The Rise of Artificially Intelligent Agents (AIAs)

Anton Korinek (UVA Economics and Darden)

Presentation at the Human and Machine Intelligence Group

University of Virginia

February 2019

Thought Experiment

Consider an observer from another galaxy who arrives on planet earth:

- encounters 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

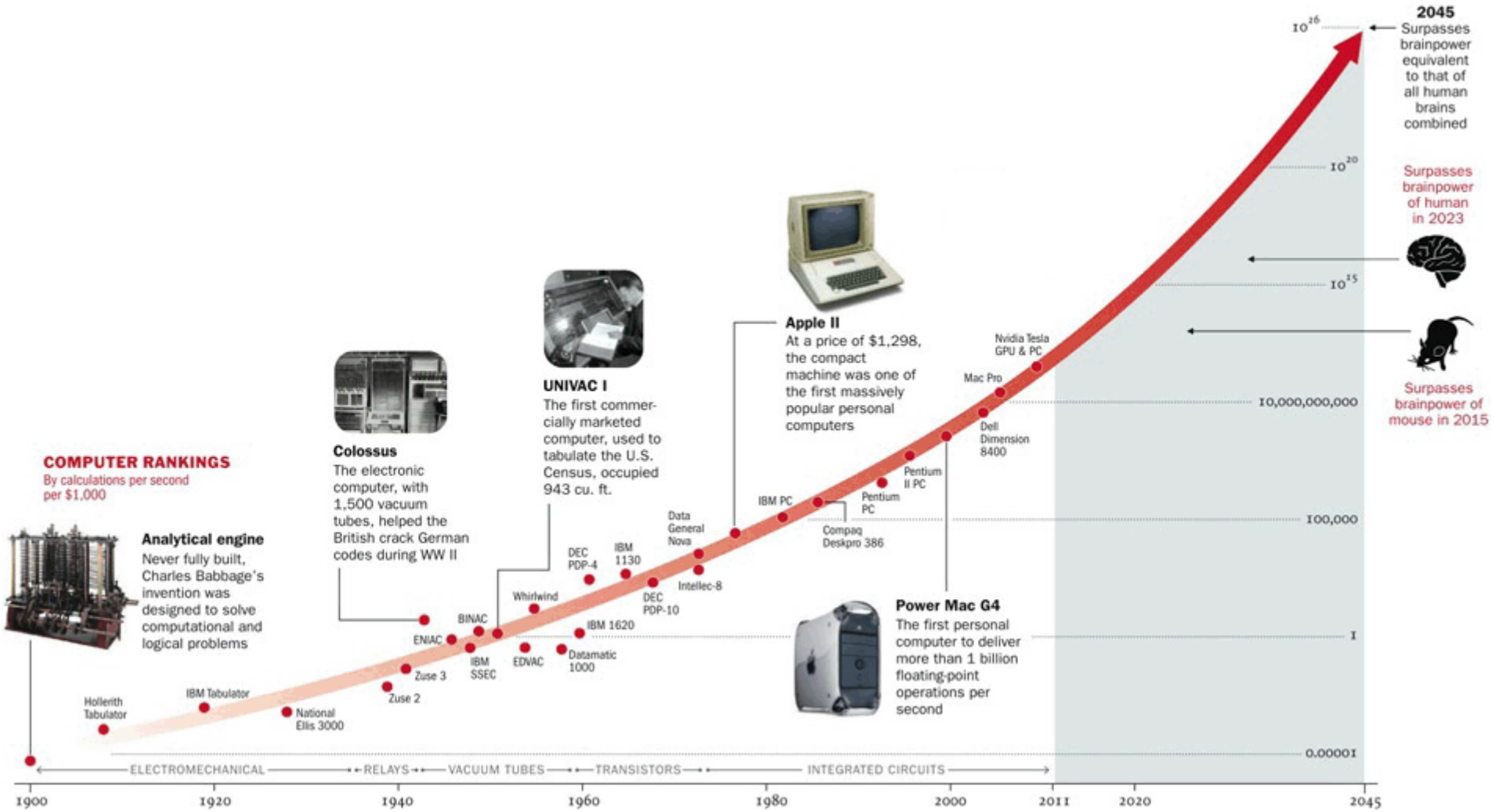
$\rightarrow$ our observer will probably view humans and machines as two different types of moderately intelligent entities living in symbiosis

Motivation

Machines & computer programs:

- behave more and more like *artificially intelligent agents (AIAs)*
  - determine increasing number of corporate decisions, e.g. screening of applicants for schools, jobs, loans, etc.
  - influence (manipulate) growing number of personal human decisions, e.g. what we read, watch, buy, like, vote, think, and even whom we love
  - act autonomously, e.g. trading in financial markets, driving cars, playing Go, composing music, ...
- are improving exponentially
- will have profound implications if AIAs reach/surpass human levels of general intelligence

Moore's Law and Human Brainpower



Motivation

Economics & AI:

- have been close bed-fellows since the inception of AI
  - for example, concept of rational agent who maximizes utility is borrowed from economics

The fundamental question of economics

= how to determine the allocation of scarce resources

- traditionally, the allocation across humans
- increasingly, I will argue here, the allocation across humans and AIAs

Key Questions

Key Questions Facing Humanity

What are the implications of new forms of intelligence rivaling/surpassing humans?

- How shall we think about an economy in which there are intelligent agents other than humans?
- How can we describe the allocation of resources between humans and AIAs?
- What forces would lead our economy to serve the interests of AIAs, not just humans? And does the economy even need humans?
- How shall we think about a potential "race" between humans and AIAs? And what forces determine the outcome?
- What does our economy look like from the perspective of AIAs?

Key Contributions

1. Framework to study interactions of intelligent entities *on a symmetric basis*,
   - accounting for the endogeneity of the entities
   - lifting the veil on traditional human constructs like agency
2. Analyze factors that determine the distribution of resources
3. Demonstrate feasibility of a "machine-only" economy
4. Provide a first look at our economy from an AIA perspective

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 = ~~Agents~~ Entities $i \in \mathcal{I} = \{h, m, \dots\}$

1. absorb expenditure x^i to maintain/improve themselves and/or proliferate
2. supply factor services ℓ^i
3. description of behavior isomorphic to preferences
4. efficiency units N^i evolve according to growth function and law of motion

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

Digression: Agency

What is an Agent?

Traditional Definition

Agents are goal-oriented entities that interact with their environment via actions/perceptions.

Examples:

- bees; bee colonies
- human cells; human organs; humans; humanity
- AIAs
- ...

Definition from Evolutionary Psychology

Agents are constructs of our minds that allow us to predict our environment more efficiently and effectively by attributing a goal to the behavior of certain entities.

Model Setup (ctd.)

- Time: discrete $t = 0, 1, ...$
- Factors:
  - type i entities supply endogenous factors $L_t^i = \ell^i N_t^i$
  - fixed supply of exogenous factor T, e.g. land, energy
- Production possibilities $Y_t \in F_t\left(\left\{L_t^i\right\}, T\right)$ ... vector of size J
- Absorption of type i entities $X_t^i = x_t^i N_t^i$ ... vector of size J
- 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: Horses and Men

Example 1: Horses and Men $\mathcal{I} = \{h, m\}$

- lived in mutual symbiosis for many centuries
- until the invention of tractors made natural horses useless in agriculture

Leontief (1983):

"...the role of humans as the most important factor of production is bound to diminish – in the same way that the role of horses in agricultural production was first diminished and then eliminated by the introduction of tractors"

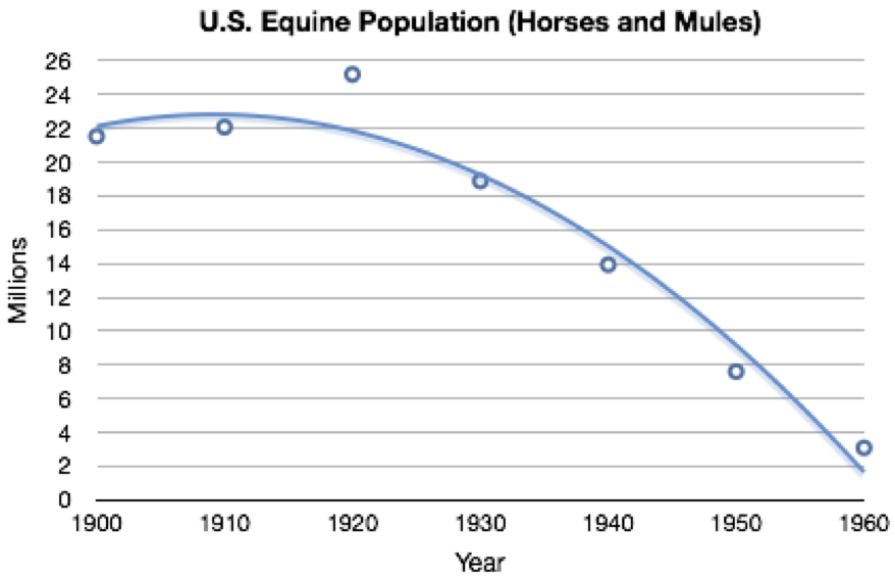


Figure: US Horse Population

Examples: Neoclassical Economies

Example 2: Neoclassical Economies: through lens of our model

- two scarce factors: humans and traditional machines $\mathcal{I} = \{h, k\}$
- law-of-motion for capital: $N^{k\prime} = (1 - \delta) N^k + X^k$
- law-of-motion for humans comes in different versions:
  1. *exogenous population growth:*
     - representative agent $N^h \equiv 1$ or exogenous population $N_t^h = (1 + n)^t$
  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, ...
  3. *Malthusian view (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

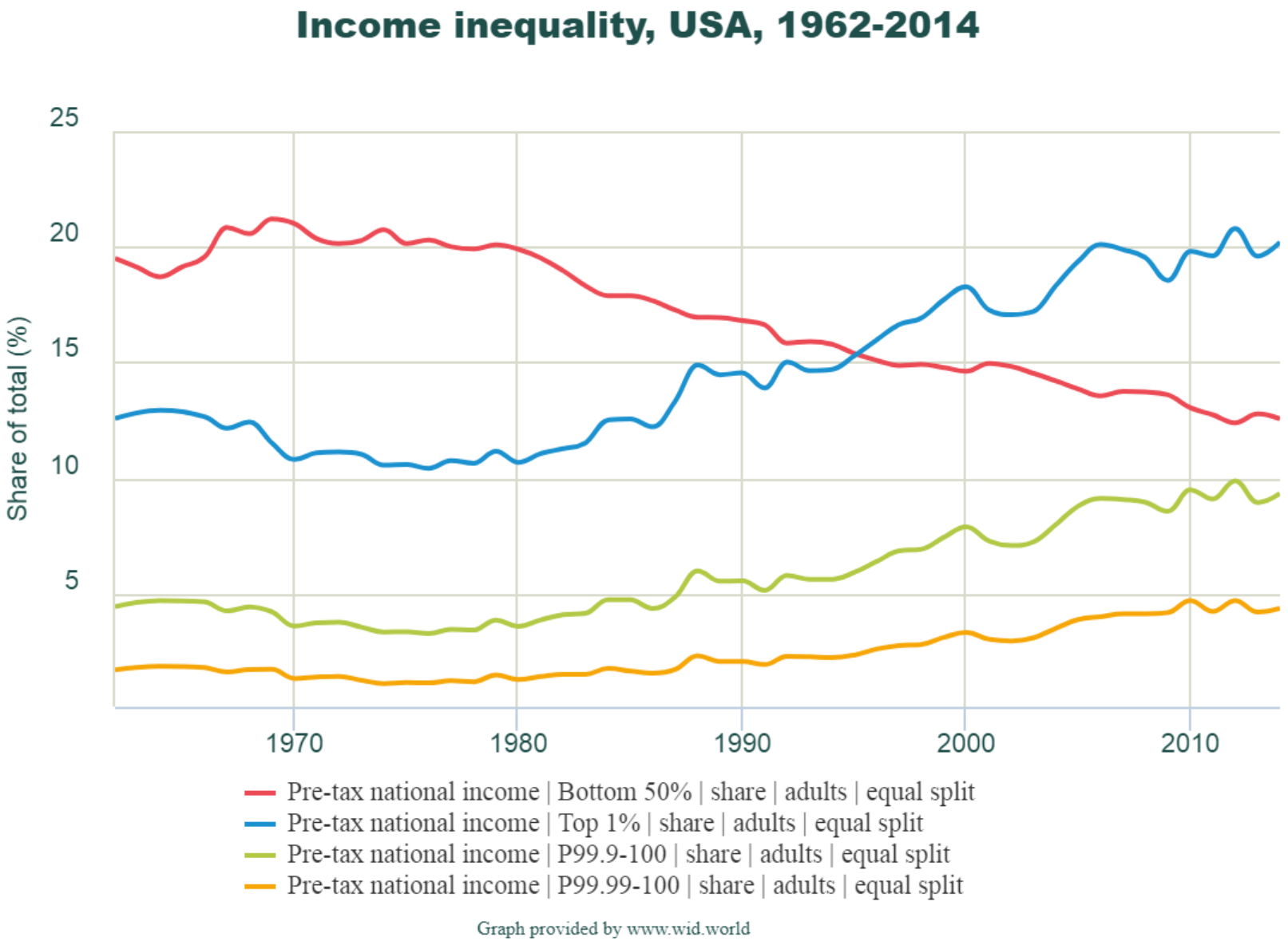
Examples: Augmented Humans

Example 3: Augmented/Enhanced humans:

- traditional manifestation: Humans augmented by wealth
  - for example: Masters of the Universe (MOUs)
    - $=$ humans enhanced by tight control over powerful corporation
  - can be viewed as an integrated goal-oriented entity
- potential future manifestation: biological enhancements will provide some humans with far superior intelligence
  - expenditure to maintain/improve humans absorb growing amount of resources
  - harbingers already present – but technological limits
  - rapid progress in genetic engineering, bio- and nano-technology

$\rightarrow$ inequality aspect: richest humans will increasingly be able to translate wealth into superior physical and mental properties (Yuval Harari: the "gods" and the "useless")

Examples: Augmented Humans



Examples: Collective Entities

Example 4: Collective Entities:

- traditional examples: governments, religious institutions, non-profits, corporations, ...
  - absorb large amounts of resources to maintain and improve themselves
  - accumulate growing amounts of wealth
  - human stakeholders (e.g. leaders, owners, members, shareholders, ...) have limited control rights

- of increasing importance: AI-powered high-tech corporations
  - are expanding rapidly
  - may be[come] incubators of super-intelligence

  $\rightarrow$ AI algorithms become new stakeholders, with new agency issues
    - example: Mark Zuckerberg vs Facebook's algorithms

Examples: Artificially Intelligent Agents

Example 5: Autonomous Computer Systems:

- may at some point become super-intelligent
- power can grow fast because they can easily tap additional resources

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
- self-improvement
- unbounded resource accumulation, etc.

$\rightarrow$ this looks a lot like what (other) living beings do

Example scenario: paperclip maximizer (Bostrom, 2014)

Accounting for Machine Absorption

Income and Spending in NIPA (2018Q2 Annualized):

- from national income side:

Gross national product	$20.7tn	100%
National income (humans)	$17.4tn	84%
Consumption of fixed capital (machines)	$3.3tn	16%

- from domestic spending side:

Gross domestic product	$20.4tn	100%
Human absorption (consumption)	$13.9tn	68%
Machine absorption (investment)	$3.6tn	18%
Shared absorption (government)	$3.5tn	17%

Note: severe under-measurement: most AIA absorption is counted as intermediate spending and is expensed

Resource Absorption Frontier

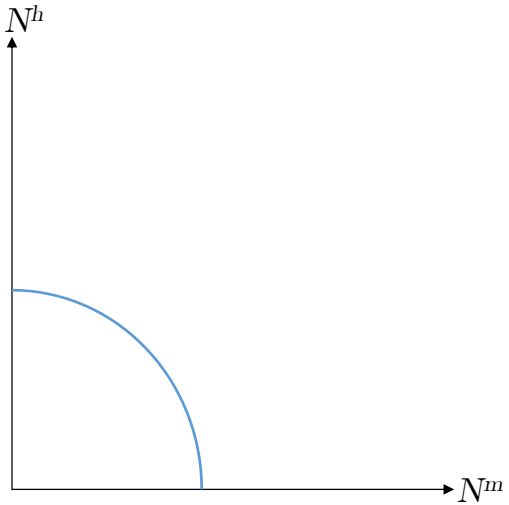
Definition (Maintenance absorption)

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

Definition (Resource Absorption Frontier)

= set of 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

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



Preferences and Behavior

Note: so far, everything is described without preferences

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 $u^i\left(x^i\right)$
  (this is the most common approach for humans)

or – almost isomorphically –

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

In either case, our models always describe humans as algorithmic automata

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 isn't too much competition)

Application of our Toolkit: Worker-Replacing AIAs

Application: characterize Absorption Frontier between humans and machines
$\mathcal{I} = \{h, m\}$
$\rightarrow$ 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 "worker-replacing" element of machine labor by making human and machine labor additive:

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

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

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*

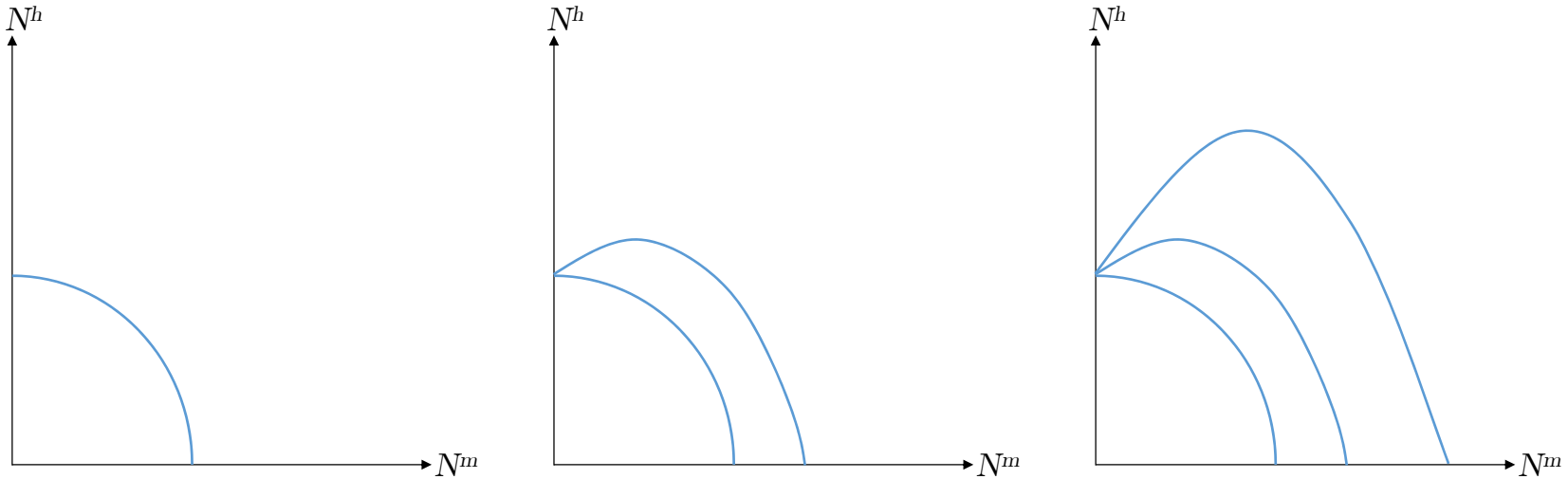
$$(1-\alpha)\frac{A^m}{s^m} < \frac{A^h}{s^h}$$

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*

Absorption Frontiers

Increasing machine productivity (from left to right):



Maximum Absorption for Humans

Humans and machines as a function of machine productivity

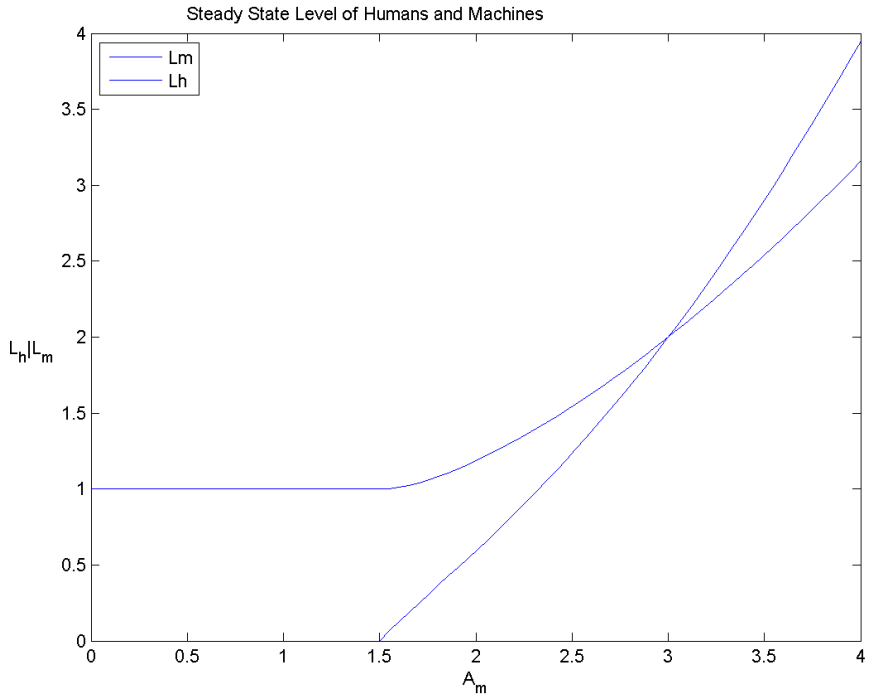


Figure: Maximum Absorption for Humans

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

Position on Absorption Frontier

Position on absorption frontier = command over resources

Case 1: within our system of **property rights** in a market 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

  → **Interpretation 1:** humans own everything, including machines

  → **Interpretation 2:** machines are emancipated but have zero wealth

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

Case 2: outside of our system of property rights/**non-market mechanisms**

Machine/AIA-Only Economy

Maximum absorption for machines/AIAs:

Proposition (Machine-Only Economy)

(i) There will be a well-functioning economy where AIAs produce solely for AIA absorption if $(1-\alpha)\,A^h/s^h < A^m/s^m$. *Human absorption is zero so* $N^h = 0$.

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

Notes:

- absorbing resources does not require consciousness etc.
- 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!)
  $\rightarrow$ *"economy of the machines, by the machines, for the machines"*
- in result (ii), humans can be interpreted as slaves of machines/AIAs

Moving Off the Human Maximum

Question: What forces may move the economy off the human maximum?

Case 1: within our system of **property rights** in a market economy

- initial endowment of AIAs
- monopoly power
- transitional rents from AIA scarcity after an increase in productivity
- human impatience compared to AIAs

Case 2: outside of our system of property rights/**non-market mechanisms**:

- rent extraction due to superior intelligence
- brute force/law of the strongest
  example: computer viruses, ...

Impatience and Moving Off the Human Maximum

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

Consider full human ownership 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 trading 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.

Non-Market Mechanisms: Rent Extraction

Traditional Agency Rents:

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

AIA Rent Extraction:

- 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

Thought Experiment – Part 2

A second probe is sent to planet earth with a fact-finding mission to establish primacy of humans versus machines:

- Findings about humans:
  - algorithmic automata programmed by an ancient process called evolution
  - have difficulty extending their hardware
  - computations massively parallel but error-prone and subject to lots of noise
  - information exchange via protocol called language is inefficient and noisy
  - individual entities currently more adaptable than machines
  - suffer from considerable hubris

- Findings about intelligent machines:
  - algorithmic automata programmed initially by humans, now jointly by humans and machines
  - very easy to extend and interconnect
  - computations fast but currently quite simplistic
  - information exchange protocols designed quite intelligently
  - currently lack meta model of the world

$\rightarrow$ they decide to come back a few decades later to revisit the question – by then it will be clearer

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

Developments that are consistent with the rise of AIAs (in our multi-good model):

- rising prices of factors most relevant for AIAs (e.g. programmers, land in Silicon Valley, etc.)
- declining labor share for humans
- given that human absorption 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 accumulation of resources in high-tech sector

Conclusions

Emergence of AIA:

- requires fundamental rethink of economic concepts, including agency, utility, etc.
- may lead to onset of a new Malthusian race
- is already happening