*What if we succeed?

Anders Sandberg
Future of Humanity Institute
Oxford University
* Human capital makes the world go around
* Eras: big data, manual dexterity, AGI
* Predictability: within eras good, between lousy
  * Amateurs and experts about as good at AI prediction!
  * Idea-dominated fields unpredictable by nature
* WBE tortoise vs. AI hare
* No reason to think human intelligence is ceiling, nor that transition village idiot to Einstein will be slow
  * Even near human-level AI is potentially a gamechanger
* Intelligence is very powerful
  * Ask the chimpanzees
* The "Omohundro-Yudkowsky hypothesis": generic motivations are unfriendly
  * The safety/motivation problem is hard and understudied
* Hence: Very, very risky, potentially terminal.
  * But benefits can be just as big!

**Best and worst cases**
* General principles
  * Try to get the safety-increasing technologies early
  * Spread risks
* Theoretical research
  * Resolve big uncertainties about intelligence explosions
  * How to transfer human values into code
  * Likely lots of low-hanging fruit in AI safety domain

*What can improve outcomes?*
* Practical research
  * Machine intelligence safety and trustworthiness
  * Better measures of progress
  * Building law abiding machines as good challenge?
  * Adding human detectors as default?

* Practical stuff
  * Useful to have real ethics discussions
  * Improve human retraining

* What can we do now?