## **Transformative AI, existential risk, and real interest rates**

Trevor Chow <sup>1</sup> Basil Halperin <sup>2</sup> J. Zachary Mazlish <sup>3</sup>

<sup>1</sup>Stanford, <sup>2</sup>Stanford, <sup>3</sup>Oxford & GPI

January 2025



## How can we forecast the development of transformative AI?

#### 1. Surveys

► Grace et al (2024): 2047



## How can we forecast the development of transformative AI?

#### 1. Surveys

► Grace et al (2024): 2047

→ more

#### 2. Models / trend extrapolation:

- ► Kurzweil (2005): 2045
- ► Cotra (2022): 2040
- ► Davidson (2023): 2043
- ► Epoch (2022): 2046



## How can we forecast the development of transformative AI?

#### 1. Surveys

► Grace et al (2024): 2047



#### 2. Models / trend extrapolation:

- ► Kurzweil (2005): 2045
- ► Cotra (2022): 2040
- ► Davidson (2023): 2043
- ► Epoch (2022): 2046



#### 3. Asset prices

- Prices aggregate dispersed wisdom (Hayek 1945)...
- ...financial market prices especially so (Fama, etc)



#### **Real interest rates**

**Central point:** short timelines for transformative AI would *increase real interest rates* 

#### **Definition (Transformative AI)**

"Artificial intelligence technology that has at least as profound an impact on the human trajectory as did the industrial revolution or agricultural revolution"

#### **Definition (Transformative AI)**

"Artificial intelligence technology that has at least as profound an impact on the human trajectory as did the industrial revolution or agricultural revolution"

Want to recognize double-edged sword nature (Jones 2023):

#### **Definition (Transformative AI)**

"Artificial intelligence technology that has at least as profound an impact on the human trajectory as did the industrial revolution or agricultural revolution"

Want to recognize double-edged sword nature (Jones 2023):

1. May rapidly accelerate growth...

#### **Definition (Transformative AI)**

"Artificial intelligence technology that has at least as profound an impact on the human trajectory as did the industrial revolution or agricultural revolution"

Want to recognize double-edged sword nature (Jones 2023):

- 1. May rapidly accelerate growth...
- 2. ...may pose an "existential risk"

#### **Definition (Transformative AI)**

"Artificial intelligence technology that has at least as profound an impact on the human trajectory as did the industrial revolution or agricultural revolution"

Want to recognize double-edged sword nature (Jones 2023):

- 1. May rapidly accelerate growth...
- 2. ...may pose an "existential risk"

#### **Definition (Aligned transformative AI)**

Technology that causes growth in global GDP in excess of 30% per year.

► Follows Davidson (2023): a 10× increase in growth

#### **Definition (Transformative AI)**

"Artificial intelligence technology that has at least as profound an impact on the human trajectory as did the industrial revolution or agricultural revolution"

Want to recognize double-edged sword nature (Jones 2023):

- 1. May rapidly accelerate growth...
- 2. ...may pose an "existential risk"

#### **Definition (Aligned transformative AI)**

Technology that causes growth in global GDP in excess of 30% per year.

► Follows Davidson (2023): a 10× increase in growth

#### **Definition (Unaligned AI)**

Technology that causes human extinction.

### **Theory**

**Empirics** 

Discussion

**Appendix** 



## Real interest rates are determined by the supply and demand for savings

#### Ramsey rule:

$$r = \rho + \frac{1}{\sigma}g$$

- ► r: real interest rate
- $\triangleright$   $\rho$ : time discounting
- ► *g*: growth rate
- $ightharpoonup \sigma > 0$ : elasticity of intertemporal substitution

## Real interest rates are determined by the supply and demand for savings

#### Ramsey rule:

$$r = \frac{\rho}{\sigma} + \frac{1}{\sigma}g$$

- ► r: real interest rate
- $\triangleright \rho$ : time discounting
- ► *g*: growth rate
- $\sigma > 0$ : elasticity of intertemporal substitution

#### 1. Time discounting and mortality risk

"Intrinsic preference for the present"

+

Probability of death

· Intuition: no reason to save if dead

## Real interest rates are determined by the supply and demand for savings

#### Ramsey rule:

$$r = \rho + \frac{1}{\sigma}g$$

- ► r: real interest rate
- $\triangleright$   $\rho$ : time discounting
- ► *g*: growth rate
- $\sigma > 0$ : elasticity of intertemporal substitution

#### 1. Time discounting and mortality risk

"Intrinsic preference for the present"

+

Probability of death

· Intuition: no reason to save if dead

#### 2. Economic growth

 Intuition: consumption smoothing ("no reason to save if going to be rich")

### Real interest rates and transformative AI

#### Ramsey rule:

Aligned transformative AI: 
$$g = 30\%$$

$$r = \rho + \frac{1}{\sigma}g$$

- ► r: real interest rate
- $\triangleright$   $\rho$ : time discounting
- ► g: growth rate
- $ightharpoonup \sigma > 0$ : elasticity of intertemporal substitution

#### **Real interest rates and transformative AI**

#### Ramsey rule:

$$r = \rho + \frac{1}{\sigma}g$$

- ► r: real interest rate
- $\triangleright$   $\rho$ : time discounting
- ► g: growth rate
- $ightharpoonup \sigma > 0$ : elasticity of intertemporal substitution

#### Aligned transformative AI: g = 30%

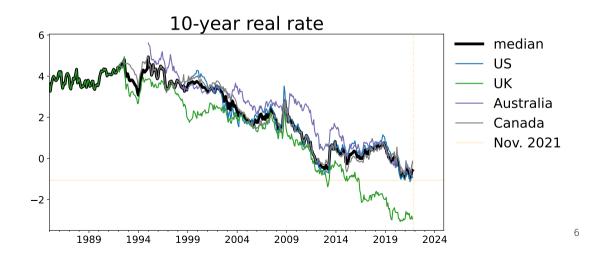
Example calibration:  $\rho=$  1%,  $\sigma=$  1, g= 1%

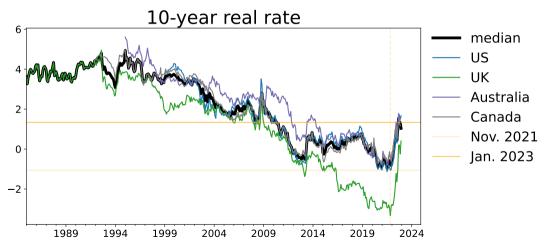
Then:

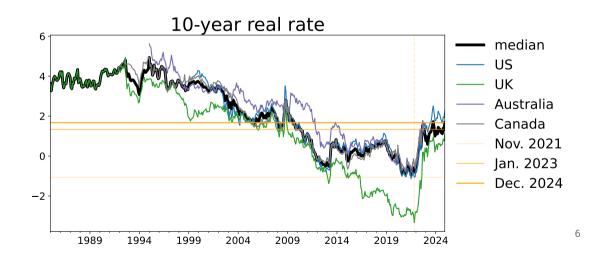
$$r = 2\%$$

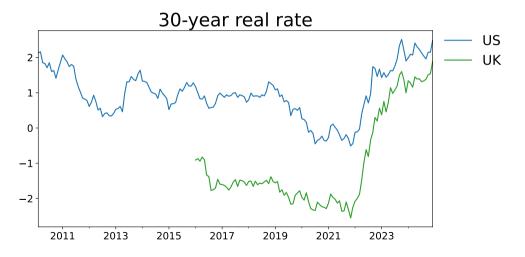
VS.

$$r = 31\%$$
!









"I'm usually on the more frugal, conservative finance side. My whole life's been budgeting, and we're like, 'Nope, we're going to take that extra, very expensive vacation this year," said Sharon Korinek.

Korinek used to do chief financial officer-type work for private companies but hasn't returned to an office job because she doesn't see the point of it. Her husband, Anton, agrees. He's an economist at the University of Virginia who researches artificial intelligence and has aired his viewed at <u>Marketplace Tech.</u>

The Korineks can fund that extra vacation partly because they're not saving for one verybig-ticket item in the future.

"I've had so many arguments about this. I'm like, 'Our kids aren't going to college," Sharon said. "Most people look at us like we're crazy."

Sharon and Anton have two kids, ages 8 and 6. The thought of opening a 529 college savings plan feels absurd to them, considering that they envision a future in which AI will be smarter than most humans.



## **Zachary Anglin** @zachanglin · Sep 26, 2022 Have you?

Q

**Ċ**Ţ







**Kelsey Piper** 

@KelseyTuoc

taken out large (>1M) long-term loans? yes

9:12 PM · Sep 26, 2022



•••

In 2017 I was convinced AI timelines were <5 years so I cashed in my 401k and blew away the money and let me tell you this particular form of intellectual consistency is Not Recommended



#### For employees

- ✓ Health, dental, and vision insurance
- ✓ Mental healthcare support and services
- ✓ Commuter benefits
- √ 401(k) with generous matching
- ✓ Domestic conference budget for each employee

### **Theory**

## **Empirics**

Discussion

**Appendix** 

$$r_t = i_t - \mathbb{E}_t \pi_{t+1}$$

$$r_t = i_t - \mathbb{E}_t \pi_{t+1}$$

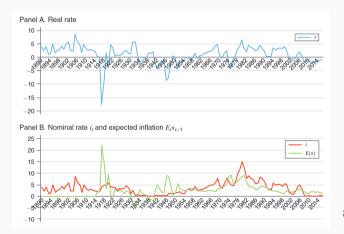
Most bonds have historically been **nominal** 

► Most research: start with  $i_t$ , estimate  $\mathbb{E}_t \pi_{t+1}$ , calcuate  $r_t$ 

# Most bonds have historically been **nominal**

- ► Most research: start with  $i_t$ , estimate  $\mathbb{E}_t \pi_{t+1}$ , calcuate  $r_t$
- Most research: estimate  $\mathbb{E}_t \pi_{t+1}$  using AR(1)
  - Not great, especially at turning points

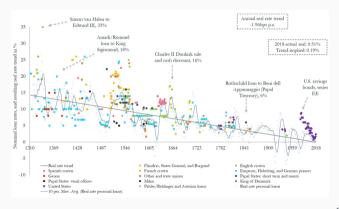
$$r_t = i_t - \mathbb{E}_t \pi_{t+1}$$



# Most bonds have historically been **nominal**

- ► Most research: start with  $i_t$ , estimate  $\mathbb{E}_t \pi_{t+1}$ , calcuate  $r_t$
- Most research: estimate  $\mathbb{E}_t \pi_{t+1}$  using AR(1)
  - Not great, especially at turning points
- ► (Other problem: credit risk)

$$r_t = i_t - \mathbb{E}_t \pi_{t+1}$$



# Most bonds have historically been **nominal**

- ► Most research: start with  $i_t$ , estimate  $\mathbb{E}_t \pi_{t+1}$ , calcuate  $r_t$
- Most research: estimate  $\mathbb{E}_t \pi_{t+1}$  using AR(1)
  - Not great, especially at turning points
- ► (Other problem: credit risk)

#### $r_t = i_t - \mathbb{E}_t \pi_{t+1}$

## Real rates should have very different drivers in the short run versus in the long run

- ► **Long run:** prices/wages are flexible. High growth ⇒ high *r*
- ► **Short run:** prices/wages are sticky. 'Excessively' high  $r \Longrightarrow$  low growth!

# Most bonds have historically been **nominal**

- ► Most research: start with  $i_t$ , estimate  $\mathbb{E}_t \pi_{t+1}$ , calcuate  $r_t$
- Most research: estimate  $\mathbb{E}_t \pi_{t+1}$  using AR(1)
  - Not great, especially at turning points
- ► (Other problem: credit risk)

#### $r_t = i_t - \mathbb{E}_t \pi_{t+1}$

## Real rates should have very different drivers in the short run versus in the long run

- ► **Long run:** prices/wages are flexible. High growth ⇒ high *r*
- ► **Short run:** prices/wages are sticky. 'Excessively' high  $r \Longrightarrow$  low growth!

$$r_t = i_t - \mathbb{E}_t \pi_{t+1}$$

Looking exclusively at long-term (10+) year horizon:

$$r_t = i_t - \mathbb{E}_t \pi_{t+1}$$

Looking exclusively at long-term (10+) year horizon:

- 1. Use *real* bonds instead of nominal bonds
  - ▶ ("inflation-linked bonds"; "TIPS")

$$r_t = i_t - \mathbb{E}_t \pi_{t+1}$$

#### Looking exclusively at long-term (10+) year horizon:

# 1. Use *real* bonds instead of nominal bonds

- ▶ ("inflation-linked bonds"; "TIPS")
- ► Why not used previously? Only ~20y of data in US; ~30y in UK/AU/CA

$$r_t = i_t - \mathbb{E}_t \pi_{t+1}$$

### Looking exclusively at long-term (10+) year horizon:

# 1. Use *real* bonds instead of nominal bonds

- ► ("inflation-linked bonds"; "TIPS")
- ► Why not used previously? Only ~20y of data in US; ~30y in UK/AU/CA

# **2. Directly measure expected inflation** using rich cross-country survey data on long-term expected inflation

- ► Source: Consensus Economics (\$\$\$)
- ► 89 countries over 30 years
- ► 10y horizon

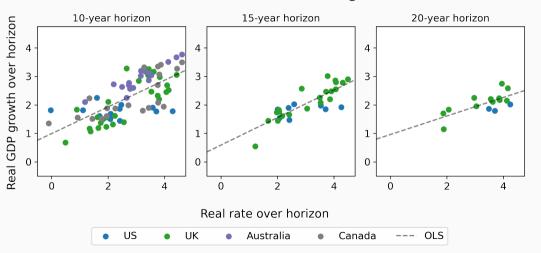
# Inflation-linked bonds and realized growth



# Inflation-linked bonds and realized growth

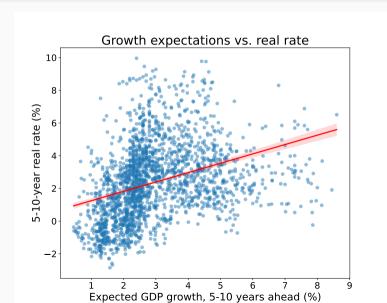


#### Real rate vs. future real GDP growth

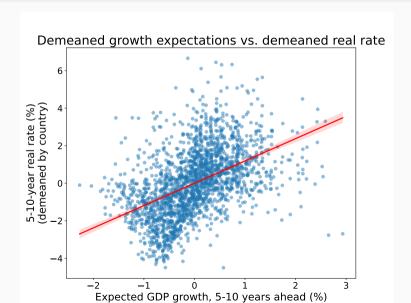




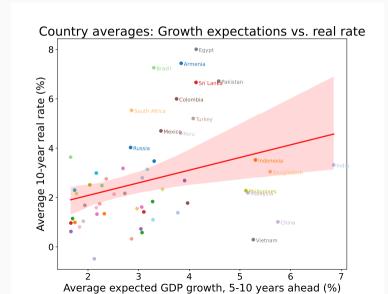




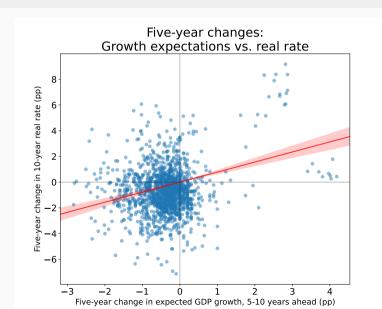












**Theory** 

**Empirics** 

Discussion

**Appendix** 

1. Non-transformative AI is consistent even with low rates

- 1. Non-transformative AI is consistent even with low rates
  - ► Precautionary savings against automation

- 1. Non-transformative AI is consistent even with low rates
  - ► Precautionary savings against automation
  - ► TAI won't affect growth

- 1. Non-transformative AI is consistent even with low rates
  - ► Precautionary savings against automation
  - ► TAI won't affect growth
- 2. Marginal utility may remain high post-singularity
  - ► New products [Trammell (2024)]
  - ► Habit formation

- 1. Non-transformative AI is consistent even with low rates
  - Precautionary savings against automation
  - ► TAI won't affect growth
- 2. Marginal utility may remain high post-singularity
  - ► New products [Trammell (2024)]
  - ► Habit formation

#### Bad arguments:

- (i) "Want to invest more to have a shot at controlling the lightcone".Maresca 2025: you want to save more, but this still pushes up r
- (ii) "High expected returns": movement along supply curve vs. shift in supply curve
  - Also distinguish between high risk-free rate (discussion here) versus high risk premium (not discussed here)

Challenge: measuring existential risk over time

**Challenge:** near-existential disasters could lower real rate!

lacktriangleright E.g. nuclear war  $\Longrightarrow$  precautionary saving  $\Longrightarrow$  lower real rate

Challenge: measuring existential risk over time

**Challenge:** near-existential disasters could lower real rate!

lacktriangleright E.g. nuclear war  $\Longrightarrow$  precautionary saving  $\Longrightarrow$  lower real rate

# Individual mortality risk and individual savings behavior

- ► ⇒ Test mechanism
- 1. New AIDS therapy rollout  $\Longrightarrow$  more savings, more education
- 2. Information treatment  $\Longrightarrow$  more investment in agriculture, livestock
- 3. Huntington's disease testing  $\Longrightarrow$  30pp less likely to go to college

**Challenge:** measuring existential risk over time

**Challenge:** near-existential disasters could lower real rate!

lacktriangleright E.g. nuclear war  $\Longrightarrow$  precautionary saving  $\Longrightarrow$  lower real rate

# Individual mortality risk and individual savings behavior

- ► ⇒ Test mechanism
- 1. New AIDS therapy rollout  $\Longrightarrow$  more savings, more education
- 2. Information treatment  $\Longrightarrow$  more investment in agriculture, livestock
- 3. Huntington's disease testing  $\Longrightarrow$  30pp less likely to go to college
- 4. Cross-sectionally: pessimistic survival beliefs are correlated with a lower savings rate

**Challenge:** measuring existential risk over time

**Challenge:** near-existential disasters could lower real rate!

lacktriangle E.g. nuclear war  $\Longrightarrow$  precautionary saving  $\Longrightarrow$  lower real rate

# Individual mortality risk and individual savings behavior

- ► ⇒ Test mechanism
- 1. New AIDS therapy rollout  $\Longrightarrow$  more savings, more education
- 2. Information treatment  $\implies$  more investment in agriculture, livestock
- 3. Huntington's disease testing  $\Longrightarrow$  30pp less likely to go to college
- 4. Cross-sectionally: pessimistic survival beliefs are correlated with a lower savings rate
- 5. Cold War evidence (sorta)

#### Stocks are harder to use to forecast timelines:

1. Only reflect aligned case (future profits)

- 1. Only reflect aligned case (future profits)
- 2. Stocks only reflect public companies (private companies, nonexistent companies)

- 1. Only reflect aligned case (future profits)
- 2. Stocks only reflect public companies (private companies, nonexistent companies)
- 3. Capped profits / nationalization

- 1. Only reflect aligned case (future profits)
- 2. Stocks only reflect public companies (private companies, nonexistent companies)
- 3. Capped profits / nationalization
- 4. TAI could lower stock prices: depends if  $\sigma \stackrel{>}{<} 1$

$$P = \frac{D}{r - g}$$

- 1. Only reflect aligned case (future profits)
- 2. Stocks only reflect public companies (private companies, nonexistent companies)
- 3. Capped profits / nationalization
- 4. TAI could lower stock prices: depends if  $\sigma \stackrel{>}{<} 1$

$$P = \frac{D}{r - g}$$

$$= \frac{D}{(\rho + \sigma \cdot g) - g}$$

$$= \frac{D}{\rho + (\sigma - 1) \cdot g}$$



# **Conclusion: Two possibilities**

#### 1. Markets are are efficient information aggregators

#### 2. Markets are wrong

- ► Trade opportunity? ("Get rich or die trying")
- ► Opportunity to borrow cheaply? ("Impatient philanthropy")

# **Conclusion: Two possibilities**

#### 1. Markets are are efficient information aggregators

#### 2. Markets are wrong

- ► Trade opportunity? ("Get rich or die trying")
- ► Opportunity to borrow cheaply? ("Impatient philanthropy")

#### Contribution:

- 1. Al safety: outside view evidence on Al timelines
- Mainline economics: fundamental question about determinants of real interest rates

### **Theory**

**Empirics** 

Discussion

**Appendix** 





► Grace et al. (2018) survey of Al researchers: "when unaided machines can accomplish every task better and more cheaply than human workers"

Median: 2061

► Grace et al. (2022):

Median: 2058



Grace et al. (2018) survey of AI researchers: "when unaided machines can accomplish every task better and more cheaply than human workers"

· Median: 2061

► Grace et al. (2022):

• Median: **2058** 

► Grace et al. (2024):

· Median: 2047



► Grace et al. (2018) survey of Al researchers: "when unaided machines can accomplish every task better and more cheaply than human workers"

Median: 2061

► Grace et al. (2022):

• Median: **2058** 

► Grace et al. (2024):

· Median: 2047

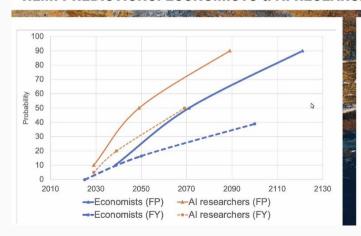
Metaculus forecasting platform

· Median: 2031



Korinek et al. (2022) survey of economists: 2070-2130+

# HLMI PREDICTIONS: ECONOMISTS & AI RESEARCHERS



# Economists ascribe a median 10% probability to HLMI never being developed.

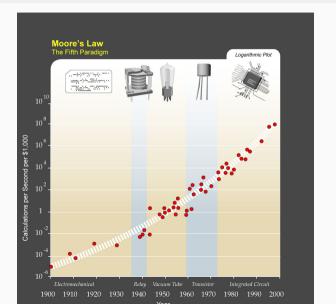
FP = fixed probability (shown probability, asked for years)

FY = fixed year (shown year, asked for likelihood)

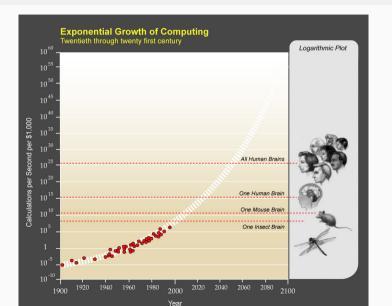
# 2. Models / trend extrapolation



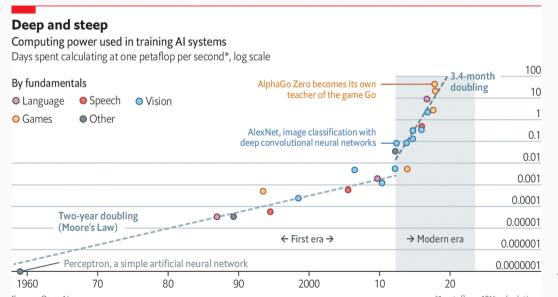














#### **Compute-centric forecasting models:**

1. Estimate # of computations brain per second



#### **Compute-centric forecasting models:**

- 1. Estimate # of computations brain per second
- Project forward trends in power of computers (e.g. Moore's Law)



#### **Compute-centric forecasting models:**

- 1. Estimate # of computations brain per second
- Project forward trends in power of computers (e.g. Moore's Law)
- 3.  $\Longrightarrow$  project date when computers can do as many calculations per second as brain



#### **Compute-centric forecasting models:**

- 1. Estimate # of computations brain per second
- 2. Project forward trends in power of computers (e.g. Moore's Law)
- ⇒ project date when computers can do as many calculations per second as brain



#### **Compute-centric forecasting models:**

- 1. Estimate # of computations brain per second
- Project forward trends in power of computers (e.g. Moore's Law)
- ⇒ project date when computers can do as many calculations per second as brain

#### Seminal paper: Cotra (2020)

► "Biological anchors" or "bio anchors"



#### **Compute-centric forecasting models:**

- 1. Estimate # of computations brain per second
- Project forward trends in power of computers (e.g. Moore's Law)
- 3. ⇒ project date when computers can do as many calculations per second as brain

- ► "Biological anchors" or "bio anchors"
- ► Performance = compute × \$ × algorithms



#### **Compute-centric forecasting models:**

- 1. Estimate # of computations brain per second
- Project forward trends in power of computers (e.g. Moore's Law)
- ⇒ project date when computers can do as many calculations per second as brain

- ► "Biological anchors" or "bio anchors"
- ▶ Performance = compute  $\times$  \$  $\times$  algorithms
  - Temporarily high growth



#### **Compute-centric forecasting models:**

- 1. Estimate # of computations brain per second
- Project forward trends in power of computers (e.g. Moore's Law)
- 3. ⇒ project date when computers can do as many calculations per second as brain

- ► "Biological anchors" or "bio anchors"
- ► Performance = compute × \$ × algorithms
  - Temporarily high growth
  - · Made up [cf Nostalgebraist, 2022]



#### **Compute-centric forecasting models:**

- 1. Estimate # of computations brain per second
- 2. Project forward trends in power of computers (e.g. Moore's Law)
- 3. ⇒ project date when computers can do as many calculations per second as brain

- ▶ "Biological anchors" or "bio anchors"
- ► Performance = compute × \$ × algorithms
  - Temporarily high growth
  - Made up [cf Nostalgebraist, 2022]
- ► Follow up work by Davidson (2022) and Epoch



#### **Compute-centric forecasting models:**

- 1. Estimate # of computations brain per second
- Project forward trends in power of computers (e.g. Moore's Law)
- ⇒ project date when computers can do as many calculations per second as brain

#### Seminal paper: Cotra (2020)

- ► "Biological anchors" or "bio anchors"
- ► Performance = compute × \$ × algorithms
  - Temporarily high growth
  - · Made up [cf Nostalgebraist, 2022]
- ► Follow up work by Davidson (2022) and Epoch

**Cotra (2020):** 2050

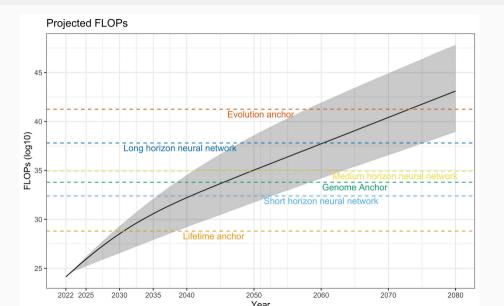
**Cotra (2022):** 2040

**Davidson (2023):** 2043

**Epoch (2022):** 2046

# **Bio anchors (Epoch version)**





# Motivation: financial markets are powerful information aggregators



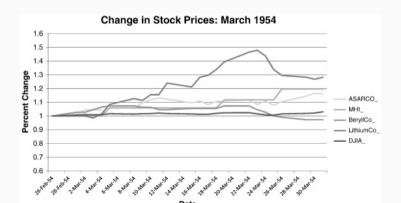
Prices aggregate dispersed information (Hayek 1945); financial market prices especially so (Fama, etc)





# Prices aggregate dispersed information (Hayek 1945); financial market prices especially so (Fama, etc)

► Alchian and the hydrogen bomb

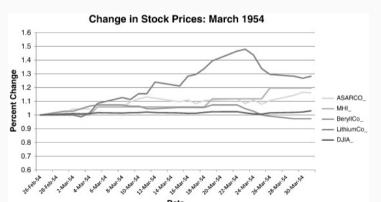


# Motivation: financial markets are powerful information aggregators

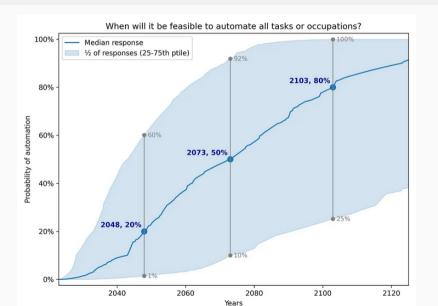


# Prices aggregate dispersed information (Hayek 1945); financial market prices especially so (Fama, etc)

- ► Alchian and the hydrogen bomb
- ► Space shuttle Columbia disaster; election markets; inflation breakevens; ...



# Weinstein-Raun (2024): Grace et al (2024) reanalysis



Euler equation:

$$1 = \beta \delta \mathbb{E}_t \left[ \frac{u'(C_{t+1})}{u'(C_t)} (1 + r_t) \right] \tag{1}$$

Euler equation:

$$1 = \beta \delta \mathbb{E}_t \left[ \frac{u'(C_{t+1})}{u'(C_t)} (1 + r_t) \right]$$
 (1)

**Uncertainty:** risk-neutral expected value. E.g. if  $\Delta C_{t+1} \sim \text{LogNormal}(g, \text{Var})$ , then

$$r_t = \rho + \frac{1}{\sigma}g - \frac{1}{2\sigma^2} Var$$

Euler equation:

$$1 = \beta \delta \mathbb{E}_t \left[ \frac{u'(C_{t+1})}{u'(C_t)} (1 + r_t) \right]$$
 (1)

**Uncertainty:** risk-neutral expected value. E.g. if  $\Delta C_{t+1} \sim \text{LogNormal}(g, \text{Var})$ , then

$$r_t = 
ho + rac{1}{\sigma}g - rac{1}{2\sigma^2} Var$$

**Incomplete markets:** slope unaffected in benchmark

$$r_t = \rho + \frac{1}{\sigma}g - \text{constant}$$

Euler equation:

$$1 = \beta \delta \mathbb{E}_t \left[ \frac{u'(C_{t+1})}{u'(C_t)} (1 + r_t) \right]$$
 (1)

**Uncertainty:** risk-neutral expected value. E.g. if  $\Delta C_{t+1} \sim \text{LogNormal}(g, \text{Var})$ , then

$$r_t = \rho + \frac{1}{\sigma}g - \frac{1}{2\sigma^2} Var$$

**Incomplete markets:** slope unaffected in benchmark

$$r_t = \rho + \frac{1}{\sigma}g - \text{constant}$$

**OLG, recursive preferences:** slope unaffected

Euler equation:

$$1 = \beta \delta \mathbb{E}_t \left[ \frac{u'(C_{t+1})}{u'(C_t)} (1 + r_t) \right]$$
 (1)

**Uncertainty:** risk-neutral expected value. E.g. if  $\Delta C_{t+1} \sim \text{LogNormal}(g, \text{Var})$ , then

$$r_t = 
ho + rac{1}{\sigma}g - rac{1}{2\sigma^2} Var$$

**Incomplete markets:** *slope* unaffected in benchmark

$$r_t = \rho + \frac{1}{\sigma}g - \text{constant}$$

**OLG, recursive preferences:** slope unaffected

► Habit formation

# Expected growth vs. realized growth



