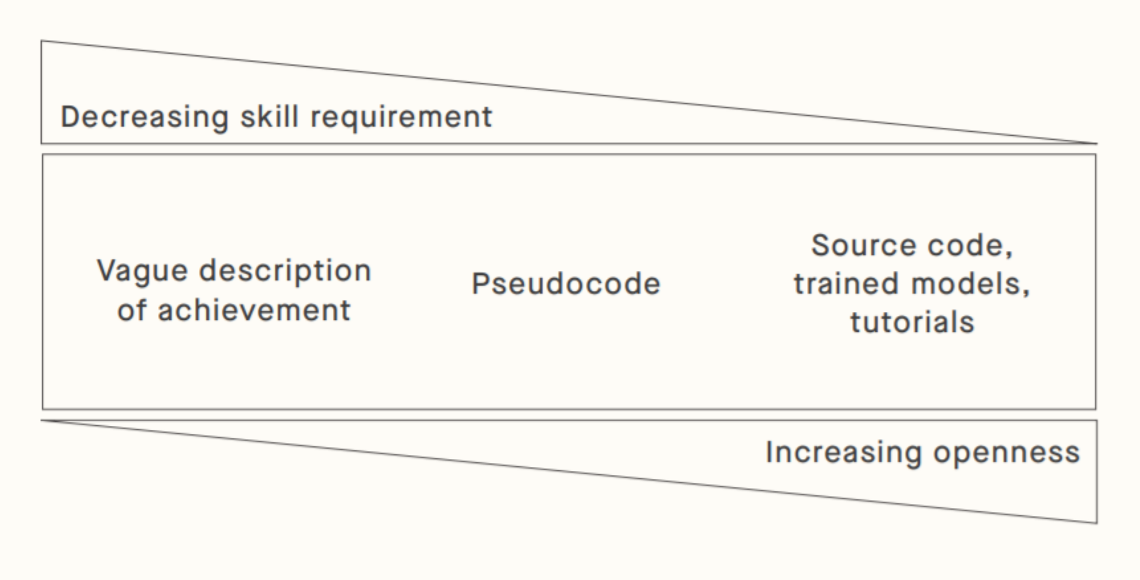
- A rapidly growing set of risks:
  - Adversarial examples/other classes of forced domain shift
  - Data poisoning
  - Neural trojans/backdoors
  - Hardware backdoors
  - Library backdoors
  - Traditional cybersecurity vulnerabilities (e.g. memory overflows)

Weaponization

- For further reading, see “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation,” Brundage et al., 2018, arXiv.

Weaponization

- Drivers of risk
  - Scale
  - Speed
  - Performance
  - Attribution
  - Distance
  - Novel attacks
  - Diffusion
  - Sociality


Decreasing skill requirement
Vague description of achievement
Pseudocode
Source code, trained models, tutorials
Increasing openness

Weaponization

- Classes of risks
  - Digital security
  - Physical security
  - Political security

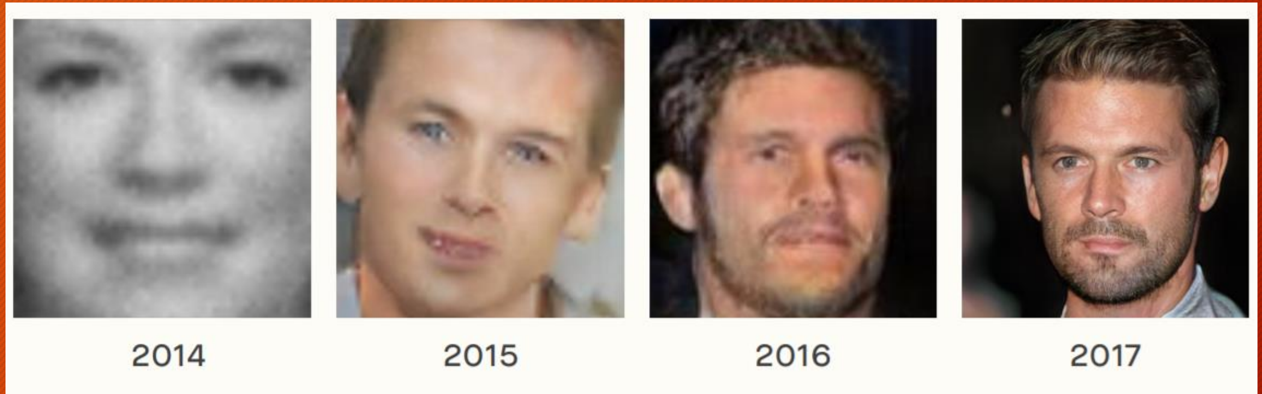


Figure 2: Increasingly realistic synthetic faces generated by variations on Generative Adversarial Networks (GANs). In order, the images are from papers by Goodfellow et al. (2014), Radford et al. (2015), Liu and Tuzel (2016), and Karras et al. (2017).

Security Solutions

- Not using AI where it doesn’t make sense
- Vetting of libraries
- Adversarial input detection (still hard)
- Preventing adversary access
- Trusted hardware
- Responsible disclosure
- Etc. (See Brundage et al., 2018)

The Role of Policy

- AI development as a collective action problem
  - Individual incentives for speed, increasing capability, etc. at the expense of robustness, human oversight
  - *The extent of future tradeoffs is unclear – more to worry about if they're sharp*
  - Corporate to corporate and country to country race dynamics as risks
- Norms and policy as competing tools for alleviating these risks
  - Norms – flexible; easy to implement conditional on will
  - Policy – potential for harmful lock-in; can be forced
  - Optimal combination unclear

Thanks!

- Miles.Brundage@philosophy.ox.ac.uk