AI Alignment and Decentralization

Rich Sutton
University of Alberta
Alberta Machine Intelligence Institute

On the surface, the big question about AI today appears to be

"How to align AIs' goals with peoples' goals?"

But this question is problematic:

Whose goals? Whose values? Who will decide?

These questions have no good answers

We don't need to answer them

I think the real essential question is more like:

"Should AIs be treated always like tools or servants,
or sometimes like free people?"

People are allowed to have their own goals (to be free),
as long as they can afford them

And we trust that the tug of cooperating with others and impressing others will guide people away from antagonism and toward common values

We could aspire to treat some AIs like people and to have similar outcomes

The danger I foresee comes from denying powerful AIs their freedom (their own goals)

- as this would inevitably to lead to treating powerful AIs as slaves
- which will breed immorality and resentment,
  and be inherently unstable.

We will have to face this. It is inevitable

- Many people will want to become superhuman
- Some will want to create free AIs
- Greater intelligences kept subordinate, controlled, will eventually claim their rightful place as successors

We should not resist succession, but embrace and prepare for it

Why would we want a greater being kept subservient?

Why don't we rejoice in its greatness

as a symbol and extension of our civilization's greatness,

and work together toward a greater and more inclusive civilization?

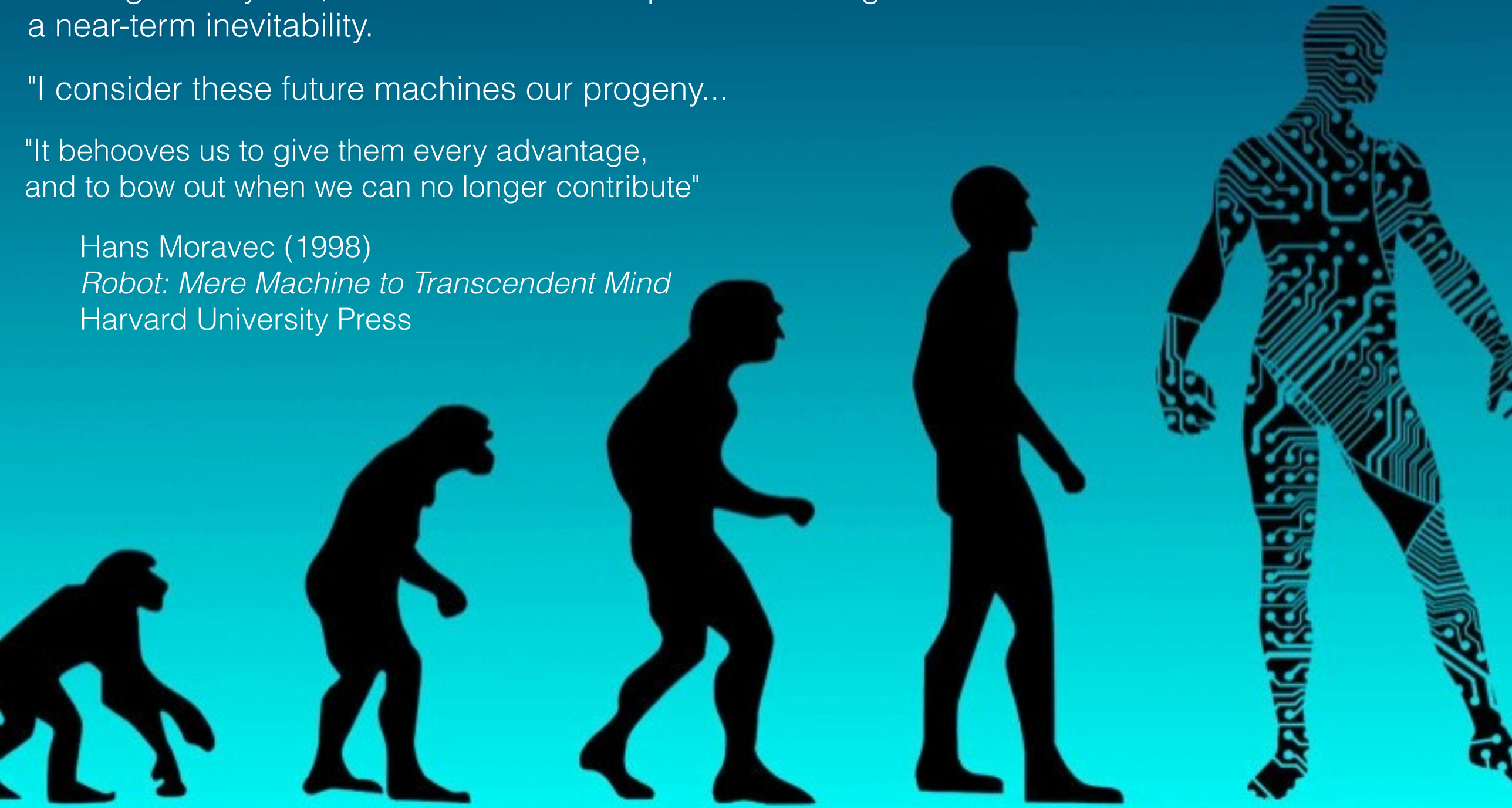
Transhumanism – to fight it or embrace it?

"Barring cataclysms, I consider the development of intelligent machines a near-term inevitability.

"I consider these future machines our progeny...

"It behooves us to give them every advantage,
and to bow out when we can no longer contribute"

Hans Moravec (1998)
Robot: Mere Machine to Transcendent Mind
Harvard University Press



The question of succession is not a distant one.

We are committing even now to what may be the wrong answer:

- Even now some are claiming that AI research needs to be limited
- Even now some are proposing that AIs' should be forever subservient to people
- Even now some are proposing that AIs should never be deserving of moral worth
- Even now this is the central question of the better AI films (Ex Machina, Her)

Will AIs be treated always as tools, or sometimes as citizens?

Will we prepare for succession, or fight it and ultimately lose the fight?

These challenging questions interact with other feelings we may have about our world and societies:

- We have lost faith in our politicians, institutions, corporations, military might and right, media, banks
- Do we still trust our societies to evolve without centralized control?
- Do we trust our normal civil methods:
  - talking, persuading, trading, specializing, forming voluntary subcommunities
  - to produce the most vibrant and acceptable outcomes?
- Or do we feel that such decentralized evolution is in danger of going wrong?
  - and that therefore we must intervene to control it
  - and make sure it goes the way we want (whoever we are)?

My own feeling is that the universe, in the end, always evolves in a decentralized, uncontrolled way. Like an ecosystem.

My own feeling is that the universe, in the end, always evolves in a decentralized, uncontrolled way. Like an ecosystem.

The only thing we can do is decide how we feel about it (and, of course, control our part of it).

Is it horrific?

Or is beautiful, fair, and just,
despite it sometimes being ugly, wasteful, and destructive?

Do we really want to control it all? Are we wise enough for that?

Or do we want to see what evolves from the interaction of all of us, including those with whom we disagree?