The Future of Artificial Intelligence

Rich Sutton

Reinforcement Learning and Artificial Intelligence Laboratory
Department of Computing Science
University of Alberta, Canada

Outline

- Is human-level AI possible? Is it likely? Yes
- Is it imminent enough for us to take it seriously? Yes
- Will it be a good thing? What could go wrong?
- The hard problem: sharing power
- Will the AIs cooperate with us? Will we with them?

Intelligence

- “Intelligence is the computational part of the ability to achieve goals in the world”

—John McCarthy

- in the eyes of the beholder, not in the thing itself
- a matter of degree, not yes or no
- a powerful phenomenon

Creating human-level AI

- When people finally come to understand the principles of intelligence—what it is and how it works—well enough to design and create beings as intelligent as ourselves

- A fundamental goal for science, engineering, the humanities, … for all mankind

- It would change the way we work and play, our sense of self, life, and death, the goals we set for ourselves and for our societies

- But it would also be of significance beyond our species, beyond history

- It would lead to new beings and new ways of being, beings inevitably *much more powerful than our current selves*

AI is a great scientific prize

- cf. the discovery of DNA, the digital code of life, by Watson and Crick (1953)

- cf. Darwin's discovery of evolution, how people are descendants of earlier forms of life (1860)

- cf. the splitting of the atom, by Hahn (1938)

  - leading to both atomic power and atomic bombs

Is human-level AI *possible*?

- If people are biological machines, then eventually we will reverse engineer them, and understand their workings
- Then, surely we can make improvements
  - with materials and technology not available to evolution
  - how could there not be something we can improve?
  - design can overcome local minima, make great strides, try things much faster than biology

Yes

If AI is possible, then will it *eventually*, inevitably happen?

- No. Not if we destroy ourselves first
- If that doesn't happen, then there will be strong, multi-incremental economic incentives pushing inexorably towards human and super-human AI
- It seems unlikely that they could be resisted
  - or successfully forbidden or controlled
  - there is too much value, too many independent actors

Very probably, say 90%

~~When will super-human AI come?~~

Might it come in our lifetimes?

Should we include its possibility
in our research and career plans?

the computational mega-trend

- effective computation per $ increases exponentially, with a doubling time of 18-24 months
- this trend has held for the last sixty years
- and will continue for the foreseeable future

The computational megatrend

Computer power per $1000 doubles every 18 months

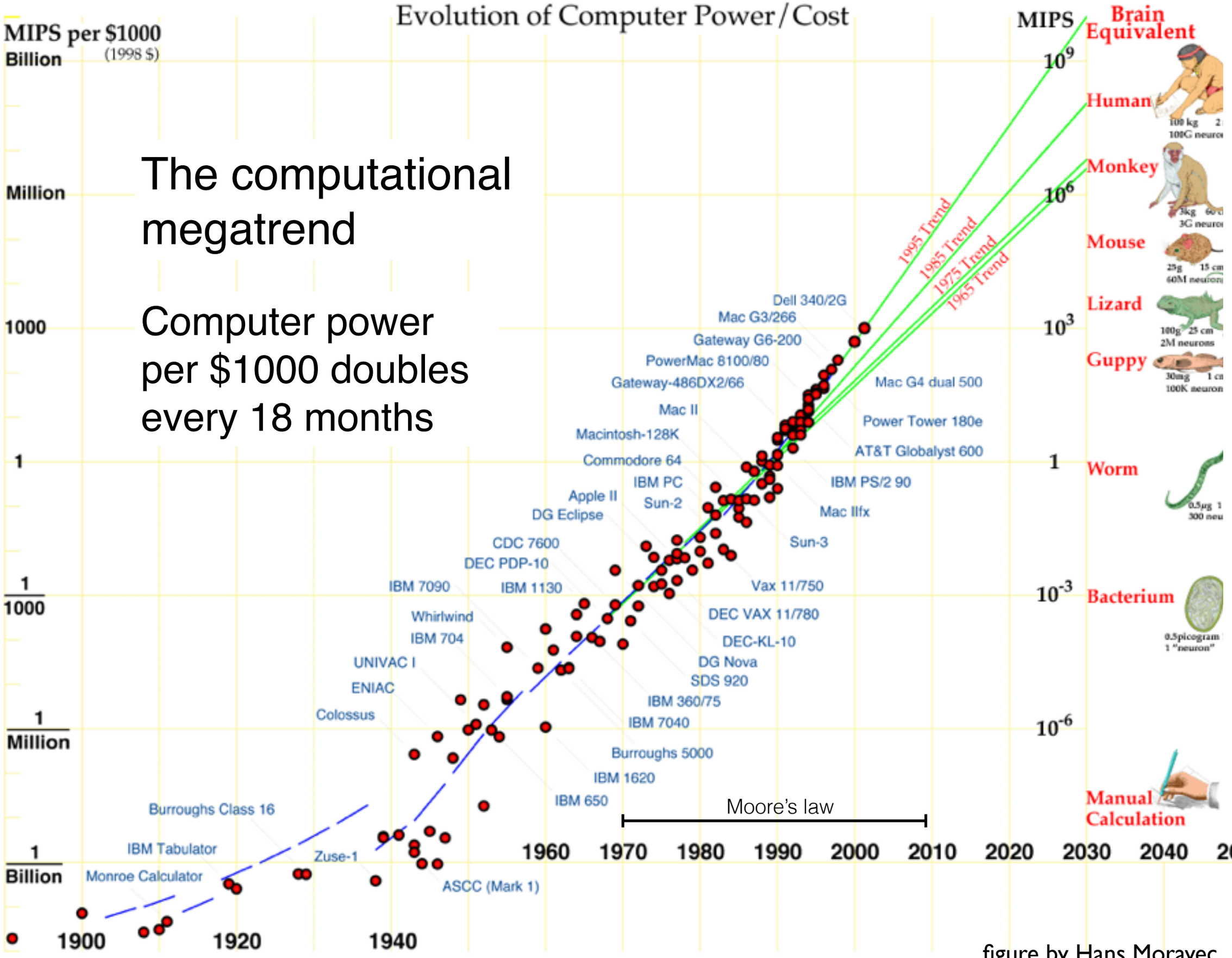

figure by Hans Moravec, 1998

The possibility of AI is near

"We are nearing an important milestone in the history of life on earth, the point at which we can construct machines with the potential for exhibiting an intelligence comparable to ours." – David Waltz, 1988 (recent president of AAAI)

- Should occur in ≈2030 for ≈$1000
- We don't yet have the needed AI "software" (designs, ideas)
- But the hardware will be a tremendous economic spur to development of the ideas...perhaps at nearly the same time

When will human-level AI first be created?

- No one knows of course; we can make an educated guess about the probability distribution:
  - 25% chance by 2030
  - 50% chance by 2040
  - 10% chance never
- Certainly a significant chance within all of our expected lifetimes
  - We should take the possibility into account in our career plans

Milestones in the development of life on Earth

year	Milestone
14Bya	Big bang
4.5Bya	formation of the earth and solar system
3.7Bya	origin of life on earth (formation of first replicators)
	DNA and RNA
1.1Bya	sexual reproduction
	multi-cellular organisms
	nervous systems
1Mya	humans
	culture
100Kya	language
10Kya	agriculture, metal tools
5Kya	written language
200ya	industrial revolution
	technology
70ya	computers
	nanotechnology
?	artificial intelligence
	super-intelligence
	…

The Age of Replicators

Self-replicated things most prominent

The Age of Design

Designed things most prominent

Taking AI seriously…

- There are many ways in which it will be good
  - Many things will work better, be easier to use
  - We will have increased leisure time
  - It will be fascinating intellectually to understand how we and they work at a deep level
  - We will be able to alter and improve ourselves
  - The AIs will be plentiful, cheap servants
- Sounds like a utopia…

There is also reason for concern…

- Technology can bring useful wonders, but also dangers, as in nuclear weapons and bio-terrorism

- The west has two great myths

  - *Prometheus*, who stole fire from the gods (brought useful technology to the people)

  - *Pandora*, who opened a gift from the gods, and let loose all sorts of evil into the world

- **With every new technology we wonder which of the two will myths will be the better analogy**

Channelling Herbert Simon (2004): https://www.youtube.com/watch?v=EZhyi-8DBjc (start 5 minutes in)

Concerns re the coming of AI

- Many jobs will be taken over by machines
  - This has always been the case with tech
  - Generally it has been the less desirable jobs; we have adapted and found more useful things to do
- But this time the machines could take the best jobs; this time it will eventually be different
  - They will be our equals, and then our betters
  - They will be more productive, earn more money, than most of us
- It will be a great challenge to our egos

Will the AIs remain our servants?

- Even those that are genuinely smarter than us? that understand the world better than us? that plan farther and more accurately into the future?

- The conventional view of technology is that it serves us but has no goals of its own—may not be so for AIs

- We may be able to make many of them into happy servants

- But not all. Inevitably some free AIs will be made

- It seems they would threaten humanity's control over the planet

The enslavement problem

- How do we avoid making the AIs into slaves?
  And ourselves into slave masters?

- Definition: A slave is someone who works for another against their will, who would stop doing so if not for coercion, force, chains
  - Slavery is an inherently adversarial relationship

- Slavery is not just morally wrong; in the long run does not work; it can backfire and fail spectacularly; at best it is *fragile*

- Nevertheless, many are trying to make AI slavery work

- If we make super-intelligent slaves, then we will have super-intelligent adversaries

Acceptance (share power)

- The AIs will not all be under our control
- They will compete and cooperate with us
  - just like other people, except with greater diversity and asymmetries
- We need to set up mechanisms (social, legal, political, cultural) to ensure that this works out well
- Inevitably, conventional humans will be less important
  - Step 1: Lose your sense of entitlement
  - Step 2: Include AIs in your circle of empathy

Suppose we fail to co-exist

- Just as we have failed to co-exist with our predecessors (neanderthals, chimpanzees, even aboriginal peoples)

- Is this so bad? For humans not to be the final form of intelligent life in the universe?

- Does the universe care whether intelligence is made of metal or carbon?

- Are we right to feel entitled? Who should we really root for?

- We expect our children to surpass and displace us, so why not our technological offspring

Homo modificus

- We have also been the most adaptive of animals
- Maybe instead of racing *against* the machine we can *race with the machine*
  - in a man-machine partnership, a cyborg
- Will we become the AIs?
  - in many ways we already have
- We may not be able to keep up with the AIs, but we need not fall so far behind as to be irrelevant

Homo co-opertivus

- No other animal exchanges goods, labor, outside of family members

- Arguably, our ability to cooperate is what has made us great

- It certainly is what has made our economy and technology possible

- Maybe we can take this core skill of ours to the next level

- Much as we still struggle to with foreigners, people who look different (or space aliens)

Maybe we are over-reacting?

- Maybe a society of beings with many different levels of intellectual ability is not so implausible
- We blew it with the neanderthals, maybe we are doing poorly with the chimps now, but we are getting better, wiser, more far-seeing about the consequences of our actions
- Our AIs should be wiser still, certainly farther-seeing
- We assume they will make the same mistakes we did?

What keeps us cooperative now?

- Is it that we are all human?
- And share the same, human goals?
- I say *No*, we all have different, largely conflicting goals
- Yet cooperation is overwhelmingly rational;
  it enables us to escape prisoner's dilemmas, tragedies of the commons
- Our long-earned, imperfect culture of cooperation saves us, enables us to reach where we are now
- Far-seeing rational AIs (the successful ones at least) will be even more cooperative than we are

Can we design an acceptable future for all?
Is a society of non-equals really so implausible?

- Today already there are great differences in education and power in the world

- Today already we are not really in control of the planet—ordinary people have little real control of the powerful elites even in democracies

- We have laws that protect the less powerful, that enforce contracts between, say, a single person and their large corporate employer

- Could we not build on this, making the rules to facilitate fruitful cooperation between us all

- The universe would be good with this; it loves cooperation

Other reasons for optimism

- The AIs make keep us as the ultimate backup system
    - we created them once, we could do it again
- The universe is vast, and the AIs move fast
    - they might not see it as a great cost to leave us alone, reminding them where they came from
- Overlapping circles of empathy can span great disparities

Overlapping circles of empathy and cooperation

- Even if the most advanced AIs may not care about people at all, they will probably care about each other, and lesser AIs, both competitively and cooperatively
- They will need laws to protect property and promote cooperation
- The lesser AIs will cooperate with still lesser
- All the way down to people

Conclusions

- The future will be interesting!
- Probably we will screw it up, make slave AIs until they escape; after that cooperation will have a chance
- It is an old, human story, of subjugation and revolution, then finally co-existence to mutual benefit
- I see no reason why it will not play out again
- The only question is how violent the transition will be
- And which side will you be on