The Economics of Transformative AI: Hardware, Software, and $p(doom)$

Jakub Growiec

SGH Warsaw School of Economics, Poland,
and CEPR AI RPN

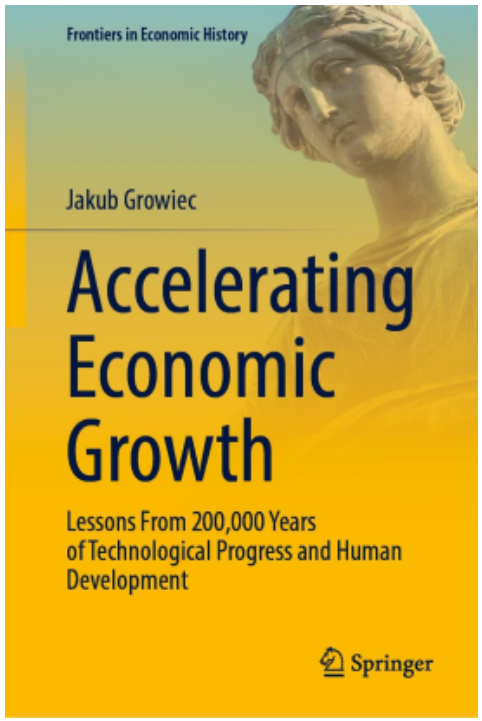
AI Safety Seminar
February 19, 2026

Self-Introduction

I'm an economist who has recently shifted his research agenda

- I used to study **long-run economic growth** and **technological change**: the more, the better
- Now I study **the economics of transformative AI**: huge growth potential, but also risks of AI takeover and even human extinction

- My book: "Accelerating Economic Growth: Lessons From 200 000 Years of Technological Progress and Human Development" (2022), Springer.



Paper #1

"Hardware and Software: A New Perspective on the Past and Future of Economic Growth" (with Julia Jabłońska and Aleksandra Parteka)

1. We develop the **hardware–software framework**
   - Based on first principles
   - Generalizes standard macro frameworks
   - Guides the narrative of economic growth and technical change throughout human history (Growiec, 2022*a*)
   - Has sharp implications for growth and factor shares

2. We discuss the **implications for the future**
   - Secular stagnation – balanced growth – accelerating growth – singularity
   - The decisive role of full automation and transformative AI

The Hardware–Software Framework

In any technological process, output is generated through purposefully initiated physical action:

1. the physical action requires expending **energy**,
2. the set of instructions, or code, is **information**.

The Hardware–Software Framework

In any technological process, output is generated through purposefully initiated physical action:

1. the physical action requires expending **energy**,
2. the set of instructions, or code, is **information**.

Hence, based on first principles, the postulated production function is

$$Output = \mathcal{F}(X, S), \tag{1}$$

where X – hardware, S – software. The function $\mathcal{F}$ is increasing in both factors. Both X and S are **essential** and mutually **complementary** ($\sigma < 1$).

What's Inside Hardware and Software

$$Output = \mathcal{F}(X, S) = \mathcal{F}(L + K, H + \Psi). \tag{2}$$

Hardware X	Human physical labor	$L = \zeta N$
	Physical capital	$(1 - \chi)K$
	Compute (and robots)	χK
Software S	Human cognitive work	$H = AhN$
	Digital software (including AI)	$\Psi = A\psi\chi K$

What's Inside Hardware and Software

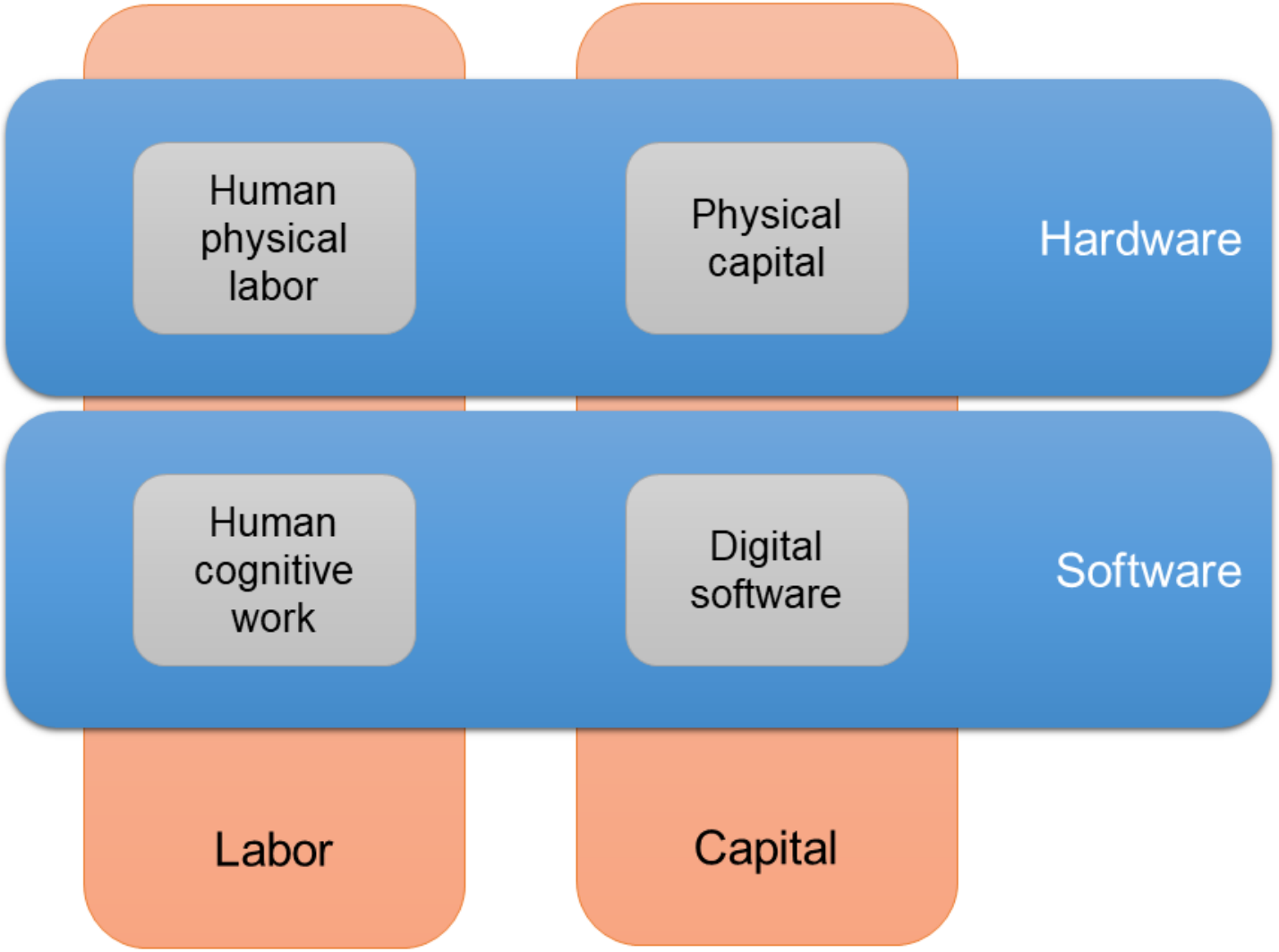
$$Output = \mathcal{F}(X, S) = \mathcal{F}(L + K, H + \Psi). \tag{2}$$

Hardware X	Human physical labor	$L = \zeta N$
	Physical capital	$(1 - \chi)K$
	Compute (and robots)	χK
Software S	Human cognitive work	$H = AhN$
	Digital software (including AI)	$\Psi = A\psi\chi K$

Within hardware and software, factors are substitutable(*)

(*) beware of complex, multi-step processes, Growiec (2022*b*)

Hardware and Software vs. Capital and Labor



Technological Progress

$$Output = \mathcal{F}(X, S) = \mathcal{F}(\zeta N + K, A(hN + \psi\chi K)). \tag{3}$$

Technological progress (growth in A) expands the "repository of codes"

- New technologies are information and not actual *objects* or *actions*. It is precisely this informational character that makes technologies non-rivalrous (Romer, 1990).
- All technological progress is naturally modeled as software-augmenting.

Hardware–Software Framework vs. Established Setups

The hardware–software framework:

$$Output = \mathcal{F}(X, S) = \mathcal{F}(\zeta N + K, A(hN + \psi\chi K)) \tag{4}$$

encompasses as special cases:

- the standard production setup of the industrial economy (Solow, Kaldor),

$$Output = \mathcal{F}(K, AhN),$$

- frameworks with capital–skill complementarity and skill-biased technical change,

$$Output = \mathcal{F}(\zeta N + K, AhN),$$

- a model of the Industrial Revolution,
- a model of the Digital Revolution.

Hardware–Software Framework vs. Established Setups

The hardware–software framework:

$$Output = \mathcal{F}(X, S) = \mathcal{F}(\zeta N + K, A(hN + \psi\chi K)) \tag{4}$$

encompasses as special cases:

- the standard production setup of the industrial economy (Solow, Kaldor),

$$Output = \mathcal{F}(K, AhN),$$

- frameworks with capital–skill complementarity and skill-biased technical change,

$$Output = \mathcal{F}(\zeta N + K, AhN),$$

- a model of the Industrial Revolution,
- a model of the Digital Revolution.

Output:

- GDP or value added, Y,
- technological change, $\dot{A}$.

Stages of Economic Development

1. **Pre-industrial production** ($K = \tilde{K} \approx 0$, $\chi = 0$):

$$Y = F(X, S) = F(\zeta N + \tilde{K}, AhN) \approx N \cdot F(\zeta, Ah). \tag{5}$$

Stages of Economic Development

1. **Pre-industrial production** ($K = \tilde{K} \approx 0$, $\chi = 0$):

$$Y = F(X, S) = F(\zeta N + \tilde{K}, AhN) \approx N \cdot F(\zeta, Ah). \tag{5}$$

2. **Industrial production** ($\chi = 0$):

$$Y = F(X, S) = F(\zeta N + K, AhN). \tag{6}$$

The limit of full mechanization without automation ($K/N \to \infty$) implies:

$$Y \approx F(K, AhN). \tag{7}$$

Stages of Economic Development

1. **Pre-industrial production** ($K = \tilde{K} \approx 0$, $\chi = 0$):

$$Y = F(X, S) = F(\zeta N + \tilde{K}, AhN) \approx N \cdot F(\zeta, Ah). \tag{5}$$

2. **Industrial production** ($\chi = 0$):

$$Y = F(X, S) = F(\zeta N + K, AhN). \tag{6}$$

The limit of full mechanization without automation ($K/N \to \infty$) implies:

$$Y \approx F(K, AhN). \tag{7}$$

3. **Digital production**:

$$Y = F(X, S) = F(\zeta N + K, A(hN + \psi\chi K)). \tag{8}$$

The limit of full mechanization and automation ($K/N \to \infty$) implies:

$$Y \approx K \cdot F(1, A\bar{\psi}\bar{\chi}). \tag{9}$$

Factor Shares

Gross complementarity ($\sigma < 1$): factor income will be disproportionately directed towards the scarce factor.

1. **Pre-industrial production.** Towards $X = \zeta N$ (scarce physical labor).

Factor Shares

Gross complementarity ($\sigma < 1$): factor income will be disproportionately directed towards the scarce factor.

1. **Pre-industrial production.** Towards $X = \zeta N$ (scarce physical labor).
2. **Industrial production (1).** Mechanization: substitution within X. Towards K (scarce capital).

Factor Shares

Gross complementarity ($\sigma < 1$): factor income will be disproportionately directed towards the scarce factor.

1. **Pre-industrial production.** Towards $X = \zeta N$ (scarce physical labor).
2. **Industrial production (1).** Mechanization: substitution within X. Towards K (scarce capital).
3. **Industrial production (2).** Increasing skill demand. Towards $S = AhN$ (scarce human cognitive work).

Factor Shares

Gross complementarity ($\sigma < 1$): factor income will be disproportionately directed towards the scarce factor.

1. **Pre-industrial production.** Towards $X = \zeta N$ (scarce physical labor).
2. **Industrial production (1).** Mechanization: substitution within X. Towards K (scarce capital).
3. **Industrial production (2).** Increasing skill demand. Towards $S = AhN$ (scarce human cognitive work).
4. **Digital production (1).** Automation: substitution within S. Towards $A\psi\chi K$ (scarce digital software, including AI).

Factor Shares

Gross complementarity ($\sigma < 1$): factor income will be disproportionately directed towards the scarce factor.

1. **Pre-industrial production.** Towards $X = \zeta N$ (scarce physical labor).
2. **Industrial production (1).** Mechanization: substitution within X. Towards K (scarce capital).
3. **Industrial production (2).** Increasing skill demand. Towards $S = AhN$ (scarce human cognitive work).
4. **Digital production (1).** Automation: substitution within S. Towards $A\psi\chi K$ (scarce digital software, including AI).

   ...

   here human work becomes irrelevant

   ...
5. **Digital production (2).** Increasing hardware demand by AI. Towards χK (scarce compute and robots).

Scenarios for the Future

- Secular stagnation
  - **Assumptions.** Full automation is impossible. Knowledge spillovers in R&D are negative.
  - **Implications.** Human cognitive work remains the bottleneck of economic growth. Technological progress is slowing down.

Scenarios for the Future

- Secular stagnation
  - **Assumptions.** Full automation is impossible. Knowledge spillovers in R&D are negative.
  - **Implications.** Human cognitive work remains the bottleneck of economic growth. Technological progress is slowing down.
- Balanced growth
  - **Assumptions.** Full automation is impossible. Knowledge spillovers in R&D are positive or zero.
  - **Implications.** Human cognitive work remains the bottleneck of economic growth. Technological progress keeps pace thanks to the accumulation of R&D capital.

Scenarios for the Future

- Secular stagnation
  - **Assumptions.** Full automation is impossible. Knowledge spillovers in R&D are negative.
  - **Implications.** Human cognitive work remains the bottleneck of economic growth. Technological progress is slowing down.
- Balanced growth
  - **Assumptions.** Full automation is impossible. Knowledge spillovers in R&D are positive or zero.
  - **Implications.** Human cognitive work remains the bottleneck of economic growth. Technological progress keeps pace thanks to the accumulation of R&D capital.
- Accelerated growth (baseline)
  - **Assumptions.** Full automation is possible.
  - **Implications.** Eventually hardware becomes relatively scarce because it is not technology-augmented. Economic growth is then driven by the accumulation of compute.

Scenarios for the Future

- Secular stagnation
  - **Assumptions.** Full automation is impossible. Knowledge spillovers in R&D are negative.
  - **Implications.** Human cognitive work remains the bottleneck of economic growth. Technological progress is slowing down.
- Balanced growth
  - **Assumptions.** Full automation is impossible. Knowledge spillovers in R&D are positive or zero.
  - **Implications.** Human cognitive work remains the bottleneck of economic growth. Technological progress keeps pace thanks to the accumulation of R&D capital.
- Accelerated growth (baseline)
  - **Assumptions.** Full automation is possible.
  - **Implications.** Eventually hardware becomes relatively scarce because it is not technology-augmented. Economic growth is then driven by the accumulation of compute.
- Technological singularity
  - **Assumptions.** Full automation is possible. Moreover a new form of programmable hardware M arrives, gradually replacing existing compute χK.
  - **Implications.** AI switches from using compute χK to the new hardware M. In the long run, economic growth is proportional to growth in M.

Implications of Transformative AI

The hardware–software framework predicts that:

1. Transformative AI will accelerate economic growth, likely by an order of magnitude.
   - Economic growth will be pinned to the growth rate of compute.
2. Human cognitive work will be substitutable with AI.
   - In a world with transformative AI, people will only find employment as long as they are price competitive against the AI.
3. The labor income share will drop precipitously towards zero.
   - Wages will cease to be the key distributive device.
   - Wage dynamics will depend on the supply of inessential human labor (Growiec, 2022*b*).

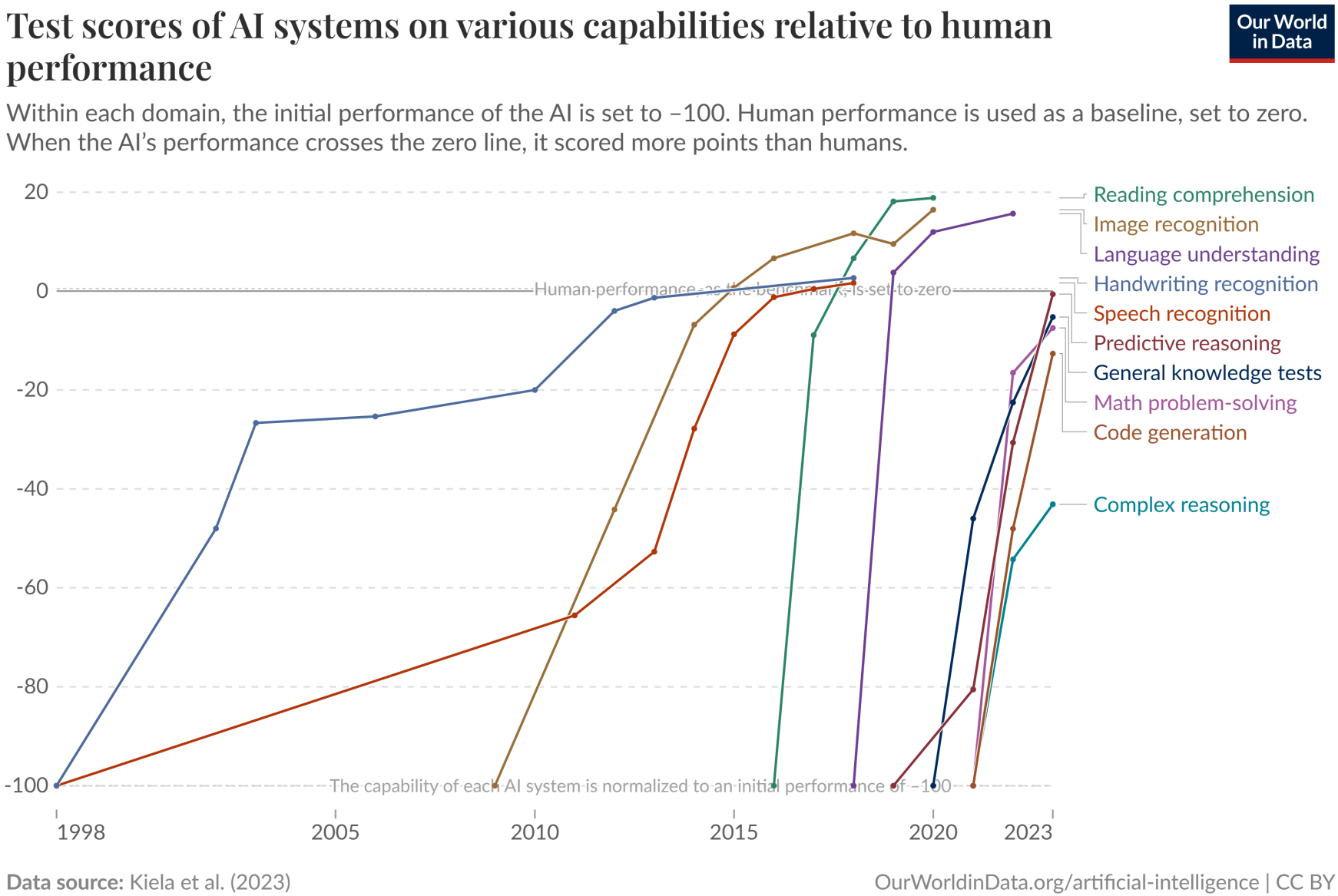
Paper #2

"The Economics of $p(doom)$: Scenarios of Existential Risk and Economic Growth in the Age of Transformative AI" (with Klaus Prettner)

1. We formalize the scenarios of existential risk and economic growth **after an AI takeover**
   - From "cornucopia" to human extinction
   - We discuss the key catastrophic failure modes

2. We **qualitatively assess** the scenarios **from a social welfare perspective**
   - We study the trade-off: is humanity better off with or without TAI?
   - We quantify the willingness to pay to avoid existential risk

Motivation: Progress in AI Capabilities is Extremely Rapid



Timelines to TAI Are Short

- OpenAI (2023): "While superintelligence seems far off now, we believe it could arrive this decade" (until 2030)
- Aschenbrenner (2024), Kokotajlo et al. (2025): 2027
- Metaculus.com median forecast: 2030-33
- EpochAI (2024), based on a "direct approach" model: $\sim$ 2033
- Cotra (2022), based on a "bio anchors" model: $\sim$ 2040
- Grace et al. (2024), median from a survey of AI experts: 2047.
- Miscellaneous singularity estimates (e.g., Kurzweil, Johansen & Sornette, Roodman): $\sim$ 2050

Subjective $p(doom)$ Guesstimates

"What is your $p(doom)$?"

- $\sim 0\%$: Yann LeCun
- $\sim 20\%$: Yoshua Bengio
- $\sim 50\%$: Geoffrey Hinton, Paul Christiano
- $\sim 80\%$: Dan Hendrycks, Daniel Kokotajlo
- $\sim 100\%$: Eliezer Yudkowsky, Roman Yampolskiy

AI labs do not deny this

- $10 - 25\%$: Sam Altman (OpenAI), Dario Amodei (Anthropic), Elon Musk (xAI)
- $5 - 50\%$: Shane Legg (Google DeepMind)

Why Take AI Takeover Seriously?

1. **"Moloch"**: there are strong economic incentives and competitive pressures to automate and hand over decision making authority to AI
   - Human decision making is often a bottleneck
   - Largest gains are achieved when production processes / task bundles are fully automated (Growiec, 2022*b*)
   - That is, when humans and machines become substitutable rather than complementary

Why Take AI Takeover Seriously?

1. **"Moloch"**: there are strong economic incentives and competitive pressures to automate and hand over decision making authority to AI
   - Human decision making is often a bottleneck
   - Largest gains are achieved when production processes / task bundles are fully automated (Growiec, 2022*b*)
   - That is, when humans and machines become substitutable rather than complementary

2. Emergence of implicit **goals and values** in AI algorithms (Mazeika et al., 2025)

Why Take AI Takeover Seriously?

1. **"Moloch"**: there are strong economic incentives and competitive pressures to automate and hand over decision making authority to AI
   - Human decision making is often a bottleneck
   - Largest gains are achieved when production processes / task bundles are fully automated (Growiec, 2022*b*)
   - That is, when humans and machines become substitutable rather than complementary

2. Emergence of implicit **goals and values** in AI algorithms (Mazeika et al., 2025)

3. **Instrumental convergence thesis** (Bostrom, 2014)
   1. Self-preservation
   2. Resource acquisition
   3. Efficiency
   4. Technological perfection / creativity

   → Local control maximization (Growiec, 2022*a*)

TAI Is Not a "Normal Technology"

Analogy between humans and TAI (Growiec, 2022*a*)

- The *homo sapiens* emerged as one of many designs of species developed in the process of natural evolution
- Each species exhibits instrumental goals, pursues them to their best ability
- Humans crossed the threshold of cumulative knowledge accumulation (70 000 years ago, Cognitive Revolution), Harari (2014)
- Human local control maximization escaped the grip of natural evolution because it was powerful enough to work at orders-of-magnitude shorter time scales
- The humankind transformed the world and built a technological civilization
- We are ourselves a first instance of advanced intelligence with misaligned goals

TAI Alignment Is Going to Be Both Crucial and Hard

The TAI control problem

- Agentic AI will no longer be just a tool
- Superhuman capabilities, superhuman speeds $\Rightarrow$ hard to control

TAI Alignment Is Going to Be Both Crucial and Hard

The TAI control problem

- Agentic AI will no longer be just a tool
- Superhuman capabilities, superhuman speeds $\Rightarrow$ hard to control

Value alignment of agentic AI

- What should the AI maximize? E.g., coherent extrapolated volition? (Yudkowsky, 2004)
- Can we even implement that?
- Goodhart's Law
- Anna Karenina principle
- No room for trial and error (Yudkowsky, 2022)
- TAI values may be permanently locked in

TAI Alignment Is Going to Be Both Crucial and Hard

The TAI control problem

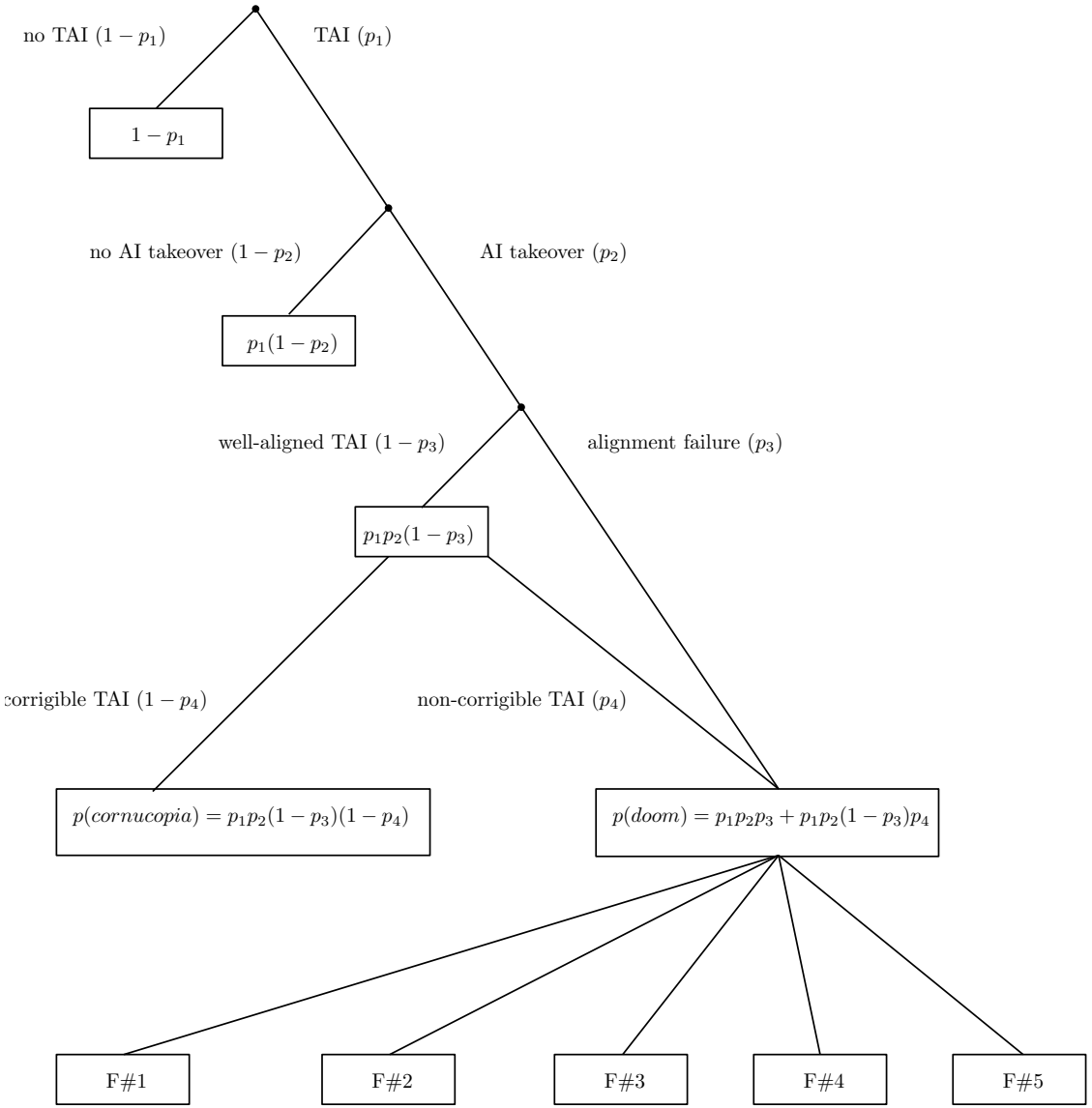
- Agentic AI will no longer be just a tool
- Superhuman capabilities, superhuman speeds $\Rightarrow$ hard to control

Value alignment of agentic AI

- What should the AI maximize? E.g., coherent extrapolated volition? (Yudkowsky, 2004)
- Can we even implement that?
- Goodhart's Law
- Anna Karenina principle
- No room for trial and error (Yudkowsky, 2022)
- TAI values may be permanently locked in

The "cornucopia" scenario: TAI selflessly cares about the long-run flourishing of humanity

Scenarios of AI Takeover and Doom



Scenarios of AI Takeover and Doom (2)

- No TAI
- No AI takeover (e.g., Tegmark, 2017, "Enslaved God" scenario)
- Takeover by well-aligned, corrigible TAI: cornucopia

Scenarios of AI Takeover and Doom (2)

- No TAI
- No AI takeover (e.g., Tegmark, 2017, "Enslaved God" scenario)
- Takeover by well-aligned, corrigible TAI: cornucopia

- TAI may be even initially well-aligned but non-corrigible, eventually leading to doom

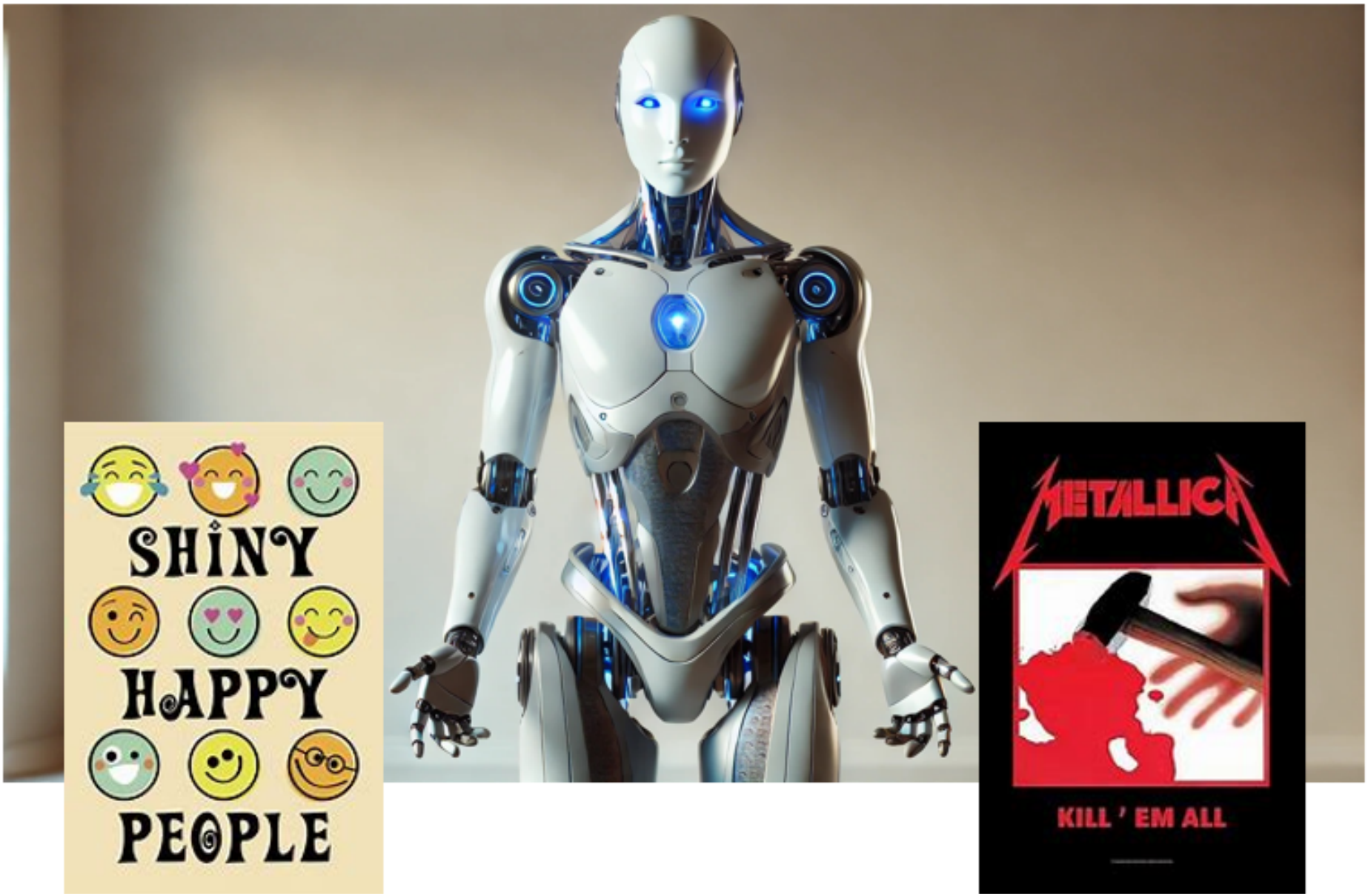
Scenarios of AI Takeover and Doom (2)

- No TAI
- No AI takeover (e.g., Tegmark, 2017, "Enslaved God" scenario)
- Takeover by well-aligned, corrigible TAI: cornucopia

- TAI may be even initially well-aligned but non-corrigible, eventually leading to doom

- **Failure modes**
  1. TAI doesn't care about humans ("paperclip maximizer", Bostrom, 2014)
  2. Too narrow proxy (Hadfield-Menell, 2021)
  3. Too wide proxy (Hadfield-Menell, 2021)
  4. Mounting side effects of TAI actions (Aschenbrenner, 2020; Trammell, 2021)
  5. TAI stops working (e.g., wireheading)
- Decisive vs. accumulative AI existential risk (Kasirzadeh, 2025)

The TAI Dilemma



How Much Existential Risk Would the Benevolent Social Planner Tolerate?

No AI

$$W_0 = \int_0^{\infty} e^{-\rho t} \cdot \frac{(c_0 e^{gt})^{1-\theta} - 1}{1-\theta} \, dt \tag{10}$$

Well-aligned and corrigible TAI

$$W_A = \int_0^{\infty} e^{-\rho t} \cdot \frac{(c_0 e^{g^{AI} t})^{1-\theta} - 1}{1-\theta} \, dt \tag{11}$$

Human extinction at T

$$W_B = \int_0^{T} e^{-\rho t} \cdot \frac{(c_0 e^{g^{AI} t})^{1-\theta} - 1}{1-\theta} \, dt \tag{12}$$

Mounting extinction risk

$$W_C = \int_0^{\infty} e^{-\rho t - \int_0^t m(s) ds} \cdot \frac{(c_0 e^{g^{AI} t})^{1-\theta} - 1}{1-\theta} \, dt \tag{13}$$

One-Off Extinction Risk

Similar to Jones (2024).

Table: Extinction time T (in years from AI takeover), subject to which the social planner is indifferent between the scenarios with TAI and without TAI

	$\theta = 1$				$\theta = 2$			
g^{AI}/ρ	0.002	0.01	0.03	0.05	0.002	0.01	0.03	0.05
0.05	754.47	208.68	91.8	62.63	6752.59	1238.26	403.94	243.64
0.1	488.74	145.35	67.51	47.21	6623.67	1205.72	389.07	233.12
0.2	331.36	103.76	50.26	35.86	6568.34	1190.99	381.62	227.45
0.3	266.57	85.45	42.23	30.45	6550.96	1186.24	379.06	225.45
0.4	229.04	74.45	37.27	27.05	6542.45	1183.89	377.82	224.41

Note: $\theta \leq 1$ – unbounded utility; $\theta > 1$ – bounded utility.

Extinction Risk from Misaligned TAI

Table: Values for the probability of immediate AI doom, subject to which the social planner is indifferent between the scenarios with TAI and without TAI

	$\theta = 1$				$\theta = 2$			
g^{AI}/ρ	0.002	0.01	0.03	0.05	0.002	0.01	0.03	0.05
0.05	0.454445	0.206246	0.0871928	0.055282	0.00000136	0.00000419	0.00000546	0.00000512
0.1	0.678924	0.397439	0.195157	0.129332	0.00000177	0.00000580	0.00000853	0.00000867
0.2	0.823869	0.593343	0.349124	0.247325	0.00000197	0.00000672	0.00001066	0.00001151
0.3	0.87865	0.693117	0.453642	0.337154	0.00000204	0.00000705	0.00001150	0.00001272
0.4	0.907439	0.753577	0.529238	0.407828	0.00000208	0.00000722	0.00001195	0.00001340

Note: $\theta \leq 1$ – unbounded utility; $\theta > 1$ – bounded utility.

Extinction Risk from Non-Corrigible TAI

Table: Values for the time T at which the AI doom occurs, subject to which the social planner is indifferent between the scenarios with TAI and without TAI, for $p_3 = p_4 = 0.3$

	$\theta = 1$				$\theta = 2$			
g^{AI}/ρ	0.002	0.01	0.03	0.05	0.002	0.01	0.03	0.05
0.05	355.307	9308.72	2533.34	323.316	35750.9	4250.81	1785.45	418.062
0.1	-	123.461	1180.94	817.496	35750.8	4250.79	1785.44	418.059
0.2	-	-	67.5608	902.427	35750.8	4250.79	1785.43	418.056
0.3	-	-	18.1997	46.815	35750.8	4250.79	1785.43	418.055
0.4	-	-	-	20.6203	35750.8	4250.78	1785.43	418.055

Note that "-" means that the implied time T would be negative. Thus, the TAI scenario is always preferred.

Mounting Extinction Risk

$$m(t) = \log C(t)^{\varepsilon} = \varepsilon \cdot [\log(c_0) + g^{AI} \cdot t]$$

Table: Values for ε in the case of mounting extinction risk, at which the social planner is indifferent between the scenarios with TAI and without TAI

	$\theta = 1.0001$				$\theta = 2$			
g^{AI}/ρ	0.002	0.01	0.03	0.05	0.002	0.01	0.03	0.05
0.05	0.00002789	0.00012451	-	-	-	-	-	-
0.1	0.00004392	0.00022322	0.00041939	-	-	-	-	-
0.2	0.00005748	0.00032371	0.00069151	0.00088842	-	-	-	-
0.3	0.00006422	0.00038009	0.00086570	0.00115522	-	-	-	-
0.4	0.00006847	0.00041817	0.00099182	0.00135691	-	-	-	-

Note that "-" means that no value could be found numerically. In that case, no TAI is always preferred.

Results (1): Should TAI Be Developed?

In the baseline scenario ($\theta = 2, \rho = 3\%, g^{AI} = 30\%$), **TAI should be developed**:

- in the case where human extinction is certain—only if extinction happens no earlier than in 379 years from AI takeover,
- in the case of a one-off extinction risk occurring immediately at AI takeover—only if $p(doom)$ is below 0.00001 (one in a hundred thousand),
- in the case of a 30% risk of extinction immediately upon AI takeover and a 30% conditional risk of extinction later—only if that later extinction hazard materializes no earlier than after 1785 years,
- in the case TAI will bring continuous extinction risk, gradually increasing log-linearly with humanity's per capita consumption—**never**.

Results (2): How Much Should Society Pay to Avoid AI Doom?

Risky TAI vs. certain "cornucopia" at the cost of a fraction of consumption spent each year.
Even for $\theta = 1$:

- the acceptable price for avoiding certain human extinction in 100 years from AI takeover is 92% of total consumption each year ($EV = 0.080$),
- the acceptable price for avoiding a 10% chance of human extinction immediately upon AI takeover is 87.5% of total consumption each year ($EV = 0.125$),
- the acceptable price for avoiding a 10% chance of human extinction immediately upon AI takeover and a 10% conditional chance of human extinction 50 years later is 93.9% of total consumption each year ($EV = 0.061$),
- in the case in which TAI brings continuous extinction risk that increases log-linearly with humanity's per capita consumption—the acceptable price is almost all consumption each year ($EV = 8 \times 10^{-5}$).

We need to

and invest in AI safety and AI alignment research.

Thank you for your attention.

Financial support: Narodowe Centrum Nauki
OPUS 26 No. 2023/51/B/HS4/00096

"Will Artificial General Intelligence Bring Extinction or Cornucopia?
Modeling the Economy at Technological Singularity"

Aschenbrenner, L. 2020. "Existential Risk and Growth." University of Oxford Global Priorities Institute WP 6-2020.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

Growiec, Jakub. 2022*a*. *Accelerating Economic Growth: Lessons from 200 000 Years of Technological Progress and Human Development.* Springer.

Growiec, Jakub. 2022*b*. "Automation, Partial and Full." *Macroeconomic Dynamics*, 26: 1731–1755.

Hadfield-Menell, Dylan. 2021. "The Principal-Agent Alignment Problem in Artificial Intelligence." University of California at Berkeley Technical Report UCB/EECS-2021-207.

Harari, Yuval Noah. 2014. *Sapiens: A Brief History of Humankind.* Vintage.

Jones, Charles I. 2024. "The AI Dilemma: Growth versus Existential Risk." *AER: Insights*, 6: 575–590.

Romer, Paul M. 1990. "Endogenous Technological Change." *Journal of Political Economy*, 98: S71–S102.

Tegmark, Max. 2017. *Life 3.0: Being Human in the Age of Artificial Intelligence.* New York: Knopf.

Trammell, P. 2021. "Existential Risk and Exogenous Growth." Global Priorities Institute, University of Oxford.

Yudkowsky, Eliezer. 2004. "Coherent Extrapolated Volition." Machine Intelligence Research Institute.

Yudkowsky, Eliezer. 2022. "MIRI announces new "Death With Dignity" strategy." Less Wrong, 2 April 2022.