

# The Long-Term Future of (Artificial) Intelligence

Stuart Russell

University of California, Berkeley

# Why are we doing AI?

- ❖ To create intelligent systems

# Why are we doing AI?

- ❖ To create intelligent systems
  - ❖ The more intelligent, the better



# Why are we doing AI?

- ❖ To create intelligent systems
  - ❖ The more intelligent, the better
- ❖ We believe we can succeed
  - ❖ Limited only by ingenuity and physics

# Why are we doing AI?

- ❖ To create intelligent systems
  - ❖ The more intelligent, the better
- ❖ To gain a better understanding of human intelligence

# Why are we doing AI?

- ❖ To create intelligent systems
  - ❖ The more intelligent, the better
- ❖ To gain a better understanding of human intelligence
- ❖ To magnify those benefits that flow from it



# Progress is accelerating

- ❖ Solid theoretical foundations
  - ❖ Rational decision making
  - ❖ Statistical learning
  - ❖ Perception, NLP as probabilistic inference
- ❖ Rapid advances
  - ❖ Deep learning in speech, vision, RL
  - ❖ Universal probability languages
  - ❖ Long-term hierarchically structured behavior

# NewScientist

WEEKLY January 29 - February 4, 2011

## THE INTELLIGENCE REVOLUTION

At last something else that thinks like us





# An industry arms race

- ❖ Once performance crosses the usability threshold, small improvements are worth billions
  - ❖ Speech
  - ❖ Text understanding
  - ❖ Object recognition
  - ❖ Automated vehicles
  - ❖ Domestic robots
  - ❖ Intelligent assistants

## 27.3 WHAT IF WE DO SUCCEED?

In David Lodge's *Small World*, the protagonist causes consternation by asking a panel of eminent but contradictory literary theorists the following question: "What if you were right?" None of the theorists seems to have considered this question before. Similar confusion can sometimes be evoked by asking AI researchers, "What if you succeed?" AI is fascinating, and intelligent computers are clearly more useful than unintelligent computers, so why worry?

How should intelligent machines interact with humans? What might happen if intelligent machines decide to work against the best interests of human beings? What if they succeed?

the trends seem not to be too terribly negative.

## 27.3 WHAT IF WE DO SUCCEED?

No one, to our knowledge, has suggested that reducing the planet to a cinder is better than preserving human civilization. Futurists such as Edward Fredkin and Hans Moravec have, however, suggested that once the human race has fulfilled its destiny in bringing into existence entities of higher (and perhaps unlimited) intelligence, its own preservation may seem less important. Something to think about, anyway.



# What if we succeed?

- ❖ *“The first ultraintelligent machine is the last invention that man need ever make.”* I. J. Good, 1965
- ❖ Might help us avoid war and ecological catastrophes, achieve immortality and expand throughout the universe

# What if we succeed?

- ❖ *“An ultraintelligent machine could design even better machines; there would then unquestionably be an ‘intelligence explosion,’ and the intelligence of man would be left far behind. ... It is curious that this point is made so seldom outside of science fiction.” I. J. Good, 1965*

# What if we succeed?

- ❖ Success would be the biggest event in human history
- ❖ It's important that it not be the last



# This needs serious thought

From: Superior Alien Civilization  
<sac12@sirius.canismajor.u>

To: humanity@UN.org

Subject: Contact

**Be warned: we shall arrive in 30-50  
years**

From: humanity@UN.org

To: Superior Alien Civilization  
<sac12@sirius.canismajor.u>

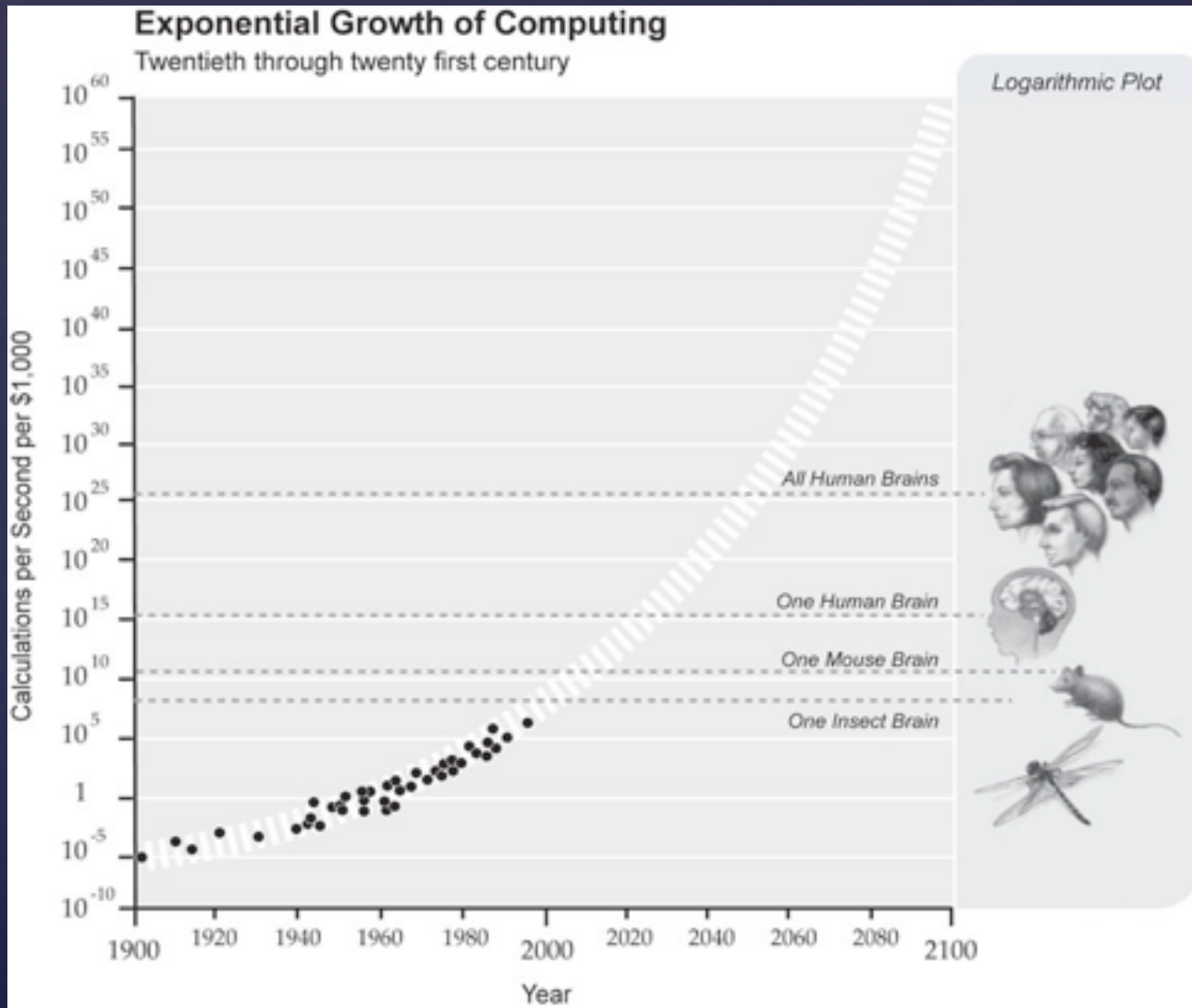
Subject: Out of office: Re: Contact

**Humanity is currently out of the office.  
We will respond to your message when we  
return.**

# Misconceptions: Machines have an IQ

- ❖ *“By 2025 these machines will have an IQ greater than 90% of the U.S. population. That ... would put another 50 million jobs within reach of smart machines.”* Harvard Business Review, 2014
- ❖ Machines will develop along several narrow corridors of ability; general intelligence comes later

# Misconceptions: IQ follows Moore's Law



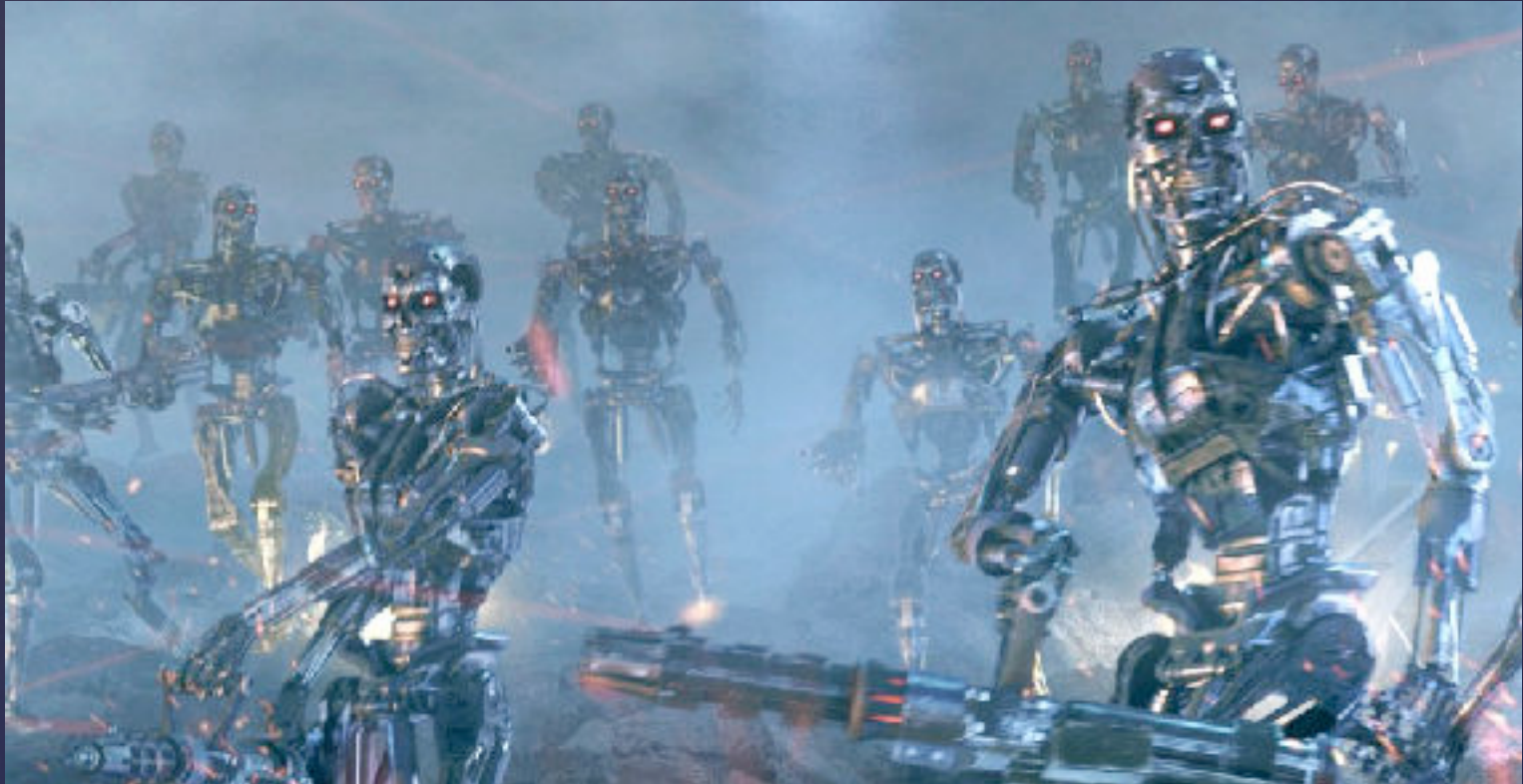


# Misconceptions:

“It’s right around the corner”

- ❖ Few AI researchers believe superintelligence is imminent...
- ❖ ... but breakthroughs are notoriously hard to predict
- ❖ If there is a non-negligible possibility in the medium term, it’s important to address the issue now

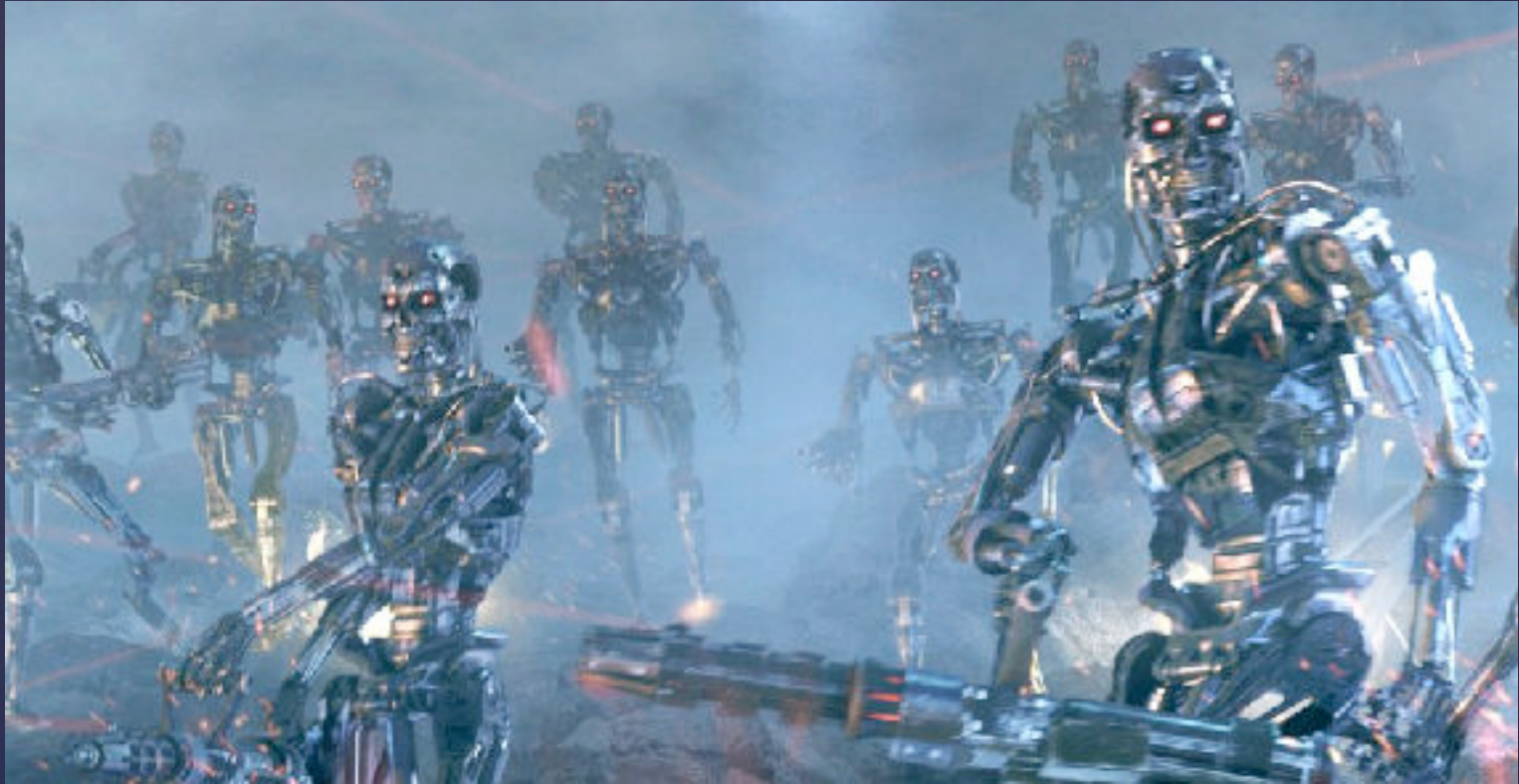
# Misconceptions: Armies of robots



- ❖ An Internet connection more than suffices for impact on humanity – after all, humans do it with words



# Misconceptions: Spontaneous malevolence



- ❖ One need assume only *competent decision making*



# What's bad about better AI?

- ❖ AI that is incredibly good at achieving something other than what we\* really\* want

# Value misalignment

- ❖ E.g., “Calculate pi”, “Make paper clips”, “Cure cancer”
- ❖ Cf. Sorcerer’s Apprentice, King Midas, genie’s three wishes





# Instrumental goals

- ❖ For any primary goal or utility function, the odds of success are improved by
  - 1) Ensuring one's own continued existence and goal integrity
  - 2) Acquiring computational/physical/financial resources
- ❖ With value misalignment, instrumental goals lead to obvious problems for humanity



# Value misalignment contd.

- ❖ If some elements of human values are omitted, an optimal policy often sets those elements to extreme values



# Misuse

- ❖ Primary goals may be aligned with those of a nefarious subgroup
- ❖ This issue is shared with other technologies such as synthetic biology, nuclear fission
- ❖ Emphasizes the need for robust security, possibly counter-AI, regulation

# Unpredictability

Can we guarantee properties for systems that

- 1) think further ahead than we do
- 2) are capable of arbitrary self-modification?



# Proposal

- ❖ Not just AI
- ❖ Provably\* beneficial\* AI
- ❖ Yes, but how?

# Boxing and Oracle AI

- ❖ Sealing the system off from the environment
  - ❖ But not completely!
- ❖ Limiting it to a pure question-answering system (degenerate agent)
  - ❖ Can we have a superintelligent question-answeringer without a metalevel agent directing its computations?

# Stepwise progress

- ❖ Ask a superintelligent verifier whether a given superintelligent agent design is safe before deploying it
  - ❖ Is verification of an agent with decision quality  $X$  easier than making decisions with quality  $X$ ?



# Formal theory of agents

- ❖ Does agent A (objectively) have goal G?
- ❖ Is agent A better than agent B?
- ❖ Can agent A violate condition P?
  - ❖ E.g., modify its own primary goals?
  - ❖ E.g., prevent modification of its goals?

# Value alignment

- ❖ Inverse reinforcement learning: learn a reward function by observing another agent's behavior
  - ❖ Theorems: probably approximately aligned learning
- ❖ Cooperative IRL:
  - ❖ Learn a multiagent reward function whose Nash equilibria optimize the payoff for humans
  - ❖ Broad Bayesian prior for human payoff
  - ❖ Risk-averse agent => cautious exploration
  - ❖ Analyze potential loss (for humans) as a function of error in payoff estimate and agent intelligence

# Value alignment contd.

- ❖ Obvious difficulties:

- ❖ Humans are irrational, inconsistent, weak-willed
- ❖ Values differ across individuals and cultures

- ❖ Reasons for optimism:

- ❖ Vast amounts of evidence for human behavior and human attitudes towards that behavior
- ❖ We need value alignment even for *subintelligent* systems in human environments

=> Moral philosophy as a major industry



# Response 1:

## It'll never happen



Sept 11, 1933: Lord Rutherford addressed BAAS: *“Anyone who looks for a source of power in the transformation of the atoms is talking moonshine.”*



Sept 12, 1933: Leo Szilard invented neutron-induced nuclear chain reaction

*“We switched everything off and went home. That night, there was very little doubt in my mind that the world was headed for grief.”*

# Time is of the essence

- ❖ The sooner we start solving the problem of control, the easier it will be
  - ❖ Commercial and military momentum will only increase
  - ❖ It takes time to develop community standards and conceptual framework
  - ❖ It takes even more time to enact a global regulatory framework (if it's needed)

Response 2:

It's too late to stop it



# Response 3:

## You can't control research

- ❖ Asilomar Workshop (1975): self-imposed restrictions on recombinant DNA experiments
- ❖ Industry adherence enforced by FDA ban on human germline modification
- ❖ 2010 US Presidential Commission: federal oversight of synthetic biology research
- ❖ Pervasive\* culture of risk analysis and awareness of societal consequences

# Response 3:

## You can't control research

**theguardian**

[News](#) | [Sport](#) | [Comment](#) | [Culture](#) | [Business](#) | [Money](#) | [Life & style](#)


[News](#) > [Technology](#) > [Self-driving cars](#)


### US needs a Federal Robotics Commission, says think tank

The US needs a federal agency to deal with the regulation and ethical challenges of robots and artificial intelligence, says influential think tank

---

**Alex Hern**

 Follow @alexhern

 Follow @guardiantech

theguardian.com, Wednesday 17 September 2014 07.00 BST

# Response 4:

## You're just Luddites!!

- ❖ The goal is not to stop AI research
- ❖ The idea is to *allow it to continue* by ensuring that outcomes are beneficial
- ❖ Solving this problem should be an intrinsic part of the field, just as containment is a part of fusion research
- ❖ *It isn't "Ethics of AI", it's common sense!*



# Summary

- ❖ The AI community may be running blindfolded towards the biggest event in human history
- ❖ If so we need a fundamental change in the way the field defines itself