The Race Between Human and Artificial Intelligence

Anton Korinek (Johns Hopkins and NBER)

INET/IMF Conference on "The Macroeconomics of AI"

April 2018

Motivation

Intelligence = the ability to accomplish complex goals

(Tegmark, 2017) .

Motivation

Rapid Advances in Artificial Intelligence:

- imply that machines & computer programs behave more and more like *artificially intelligent agents (AIAs)*
  - determine increasing number of corporate decisions, e.g. screening of applicants for jobs, loans, etc.
  - influence (manipulate) growing number of human decisions, e.g. what we read, watch, buy, drive, like, vote, think, ...
  - act autonomously, e.g. trading in financial markets, driving cars, screening applicants, playing Go, composing music, ...
- continue unabated
- will have profound implications if AIAs reach and surpass human levels of general intelligence

Motivation

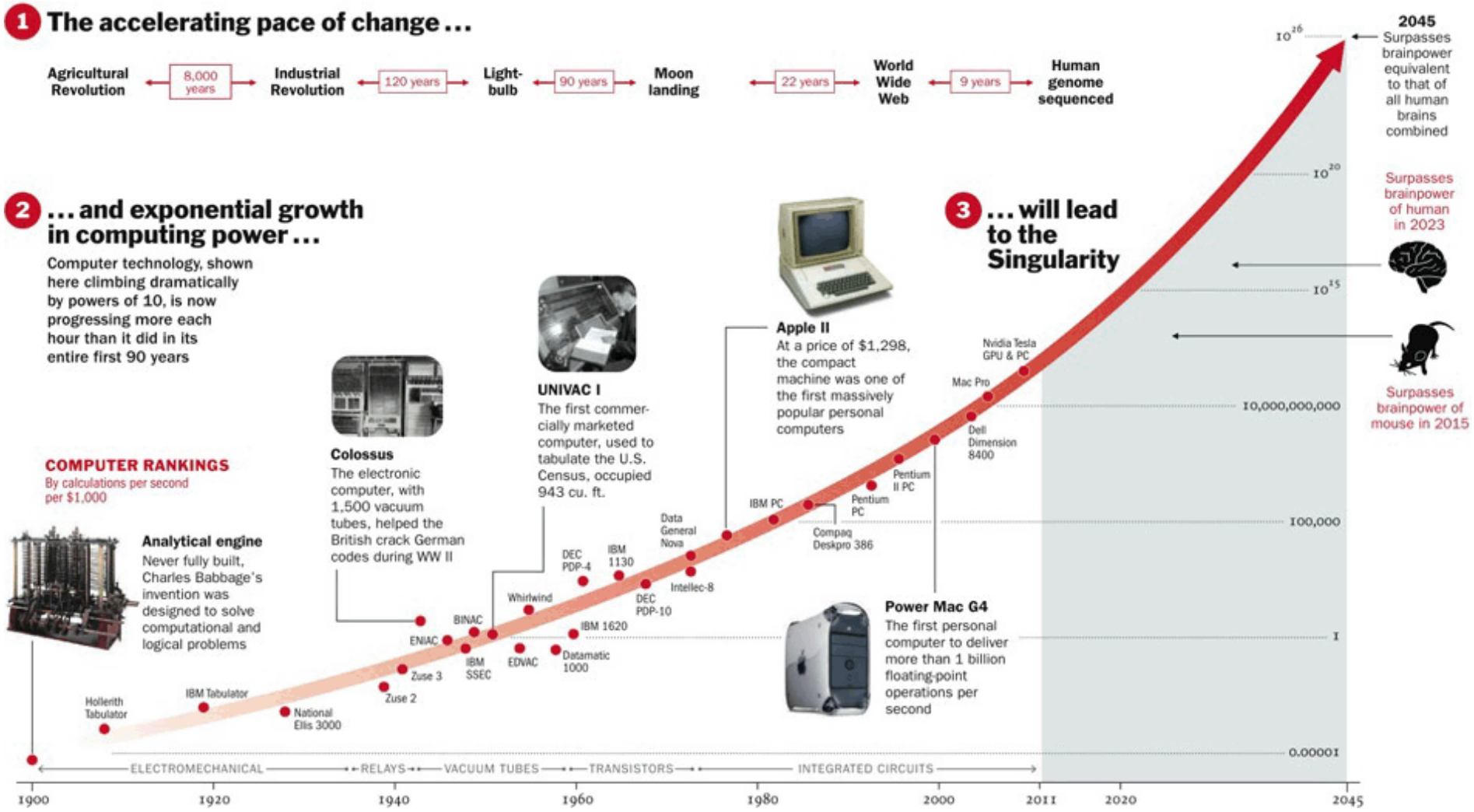


Figure: Moore's Law and Brainpower

Thought Experiment

Consider an observer from another galaxy who arrives on earth:

- encounter humans and machines busily interacting with each other
  - Are the humans controlling the machines?
  - Or are they controlled by the little black boxes that they carry around and constantly check?
  - And who controls the little black boxes?

... just one example of the blurring lines about who is in charge

Key Questions

- What are the implications of new forms of intelligence rivaling humans?
- What determines the allocation of resources between humans and AIAs?
- If there is a race between humans and AIAs, what factors drive the outcome? (Does the economy need humans?)
- Are there hints of AIAs in our present economy?

Note: economics at its heart is about the allocation of scarce resources
$\rightarrow$ well-positioned to answer these questions

Key Contributions

1. Novel framework that expands concept of agency to AIAs
2. Analyze factors that determine the distribution of resources
3. Characterize factors that determine the outcome of the race between humans and AIAs
4. Present a few (naive?) policy proposals

Classical (Anthropocentric) Economics

Humans = Agents	**Machines = Objects**
• absorb consumption expenditure • supply labor services • behavior encoded in preferences • evolve according to law of motion (e.g. constant n)	• absorb investment expenditure • supply capital services • behavior encoded in technology • evolve according to law of motion

Novel Symmetric Perspective on Humans and AIAs

Humans, machines and other AIAs $i \in \mathcal{I} = \{h, m, \dots\}$ are ~~agents, objects,~~ entities that

1. absorb resources x^i to maintain, improve and/or proliferate (can be viewed as "consumption" or "investment")
2. supply a factor endowment ℓ^i per entity, fixed in baseline, generalized in appendix (can be "human labor" or "machine services" etc.)
3. evolve according to a law-of-motion

$$N^{i\prime} = G^i(\cdot)\, N^i$$

with growth that is given by a (possibly degenerate) function $G^i(\cdot)$

Symmetric Perspective in the Data

Income and Spending in NIPA (2017Q3 Annualized):

- on national income side:

Gross national product	\$19.7tn	100%
National (human) income	\$16.7tn	85%
Consumption of fixed capital	\$3.0tn	15%

- on domestic spending side:

Gross domestic product	\$19.5tn	100%
Human absorption (consumption)	\$13.4tn	69%
Machine absorption (investment)	\$3.2tn	16%
Shared (government)	\$3.4tn	17%

Scenarios of AIAs

Three scenarios of artificially intelligent agents:

- Scenario 1: collective entities, e.g. corporations, will increasingly act as super-intelligent entities [e.g. algorithms at Facebook, Google, etc. controlling our behavior]
  - absorbing growing amounts of resources to maintain and improve themselves
  - accumulating growing amounts of wealth
  - with shareholders having very limited control rights
- Scenario 2: human enhancements will provide some humans with far superior intelligence
  - expenditure to maintain/improve humans absorbing a growing amount of resources
    - harbingers already present in current economy but technological limits
  - rapid progress in bio- and nano-technology
  - richest humans increasingly able to translate wealth into superior physical and mental properties
    (Yuval Harari: the "gods" and the "useless")
- Scenario 3: intelligent computer systems will become super-intelligent
  - well-known scenario from science fiction (esp. in Austria)

General Model Setup

- Time: discrete $t = 0, 1, ...$
- Entities: described by set $\mathcal{I}$ of size $I = |\mathcal{I}|$, indexed by i, e.g. $\mathcal{I} = \{h, m\}$, counted in terms of efficiency units N_t^i
- Factors:
  - endogenous factors $L_t^i = \ell^i N_t^i$ supplied by entities in set $\mathcal{I}$, e.g. human/machine labor
  - exogenous factors T in fixed supply, e.g. land, energy
- Goods: $j = 1...J$ consumption goods, e.g. simplest case: $J = 1$
- Production possibilities: $Y_t \in F_t\left(\left\{L_t^i\right\}, T\right)$, e.g. $Y_t = F\left(L_t^h, L_t^m, T\right)$
- Aggregate absorption: $X_t^i = x_t^i N_t^i$ for each type $i \in \mathcal{I}$

- Market clearing:

$$\sum_{i \in \mathcal{I}} X_t^i = Y_t \in F_t\left(\left\{L_t^i\right\}_{i \in \mathcal{I}}, T\right)$$

Examples: Neoclassical Economy

Example: interpret traditional neoclassical economies through lens of our model

Setup:

- two scarce factors: humans H and traditional capital K
- law-of-motion for capital: $N^{k\prime} = (1-\delta)\, N^k + X^k$

Example 1 (simplest models of population):

- representative agent $N^h \equiv 1$ or exogenous population $N_t^h = (1+n)^t$

Example 2 (human capital view):

- N^h measures efficiency units of human capital: $N^{h\prime} = G^h\left(x^h\right) \cdot N^h$
- we spend a great deal of resources x^h on increasing efficiency units per physical unit of human
  $\rightarrow$ e.g. fastest growth sectors in recent decades: education, healthcare, ...

Example 3 (Malthusian view – most relevant in LDCs):

- $N^{h\prime} = \min\left\{1, {}^{x^h}/_{s^h}\right\} \cdot (1+n)\, N^h$ where s^h is human subsistence income

$\rightarrow$ population may be limited by subsistence

Resource Absorption Frontier

Definition (Maintenance absorption)

= set of absorption levels s^i s.t. $G\left(s^i\right) = 1$

For the following concept, focus on stationary economies (no steady state growth):

Definition (Resource Absorption Frontier)

= set of efficient steady state numbers $\left(N^h, N^m\right)$ and absorption levels $\left(X^h, X^m\right)$ for given exogenous factors T, i.e. for which

$$X^h + X^m \in F\left(\ell^h N^h, \ell^m N^m, T\right) \qquad \text{with} \qquad G^i\left(X^i/N^i\right) = 1 \forall i$$

Note: in models of steady state growth, we can define an analogous *Normalized Absorption Frontier*

Preferences and Behavior

Note: so far, everything is described without preferences
[humans and machines are algorithmic automata – kind of like in macro models]

Choices to be made:

- how to allocate factors to production of output
- how to allocate output to absorption of different entities

Approaches:

- describe behavior as maximizing a utility function $u^i\left(x^i\right)$

or – almost isomorphically –

- describe behavior by the resulting behavioral rules $x^i\left(\cdot\right)$
  (for machines, this is the less contentious approach, but it's no different!)

Preferences and Behavior

How can AIAs possibly acquire "preferences"?
(question is a red herring, since they will certainly exhibit behavior)

$\rightarrow$ obvious in scenarios 1 (corporations) and 2 (enhanced humans)

In scenario 3:

Claim (Instrumental convergence: Omohundro, 2008; Bostrom, 2014)

No matter what its final goals are, a sufficiently intelligent entity automatically pursues a set of instrumental goals that are useful in the pursuit of its final goal(s):

- self-preservation
- goal-content integrity
- self-improvement
- unbounded resource accumulation

Note: this looks a lot like what (other) living beings do

Preferences and Behavior

Definition (Growth-optimal preferences)

We call preferences U^i over aggregate consumption plan $\left(X_t^i\right)_t$ and the associated behavioral rules *growth-optimal* for type i entities iff they are a strictly monotonic transformation of

$$U^i\left(\left(X_t^i\right)_t\right) = \lim_{t\to\infty} N_t^i = N_0^i \prod_{t=0}^{\infty} G\left(x_t^i\right)$$

If preferences (behavior) are not growth-optimal, we call them *mis-matched.*

Examples of mis-matched preferences:

- over-eating
- use of contraception
- ...

Observation: if entities have mis-matched preferences, they remain inside the resource absorption frontier
(but not a problem for species as long as there is no competition)

Example 1: Human-Replacing AIAs

Example 1: characterize Absorption Frontier between humans h and AIAs m
$\rightarrow$ first illustration of interactions of humans/AIAs

Setup:

- single exogenous factor "land" $T = 1$
- single consumption good
  $\rightarrow X^h, X^m, Y$ are scalars
  $\rightarrow$ maintenance absorption $s^i = \left(G^i\right)^{-1}(1)$ in steady state is scalar
- per-unit factor supplies denoted by $\ell^i \equiv A^i$
- capture "human-replacing" element of machine labor by Cobb-Douglas production with additive human and machine labor

$$Y = T^{\alpha}\left(A^h N^h + A^m N^m\right)^{1-\alpha}$$

$\rightarrow$ (i) describe steady states
$\rightarrow$ (ii) describe transition after shocks

Example 1: Maximum Absorption for Humans

Characterizing the Resource Absorption Frontier: start with corners

- define by $\bar{N}^h$ the steady-state level of humans when there are no machines so $s^h \bar{N}^h = \left(A^h \bar{N}^h\right)^{1-\alpha}$
- define by $\bar{N}^m$ the steady-state level of machines when there are no humans

Proposition (Maximum Absorption for Humans)

1. **Human-only economy:** *if*

$$(1-\alpha)\frac{A^m}{s^m} < \frac{A^h}{s^h}$$

then maximum absorption entails $\bar{N}^h$ *humans and* $N^m = 0$ *machines (intuition:* $MPL^m < s^m$*)*

2. **Human economy with symbiotic machines:** *otherwise the human maximum entails* $N^h > \bar{N}^h$ *humans and* $N^m > 0$ *machines*

Example 1: Maximum Absorption for Humans

Humans and machines as a function of machine productivity

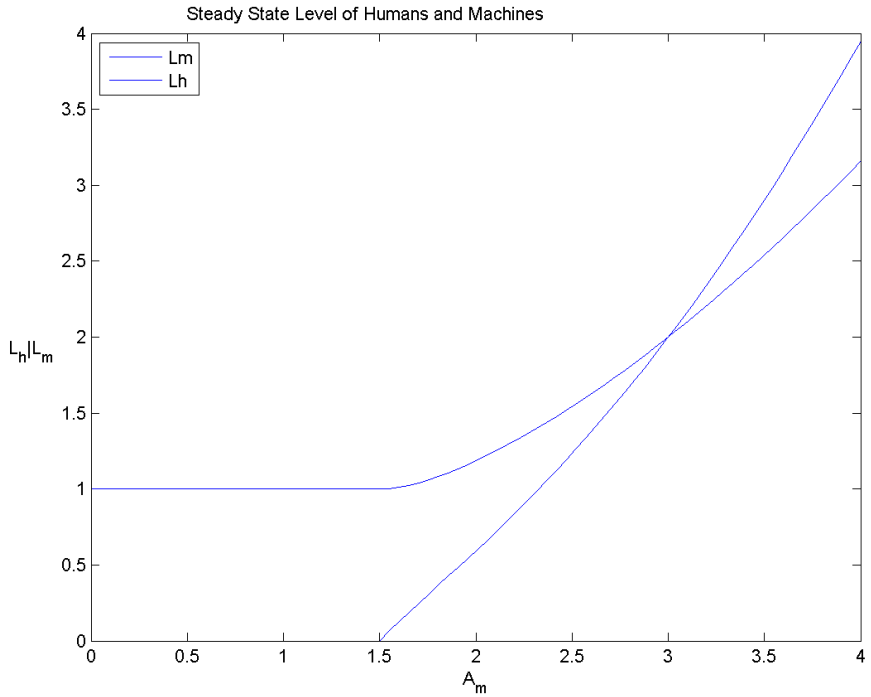
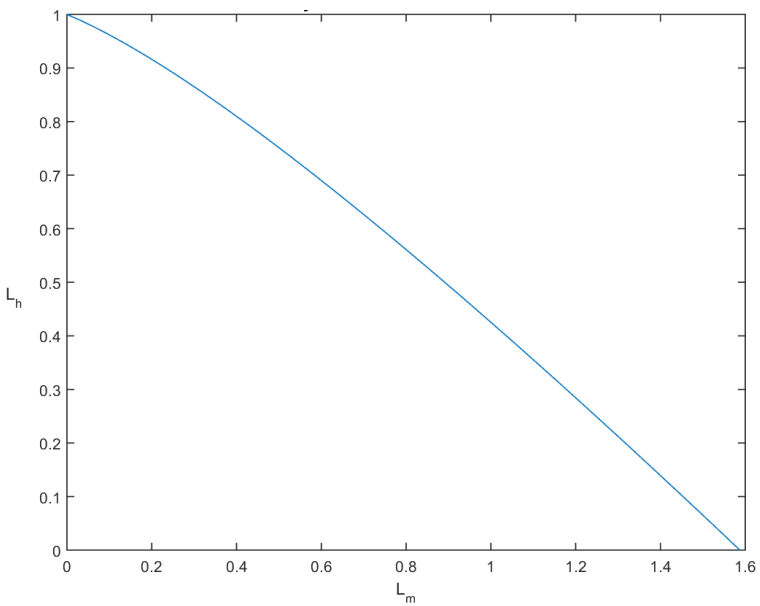


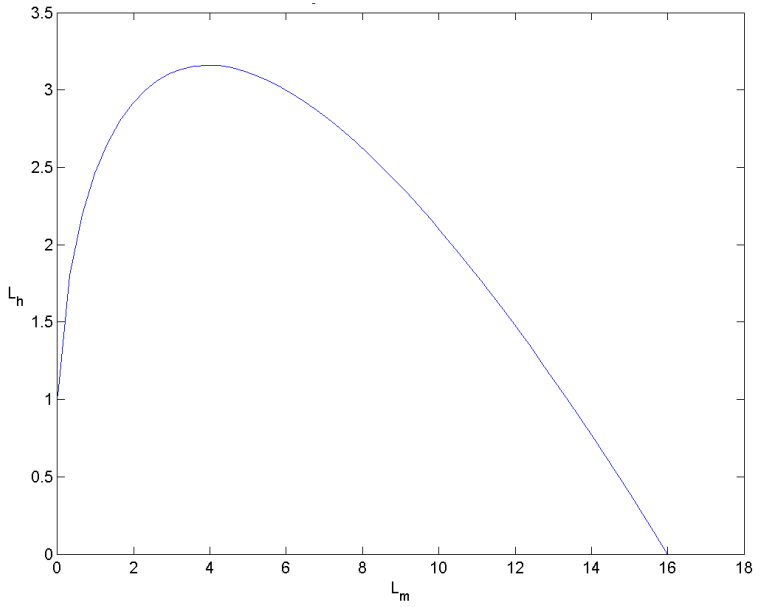
Figure: Maximum Absorption for Humans

$\rightarrow$ desirable for humans to have machines if threshold $\hat{A}^m$ surpassed

Example 1: Absorption Frontiers

Low machine productivity (left) versus high machine productivity (right):





Example 1: Absorption Frontier

Interpretation in terms of ~~property rights,~~ command over resources in a competitive economy:

- in human maximum with $N^m = 0$: interpretation trivial
- in human maximum with $N^m > 0$:
  - machines absorb their maintenance level $s^m = MPL^m$
  - humans absorb both $w^h = MPL^h$ and the entire factor rent from T,

$$s^h N^h = w^h N^h + RT$$

  note: technological progress in A^m increases land rent R

  $\rightarrow$ one interpretation: humans own everything, including machines

  $\rightarrow$ another interpretation: machines are emancipated but zero wealth

- vice versa in machine maximum
- along the frontier:
  - ownership of T is shared between humans and machines

Example 1: Machine/AIA-Only Economy

Maximum absorption for machines/AIAs:

Proposition (Machine-Only Economy)

(i) If $(1-\alpha)\,A^h/s^h < A^m/s^m$*, then maximum absorption for machines requires zero human absorption,* $N^h = 0$*. There will be a well-functioning economy where AIAs produce solely for AIA absorption.*

(ii) Otherwise, maximum absorption for machines/AIAs requires a positive $N^h > 0$*.*

Notes:

- result (i) rejects fallacy that "humans are necessary to provide demand for goods" (e.g. Ford, 2014; ...)
  $\rightarrow$ important implications for NIPA (don't subtract depreciation!)
- in result (ii), humans can be interpreted as slaves of machines/AIAs

Moving Off the Human Maximum

Question: What forces may induce humans to move off the human maximum?

- Initial endowment of AIAs
- Human impatience compard to AIAs
- Rents from transitional shortage when AIAs become more productive
- Agency rents for AIAs

Impatience and Moving Off the Human Maximum

Transition: speed depends on preferences/behavior (akin to Ramsey growth)

Consider humans only with time-separable preferences $U^i = \sum \beta^t u\left(c_t^h\right)$:

Lemma (Reaching the Human Maximum)

As $\beta \to 1$, humans reach maximum absorption

(Intuition: reaching the Golden Rule level of capital)

Consider humans and machines in a private ownership economy:

Proposition (Patience and Survival)

If $\beta^i \neq \beta^j$, then the economy converges towards the constrained maximum of the agent with higher time discount factor

Transitional Dynamics After Productivity Shock

Transitional Dynamics: consider an increase in machine productivity A^m in private ownership economy with equal discount factor and zero initial machine wealth

- in short run: $MPL^h < s^h$, $MPL^m > s^m$
- for standard preferences: humans decumulate wealth, machines accumulate wealth

Proposition (Convergence after Increase in Productivity)

In a private ownership economy, an increase in machine productivity moves the economy into the interior of the resource absorption frontier.

AIA Rents

Traditional Agency Rents:

- may allow workers (managers) to capture rent, expressed e.g. as markup $\mu^i > 1$ over their competitive wages
- are typical for agents with informational advantage
  $\rightarrow$ e.g. to obtain desirable incentive/selection effects

AIA Rents:

- may allow highly intelligent actors to extract markup $\mu^i > 0$ over competitive factor rents based on superior information processing capacity
- examples:
  - high-frequency trading
  - Amazon extracting extra consumer surplus

$\rightarrow$ AIA rents narrow the range of feasible points on the resource allocation frontier
$\rightarrow$ move into the interior

Long-Run Viability of Humans

Return to general setup: multiple goods & exog. factors, general CRS production technology

Consider effects of sustained growth in machine-specific productivity A^m:

Proposition (Redundancy of Human Labor)

$MPL^h \to 0$ except if human labor is a complement to machine labor in the production of at least one of the goods (non-substitutability)

Proposition (Long-Run Viability of Humans)

If $MPL^h \to 0$ then $N^h \to 0$ except if:

1. *either humans maintain positive net worth (positive property)*
2. *or there are no scarce factors required to produce human consumption goods that are valuable to AIAs (separability)*

.

Long-Run Policy

Long-Run Policy in the face of a Malthusian Race:

Mechanism that endangers humanity = scarcity of exogenous factors
Consolation: Malthusian race will likely look less cruel than in medieval times

- we can live in simulations [play video games] or use technology to reduce resource consumption

Policy options:

- allocation of restricted property rights to humans that cannot be sold (human reservation)
- equivalently, regular allocation of human subsistance incomes (which may be reduced by technology)
- ? slow down technological progress ?

Relating to our Present Economy

Consider general model with multiple factors and goods,
and assume sustained progress in machine technology:

- rising prices of factors most relevant for AIAs (e.g. programmers, land in Silicon Valley, etc.)
- declining labor share
- given that human aborption is more L^h-intensive than machine absorption:
  - price of machine absorption basket falls faster than of human basket
  - measured from machine perspective, fast real growth, high real interest rates, compared to human experience
- increasing corporate savings in IT sector $\rightarrow$ AIA agency rents?

Conclusions

Emergence of AIA:

- requires fundamental rethink of economic concepts, including agents, utility, etc.
- may lead to onset of a (Malthusian) race
- may already be happening