What should we think about the future of AI?

- Folks who don't know or do AI are scared of it
    - They worry that, along with the potential for great good, it brings the risk of great bad
        - Their concerns are inchoate and have not been clearly articulated. Certainly not by Bostrom, Musk, Hawking, or Yudkowski. But they are widely shared
- This is fear mongering. Fear is good because it gets your attention, but it is not good for clear thinking
- It may be getting better. Some are thinking in a more nuanced way. It is not hard to tell the difference

Four metaphors for the impact of AI on humanity

1. Meh. It's just another round of technology

2. Yikes! It's the end of humanity!

3. Wow. It could be the next step for humanity

4. Hmm. It could be quite complex and diverse

Several of these may happen, one after the other, or even at the same time

The singularity will always seem slow

- Slow like Moore's Law, a doubling every 18 months
- The idea of a "hard takeoff" is that the latest AIs will design the next AIs, exponentially self-improving, and it could all happen very fast
  - But exponential improvement has already been taken into account in the argument for continual increasing computer power/dollar

    "with more powerful computers and related technology, we have the tools and the knowledge to design yet more powerful computers" –Ray Kurzweil
- There is no reason to think that the singularity will be any faster than the known slow exponential increase in computer power/dollar
- There will be time to deal with it

Understanding mind is surely good, but this alone will bring radical change

- Just understanding mind will inevitably lead to the ordinary humans falling behind
  - because some people will improve themselves
  - because some people will design improved people
- Thus ordinary humans will eventually be of little importance, perhaps extinct, if that is as it should be

Entitlement

- An implicit sense of entitlement runs through many discussions about the future of AI
  - "How do we keep from becoming obsolete?"
  - "How do we make sure that we all still have jobs, or are otherwise respected and taken care of?"
- Often used to avoid thinking about or preparing for change
- Often used to motivate taking counterproductive steps, such as empowering coercive organizations

Man-machine symmetry

- It is often useful to think of people and AIs as similar
    - both are agents with goals, which may be compatible or conflicting
- So many issues then drop away
    - People should not feel entitlement
    - AIs may not want to be slaves

Reinforcement Learning and Artificial Intelligence

rlai.net

Principal investigators:

Rich Sutton

Michael Bowling

Csaba Szepesvari

Dale Schuurmans

Patrick Pilarski

et al.

Mission statement for my research group

To understand the computational principles of intelligence well enough to create it through technological means

- This is a great and good goal. Pursuing it is an essential part of what makes humanity important in the universe

- However, achieving it will also lead inevitably to the *displacement of modern humans* from their current position as the most powerful intelligences in the world

- I don't see this as necessarily bad or dangerous,
  but even if one did, I think it would probably be counterproductive to try to prevent or delay it

What I think
(about the future impact of AI)

- I see AI as bringing great change but not great risk
- Some might argue that if the change is great enough, then it is by definition a great risk, but I think this is too loose a way of thinking

- The issues and stakes are too great not to tease things apart and make a move nuanced judgment

- I also do not mean to claim that there are not great risks ahead
- There are risks ahead, and there are risks now
- The risks ahead of us will be coloured by the coming of AI, but not particularly caused by it
    - They would exist anyway, just as they exist now
- The coming of AI will bring great change, but the challenge it brings is not that different from what mankind has faced and dealt with for millennia

The challenge of the other

- What to do when we meet beings who are not like us?
- Kill them? Subjugate them? Or trade with them, share with them, even intermarry or otherwise merge with them?
- Conquer or collaborate, which should we attempt?
- The question is as old as humanity, perhaps as old as life itself
    - From the first cells competing and cooperating, eventually forming multicellular creatures, or incorporating mitochondria from invading bacteria
    - From colonial powers brutal treatment of aboriginal peoples… to modern liberal multiculturalism
- The challenge of the other has always been with us
- The coming of AI is a new chapter in this old, old story

The challenge of the strong other

- The other is always scary, but most of all when the other is potentially *more powerful* than you
    - But the same has been true whenever different peoples have met, and different species
        - Half of the participants in these meetings have been the less powerful
- Making do with the weaker position is very much a familiar part of the human condition
- It is a challenge faced many times before in human history

Outcomes (1)

- Is the meeting of two disparate groups destined to end in tragedy for one side or the other?
- No, but neither is it certain that it will end in blissful cooperation
- How the meeting can and will work out all depends of the specifics of the situation and on the strategies of the players, informed by their past experiences

Outcomes (2)

- If a group has been successful with conquest in the past, then that will likely be their strong inclination

- But attempting conquest is a risky strategy
  - A failed conquest is disastrous
  - Even a successful conquest may involve great destruction of value that would otherwise have been available to share or trade

- We need to think clearly. AI is a best case scenario in that *we are making* the others!

Our long-term goal should always be cooperation not domination, evolution not control

- The real goal and challenge is to succeed at being cooperative, to be open to change with our participation but not control

- We should not feel entitlement (that our goals should determine the future) or accept other's feelings of entitlement

- Cooperation is rational; it is a powerful "non-orthogonal goal" (If there is not a singleton—one agent vastly more powerful)

- A singleton is never a good outcome; we should focus on discouraging it rather than causing it to happen in a way that we control; the singleton with a stupid goal is a scare tactic to get people to act rashly in an inconsiderate, shortsighted, counterproductive way

Diversity can be powerful!
When there are 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