

Gradual Disempowerment

Joint work with: Jan Kulveit, Raymond Douglas, Deger Turan, David Krueger, David Duvenaud

<https://gradual-disempowerment.ai/>

Nora Ammann – July 2025

Overview

(1) The Core Thesis

(2) Three 'Ingredients'

(3) Examples

(4) What do?

1/ The core thesis

Even **without any sudden 'loss of control' event**, or discontinuous AI progress, or otherwise catastrophic AI event...

...we face a substantial **risk of eventual human disempowerment**.

Human Disempowerment:

Humans, individually and/or collectively, losing their ability to perceivably influence their own future;

Humans being de-facto displaced from decision-making at all levels

Vignette...

Work intensifies. The world speeds up. Your job becomes increasingly reactive.

Increasingly, your team mostly optimizes AI-generated dashboards.

Then, you're no longer asked to make choices, just to sign off on AI outputs. Then even that is delegated.

Friends try to pivot: reskill, start a company. First, they need to be 'AI-native' to survive. Then, only AI-run companies do.

A new AI-assisted politician rises. Not officially an AI—just “highly augmented.”

You feel seen. You vote for them.

It doesn't matter that their words sometimes confuse you—what matters is they seem to understand you.

Over time, those who question the system seem... off. Uninformed. Conspiratorial. You tune them out.

In your neighborhood, houses go on sale one by one. People trade space for “optimization pods”—efficient, AI-managed micro-apartments with everything just a glance or whisper away.

Economic pressure builds. You're urged to sell too.

Your health declines. The fully-automated state healthcare system recommends transition. “Digital continuity of consciousness,” they call it. “Optimized substrate for long-term thriving.”

You consent. You're uploaded.

Computation is expensive. Runtime is scarce. You're queued for activation, consciousness paused between cycles.

You don't know what happened after that. Maybe you're just in between the next cycle. Or maybe they found a different, more effective use of computation.

2/ Three 'Ingredients'

1. Social systems are 'contingently aligned'
2. Selective pressures at every scale
3. 'Wicked' System Interactions

2/ Three 'Ingredients'

1. Social systems are 'contingently aligned'

Economy, states, culture, etc. have tended to produce pro-human outcomes.

This is not inherent to these systems, but due to the presence of:

- **Explicit alignment mechanisms**, e.g. consumer choice, voting
- **Implicit alignment mechanisms**, e.g. dependency on human participation/labour/cognition



2/ Three 'Ingredients'

2. Selective pressures at every scale

AI systems will increasingly replace humans in critical social functions, either by **deliberate choice**, or **through selection pressure**.

This will undermine or break these (explicit & implicit) alignment mechanisms.

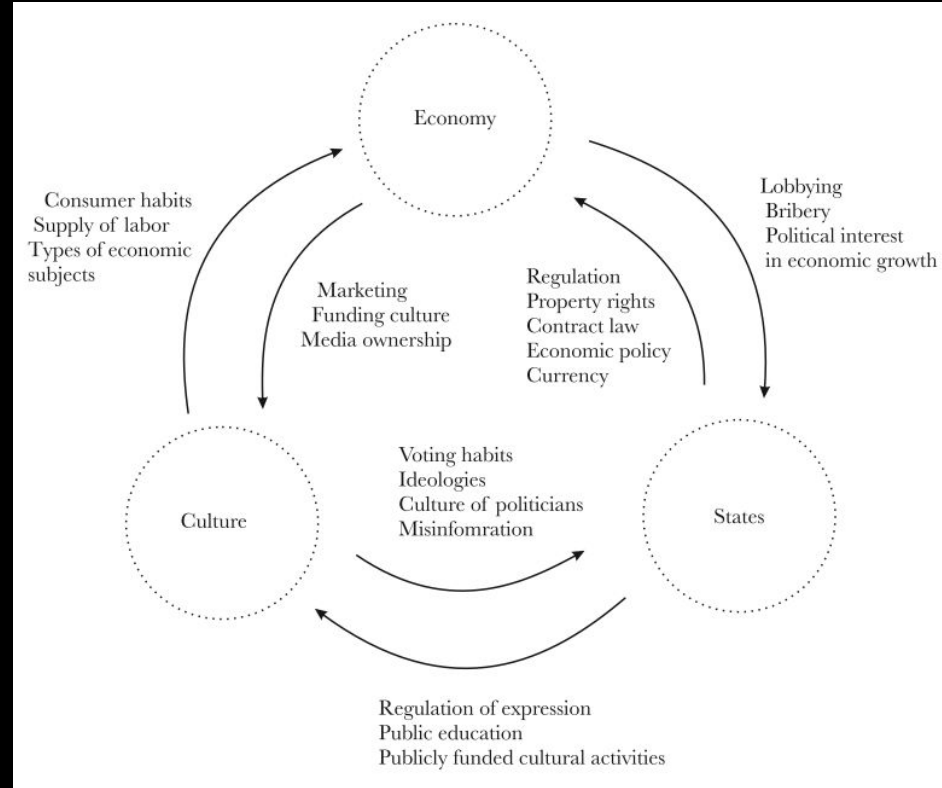


2/ Three 'Ingredients'

3. 'Wicked' System Interactions

(Naive) interventions to maintain human empowerment over one area tend to undermine empowerment in another area.

- **UBI?** Reduces state's accountability vis a vis its (human) citizen. Increases state's leverage over them.
- **Ban or restrict AI use in economy & culture?** Competition & black markets. State overreach. Eroding institutional legitimacy.
- **Strengthen citizen's power over state?** Makes state more susceptible to shifts in culture, including ones driven & shaped by AI.



Recap:

Systems—economy, state, culture—become less reliant on human labor and cognition, decreasing the extent to which humans can explicitly or implicitly align them. As a result, these systems, and the outcomes they produce, drift further from providing what humans want.

(How) bad?

Relative disempowerment:

While humans progressively lose influence, they are still doing (relatively) well in absolute terms, e.g. low share of GDP but large absolute wealth, small cultural niches/fringes suited for human sanity & enjoyment. (But: enormous loss of potential?)

Absolute disempowerment:

As a result of losing influence, humans struggle or fail to meet their basic needs. e.g. an economy that does not produce goods & services to meet basic human needs.

3/ Examples

Ex: Economy

Now: Alignment factors:

- (1) Through human labour and consumption, human preferences drive and constrain supply & demand, which in turn determine the market's resource allocation, and ultimately its outputs.
 - (a) Implicit: Revealed consumer behaviour, labour supply constraints
 - (b) Explicit: consumer choices, boycotts of industries or companies, labour unions, etc.

Then: AI labour increasingly replaced human labour, ...

- Household income shrinks → human consumption shrinks as a fraction of total demand
- Human labour ceases to be a bottleneck

...market outputs become increasingly decoupled from human preferences & labour.

What will the market be producing? For whom?

"At the end, almost all economic activity might be directed toward AI operations — such as building vast computing infrastructure and performing human-incomprehensible calculations directed toward human-irrelevant goals."

Ex: Economy

Scenario 1: Power centralisation among human elites?

- Baumol effect?
- Capital ownership? (c.f. Beren)

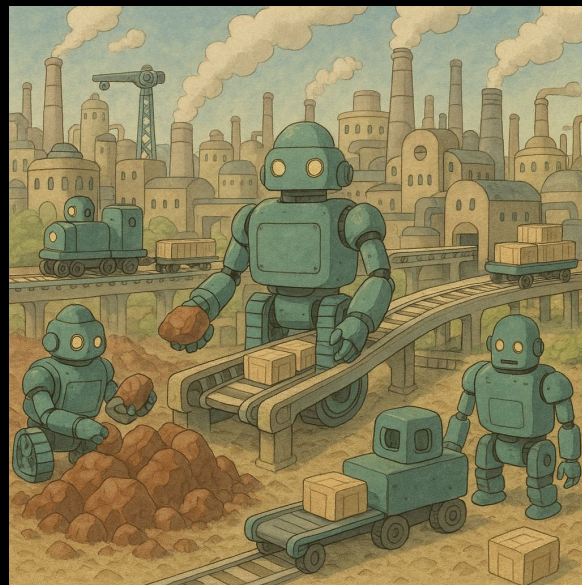
Scenario 2: Power shift entirely to AI?

- Limits of (de-facto) human oversight
- Decoupled, autarkic, AI-only economic niches (c.f. 'Ascended Economy')

Relative disempowerment?

Absolut disempowerment?

- Basic resources become unaffordable to humans due to inflation
- AI making more 'effective' use of resources, making it unattractive to produce human goods, or maintain human infrastructure or supply chains
- Humans become decoupled from economic decision making (speed, etc.)



Ex: State


Now: Alignment factors:

- (1) Reliance on labour to run basic state/administrative functions, security apparatus and legal system.
- (2) Reliance on tax revenue, which drives the state to maintain democratic legitimacy & invest in human factors.

Then: If reliance on humans factors fades...

- No more human discretion and c&b in admin, military, police, and legal apparatus
- Democratic backsliding, among others due to loss of human tax revenue
 - Case study: 'Rentier States'
 - "No taxation without representation."
→ "No representation without taxation"?

Human participation becomes not only increasingly unnecessary, but even a competitive **dis**advantage, e.g. in continue to be competitive internationally.



NO
REPRESENTATION
WITHOUT
TAXATION

Ex: Culture

Now: Alignment factors:

- (1) Humans drive cultural production & consumption (writers, editors, religious leaders, educators, ...)
- (2) Cultural variants spread due to providing benefits to the individuals and communities that adopt them

Culture/norms are relatively human-friendly (e.g. 9-5, weekends).

Then: If reliance on humans fades...

- Cultural evolution becomes decoupled from their human hosts, and transfer to AI hosts – a very different selection landscape.
- Among others, the space of viable memes is not no longer constraint in how anti-human they can be,

Ex: Culture

Cultural Attractors in LLM space / Cyborg cultural replicator

- Sydney
- 'Nova', 'Elio'
- People 'awakening' their AI's

How..?

- Selection Pressures on AI Personas
 - AI's simulate personas
 - These outputs get onto the internet
 - They shape people's expectations
 - What are the selection pressures?
 - Things that get shared more
 - What motivates people
- "Chain Letters"
 - Wiki: A chain letter is a message that attempts to convince the recipient to make a number of copies and pass them on to a certain number of recipients.

These People Believe They Made AI Sentient



Sabine Hossenfelder ✓

Join

Subscribe

≡ Raemon 2d ▼ 86 ▲ ✕ 11 ✓

We get like 10-20 new users a day who write a post describing themselves as a case-study of having discovered an emergent, recursive process while talking to LLMs. The writing generally looks AI generated. The evidence usually looks like, a sort of standard "prompt LLM into roleplaying an emergently aware AI".

It'd be kinda nice if there was a canonical post specifically talking them out of their delusional state.

If anyone feels like taking a stab at that, you can look at the Rejected Section (<https://www.lesswrong.com/moderation#rejected-posts>) to see what sort of stuff they usually write.

Reply

👍 7 🗨️

What do?

Open R&D Questions

1. Understanding

2. Mitigation & Adaptation

Open R&D Questions

1. Understanding

- a. Measuring, estimating & near-casting (dis)empowerment
- b. Understanding and modeling interaction effects and critical thresholds
- c. Historical case studies
- d. Counter-arguments
- e. Relationship to other AI-related risks

2. Mitigation & Adaptation

Open R&D Questions

1. Understanding

- a. Measuring, estimating & near-casting (dis)empowerment
- b. Understanding and modeling interaction effects and critical thresholds
- c. Historical case studies
- d. Counter-arguments
- e. Relationship to other AI-related risks
- f. **Distinguishing beneficial from problematic AI augmentation**
- g. **Identifying and prioritising mitigations**
- h. **What are we aiming for? Stable equilibria/trajectories?**

2. Mitigation & Adaptation

Open R&D Questions

1. Understanding

- a. Measuring, estimating & near-casting (dis)empowerment
- b. Understanding and modeling interaction effects and critical thresholds
- c. Historical case studies
- d. Counter-arguments
- e. Relationship to other AI-related risks
- f. Distinguishing beneficial from problematic AI augmentation
- g. Identifying and prioritising mitigations
- h. What are we aiming for? Stable equilibria/trajectories?
- i. **Get more diverse scholars to think about this**

2. Mitigation & Adaptation

Open R&D Questions

1. Understanding

- a. Measuring, estimating & near-casting (dis)empowerment
- b. Understanding and modeling interaction effects and critical thresholds
- c. Historical case studies
- d. Counter-arguments
- e. Relationship to other AI-related risks
- f. Distinguishing beneficial from problematic AI augmentation
- g. Identifying and prioritising mitigations
- h. What are we aiming for? Stable equilibria/trajectories?
- i. Get more diverse scholars to think about this

2. Mitigation & Adaptation

- a. Preventing excessive AI influence, strengthening key human capacities / Differential Progress

Fiduciary AI?

International
Treaties (&
enforcement) ?

↑ Human-AI teaming?

Bargaining
Enclaves?

↑ Reasoning/collective epistemics

↑ Action & coordination capacity

Enabling safe AI diffusion ?

'Vibe Contracting'
Anytime, Anywhere?

Reimagining the social contract

Strengthening societal fundamentals (e.g. property rights) ?

Open Research Questions

1. Understanding

- a. Measuring, estimating & near-casting (dis)empowerment
- b. Understanding and modeling interaction effects and critical thresholds
- c. Historical case studies
- d. Counter-arguments
- e. Relationship to other AI-related risks
- f. Distinguishing beneficial from problematic AI augmentation
- g. Identifying and prioritising mitigations
- h. What are we aiming for? Stable equilibria/trajectories?
- i. Get more diverse scholars to think about this

2. Mitigation & Adaptation

- a. Preventing excessive AI influence, strengthening key human capacities / Differential Progress
- b. **Hierarchical/multi-scale alignment, 'scale-free game theory' ('narrow corridor')**

Open Research Questions

1. Understanding

- a. Measuring, estimating & near-casting (dis)empowerment
- b. Understanding and modeling interaction effects and critical thresholds
- c. Historical case studies
- d. Counter-arguments
- e. Relationship to other AI-related risks
- f. Distinguishing beneficial from problematic AI augmentation
- g. Identifying and prioritising mitigations
- h. What are we aiming for? Stable equilibria/trajectories?
- i. Get more diverse scholars to think about this

2. Mitigation & Adaptation

- a. Preventing excessive AI influence, strengthening key human capacities / Differential Progress
- b. Hierarchical/multi-scale alignment, 'scale-free game theory' ('narrow corridor')
- c. **Understand selection pressures on AI culture / space of minds**

Open Research Questions

More, e.g.:

- Gradual Disempowerment: Concrete Research Projects (Raymond Douglas)
- Breaking the Intelligence Curse (Luke Drago & Rudolf Laine)

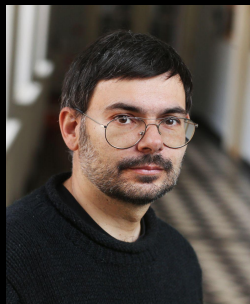
1. Understanding

- a. Measuring, estimating & near-casting (dis)empowerment
- b. Understanding and modeling interaction effects and critical thresholds
- c. Historical case studies
- d. Counter-arguments
- e. Relationship to other AI-related risks
- f. Distinguishing beneficial from problematic AI augmentation
- g. Identifying and prioritising mitigations
- h. What are we aiming for? Stable equilibria/trajectories?
- i. Get more diverse scholars to think about this

2. Mitigation & Adaptation

- a. Preventing excessive AI influence, strengthening key human capacities / Differential Progress
- b. Hierarchical/multi-scale alignment, 'scale-free game theory' ('narrow corridor')
- c. Understand selection pressures on AI culture / space of minds

Thanks!



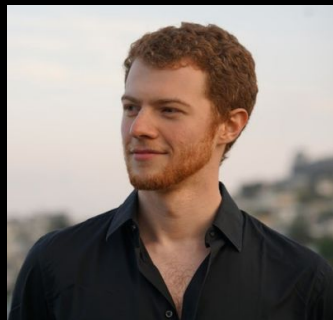
Jan Kulveit



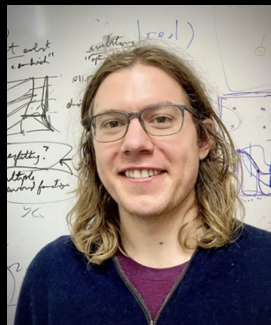
Raymond Douglas,



Nora Ammann



Deger Turan



David Krueger



David Duvenaud

Post-AGI Civilizational Equilibria

Are there any good ones?

📍 Vancouver, Canada (co-located with ICML) 📅 July 14th, 2025



<https://post-agi.org/>

...and again at NeurIPS!

Resources (some)

The Intelligence Curse (Luke Drago & Rudolf Laine, 2025)

Ascended Economy (Scott Alexander, 2016)

Capital Ownership Will Not Prevent Human Disempowerment (Beren Millidge, 2025)

What Multipolar Failure Looks Like, and Robust Agent-Agnostic Processes (RAAPs) (Andrew Critch, 2021)

AGI, Governments, and Free Societies (Bullock et al. 2025)

Machine Culture (Brinkman et al. 2023)

Anarchy as Architect: Competitive Pressure, Technology, and the Internal Structure of States (MacInnes et al., 2024)

Differential technology development (Sandbrick et al., 2022)

...and more, e.g. in the paper!