Feedback Loops All the Way Down

The Economics of AI Safety

AI Safety Community Berlin — March 26, 2026

Philipp Petermeier

Twitter / Substack: @infornomics

- Software Engineer with ML focus
  - Python ecosystem, some JS & cloud
- Studied Philosophy & Social Science
  - Extremely curious
  - Happy to engage!
  - Want a cybernetics reading group

www.infornomics.de

p.petermeier@posteo.de

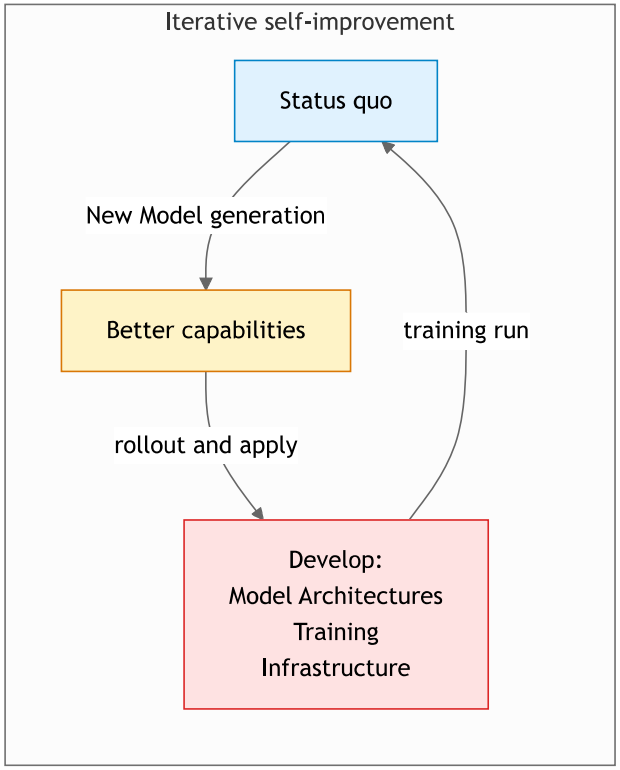
Why focus on economic channel when thinking about social impact?

- Legibility, monetary flows already have a good data gathering infrastructure
- Markets provide a good scaffolding for the social space (rationality assumptions)
- Money is one of the primary ways societies allocate resources
- Declared vs- revealed preferences often differ and markets reveal those differences
- It’s my comfort zone

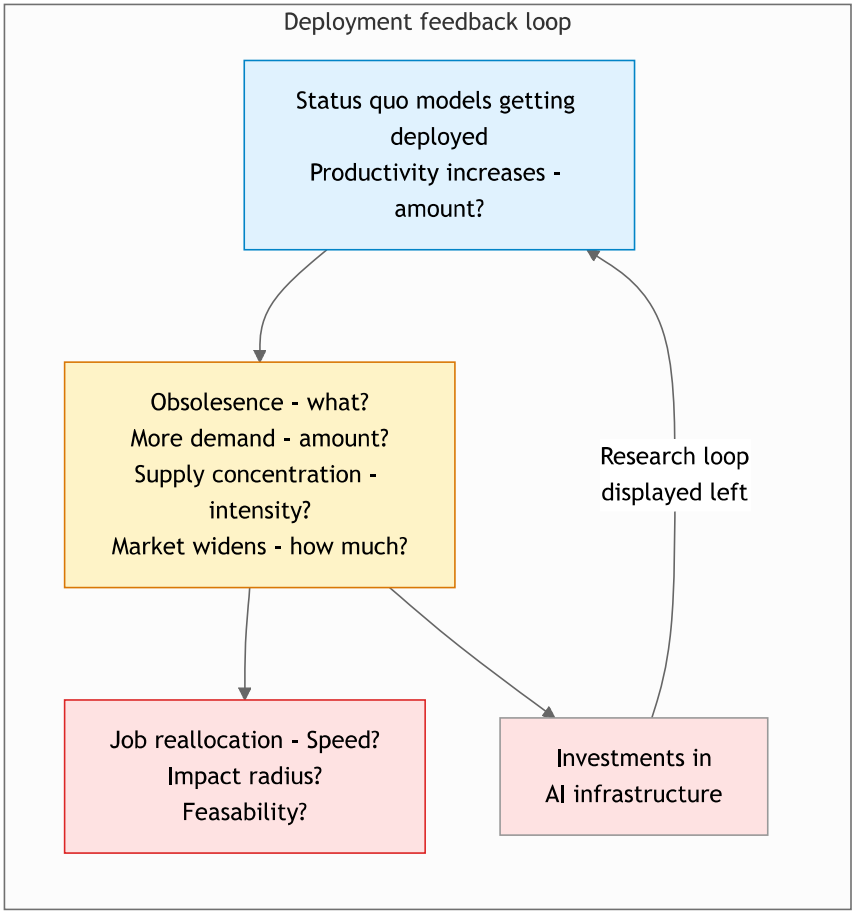
Goals of the talk

- Contextualisation of AI safety concerns into broader society
- Conceptual clarification of AI deployment impact on society
- Getting a clear picture of the status quo
- Generate interest into the topic, give directions if successful

The capabilities feedback loop

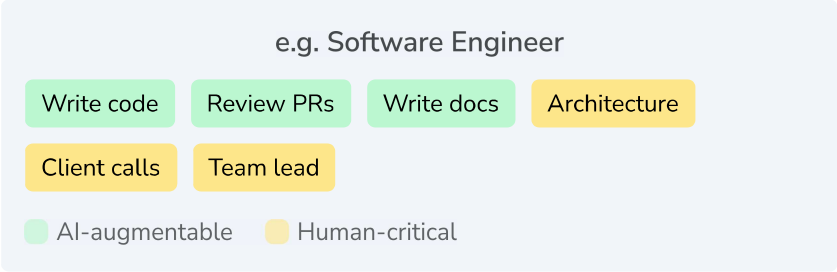


The deployment feedback loop



Jobs as task bundles

A job = a bundle of tasks that, combined, create value. AI doesn't replace jobs — it replaces **tasks**.

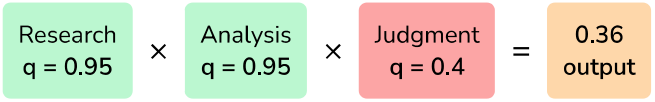


O-Ring Jobs

Named after the Challenger disaster: **one failing component nullifies everything else.**

In high-skill jobs, tasks are **multiplicative**, not additive:

$$\text{Output} = q_1 \times q_2 \times q_3 \times \ldots \times q_n$$



Implication: If AI boosts most tasks but leaves one human-critical task unimproved — or introduces over-reliance there — gains are capped or reversed.

Labor enhancing vs. displacing

A technology is **labor enhancing** if the worker remains critical to the output — their productivity rises, wages follow. It is **labor displacing** if the worker is substituted out entirely.

Enhancing

AI handles subtasks → worker produces more → still needed → wages rise

Displacing

AI handles the whole task → worker substituted → headcount falls

Exposure ≠ displacement. Most metrics (e.g. O*NET task exposure scores) measure which jobs *could* be affected — not whether workers are actually replaced. Acemoglu's task-based model is the rigorous framework for making this distinction.

Jaggedness

Human competence has a **shape** — a recognisable profile of skills that fits the task-bundle of a job. We have evolved good intuitions for reading that shape in other people.

AI capabilities don't follow that shape. They can exceed expert-level performance on some sub-tasks while failing at things a junior would handle easily.

Jevons Paradox

William Jevons (1865): as steam engines got more efficient, coal consumption went **up** — because cheap energy made more uses economically viable.

Applied to AI: if coding gets 10× faster, companies may just build 10× more software. The demand for the task *expands to absorb the efficiency gain*.

Full rebound	Partial rebound	No rebound
No net job loss — demand scales with productivity	Some displacement, but fewer workers do more	Gains captured by capital — labor share falls

Historical pattern favours rebound — ATMs increased bank teller headcount; looms grew the textile workforce. **But distribution of gains is a separate question.**

From concepts to data

These are not abstract questions. The economic feedback loop is already running — **we just can't see it clearly yet.**

Blast radius — which occupations, geographies, and career stages are actually affected, and by how much?

Timing — are we seeing a temporary shock, a structural shift, or a J-curve that hasn't bent yet?

Distribution — do productivity gains go to workers, firms, or capital owners?

Just as safety researchers need **interpretability** to understand what a model is doing, economists need **measurement infrastructure** to understand what the economy is doing. We're building both at the same time.

From Vibes to Data

What do we actually know?

Micro

RCTs & firm studies — what happens when workers get AI tools

Revealed Usage

What people actually do with AI (Anthropic Economic Index, OpenAI Signals Data)

Macro

Productivity, employment, wages at the aggregate level

We're moving from speculation to measurement, but measurement infrastructure is still being built.

This is the **"pre-interpretability" era** of AI economics.

AI at the Workbench

The RCT picture

Brynjolfsson, Li & Raymond (2025, QJE) 5,179 customer support agents. **14% productivity gain** average, **34% for novices**, minimal for experts. Gold standard — top journal.

Cui et al. (2025) — Three RCTs at Microsoft, Accenture, Fortune 100 ~5,000 developers. **26% increase** in completed pull requests.

Noy & Zhang (2023, Science) — Writing tasks. Large gains for lower performers.

Dell'Acqua et al. (2023, HBS) — BCG consultants and the **"jagged frontier"**: inside the frontier, AI helps. Outside it, AI *hurts*. People can't tell which is which.

Pattern: Large gains, especially for less-skilled workers, but with a jagged frontier. The overreliance problem is an **alignment-adjacent issue**.

The Anthropic Economic Index

What people actually do with AI — not what they say

Concentration Top 10 tasks = 24% of conversations Computer & math = ~36% Deep and narrow, not broad-based.

Augmentation vs Automation Consumer (Claude.ai): 52% augment / 45% automate API traffic: automation-dominant The consumer and enterprise stories diverge.

Geographic unevenness Usage ∝ GDP per capita globally Within US: workforce composition > income

Economic Primitives (V4, Jan 2026) Complex tasks get *larger* speedups:

- College-level → **12x** speedup
- High-school level → **9x** speedup
- But complex tasks have **lower success rates**

The safety community worries about systems that are powerful but unreliable. The economic data shows exactly that: AI is most transformative on complex tasks — and that's where it fails most.

Open-source on HuggingFace: huggingface.co/datasets/Anthropic/EconomicIndex — 4 reports, Feb 2025 – Jan 2026

The Productivity Debate

Is the J-curve finally bending upward?

The Optimist — Brynjolfsson (FT, Feb 2026)

- US productivity ~**2.7%** in 2025 (2× prior decade)
- GDP robust (3.7% Q4) while jobs revised sharply down
- "Harvest phase" of the J-curve beginning
- Furman (previously skeptical) now agrees

The Skeptic — Acemoglu (Econ Policy, 2025)

- Task-based model + Hulten's theorem
- At most **0.66% TFP** gain over 10 years
- Only ~20% of tasks exposed; ~23% profitably automatable
- Capital-labor income gap widens regardless

CEPR middle ground: EU firm-level data shows ~4% productivity increase from AI adoption, no short-run job losses — but gains concentrate in medium-to-large firms with intangible assets and human capital.

The Canaries

Early warning signals

Brynjolfsson, Chandar & Chen (Nov 2025) — "Canaries in the Coal Mine"
ADP payroll data (millions of workers): early-career employment in AI-exposed occupations fell ~**13%** on a relative basis since late 2022.

Brynjolfsson et al. (Feb 2026) — Minimum wages + robots
Automation pressure from below simultaneously. Blue-collar squeeze from robots, white-collar squeeze from AI.

CEPR / Dean Baker (Jan 2026) — The AI bubble argument
Labor compensation / consumption ratio fell to 71.6% — roughly **$1 trillion** in consumption supported by stock market wealth. If bubble bursts: ~3% of GDP in lost consumption.

The Theoretical Frontier

Demand collapse, agent economies, and comparative advantage

Séb Krier

DeepMind · Policy

"The Cyborg Era" (Jan 2026): Comparative advantage persists longer than expected. Full substitution requires ASI so superior humans add negative value + abundant compute + zero demand for human involvement. Very stringent.

Multi-agent systems (MR, Dec 2025): "Most innovations are the product of social organisations, not a single genius savant." Products, not models, create value.

Alex Imas

UChicago Booth · "Ghosts of Electricity"

Demand collapse (Jan 2026): If automation kills wages, who buys the output? Island economy model → full automation can lead to negative growth.

Agent alignment drift (Feb 2026): "Does Overwork Make Agents Marxist?" — agents accumulate experience that shifts orientations. Skills files propagate changes. Bridges econ ↔ safety.

Anton Korinek

UVA · NBER · Anthropic Advisory

"Nine Grand Challenges" framework: growth, innovation, distribution, decision-making, geoeconomics, information, safety, well-being, transition.

Public finance (Jan 2026): What do you tax when labor income erodes?

With Stiglitz: Can we steer AI toward labor-complementing innovation?

We Need Interpretability for the AI Economy

AI Safety	AI Economics
Model interpretability	Economic measurement infrastructure
Evals & benchmarks	Anthropic Economic Index, BLS projections
Alignment monitoring	Labor market monitoring, distributional tracking
Red-teaming	Stress-testing scenarios (demand collapse, bubble)
Governance frameworks	NBER Econ of Transformative AI, CEPR AI RPN

Institutions being built right now: Anthropic Economic Index (open-source) · NBER volume (Agrawal, Brynjolfsson, Korinek 2025) · CEPR AI Research & Policy Network (Korinek chair) · EconTAI Initiative at UVA · BLS incorporating AI into 10-year projections

Just as you wouldn't trust a model you can't interpret, you shouldn't trust an economic transition you can't measure.

Where to Go from Here

The economic feedback loop is **real, measurable, uneven, and accelerating.**

The safety community's conceptual tools **apply.**

The institutions for legibility are **being built now.**

Resource list: Annotated, tiered by depth — papers, data, substacks, podcasts.

Open data: huggingface.co/datasets/Anthropic/EconomicIndex

Key follows: @alexolegimas · @sebkrier · @akorinek · @jasonfurman · @erikbryn

Key reads: Imas's "Ghosts of Electricity" · Korinek's korinek.com/research

Questions?