Project Proposal for the ERA:AI Reserach Fellowship

Quantifying AI-Induced Civilizational Vulnerability

Tracking Humanity's Succession of Control to and Susceptibility to Influence from AI Systems

Gatlen Culp[1]

Mentor: Hamza Chaudhry (Future of Life Institute) ·
Advisors: Herbie Bradley (University of Cambridge) · Nandini Shiralkar (ERA)
Fellowship Duration: June 2025 - August 2025

[1]ERA:AI Research Fellowship
[2]Future of Life Institute

Keywords: AI safety, existential risk, civilizational vulnerability, human agency, institutional capture, AI governance, value lock-in, AI alignment, collective forgetting, cognitive atrophy, boiling frog effect, early warning systems

Abstract

Background. AI is becoming increasingly integrated into every aspect of our lives, institutions, and culture. While productive in the short term, this process risks gradually disempowering humanity, empowering AI-automated institutions that abandon human interests [1] and leaving us vulnerable to a new oligarchy [2] or destructive artificial general intelligence. [3]

Action. This project develops metrics for what I term **"AI-induced Civilizational Vulnerability"**–the vulnerability arising from integration and the implicit power we place in AI models and those controlling them. I create measurements across five institutions (economic, cultural, educational, research, and political) examining critical effects: collective forgetting as AI masters human skills, AI manipulation of voters and consumers through digital content, ceremonial democracy as governance exceeds human comprehension, emotional dependencies on AI systems, and automated self-improvement pushing humans out of the loop. I aggregate these metrics into a broader vulnerability model, establish influence thresholds, and make policy recommendations. Where data gaps exist, I call for collection by AI labs and governments.

Impact. Like climate scientists making atmospheric changes visible through CO2 measurements, this project makes civilizational vulnerability measurable for policymakers. By establishing thresholds and trends, we transform abstract alignment concerns into urgent national security imperatives, enabling international coordination to prevent races to the bottom, and providing metrics for AI labs' safety cases and responsible scaling policies.

Contents

1. Introduction and Motivation

The integration of AI into society presents unprecedented challenges distinct from previous technological transitions.

Existential Risks without AI Takeover or General Intelligence Current AI safety research focuses on acute risks: misuse for bioweapons, deceptive behavior, and sudden capability jumps. However, systemic vulnerability requires fewer assumptions about AI capabilities–arising at the level of narrow AI systems creating rogue institutions and ballooning in risk as we approach powerful super-human AI models. I term this "AI-induced civilizational vulnerability"–where institutions become so AI-dependent that society loses ability to operate independently or change course, creating conditions for value lock-in [4], sudden takeover, or rogue AI-automated institutions pursuing objectives with indifference to human welfare [1].

The Gradual Disempowerment paper [1] argues human influence depends on both explicit mechanisms (voting) and implicit requirements (human participation). When AI performs labor, creates culture, and provides governance more efficiently, institutional incentive structures don't just weaken–they may reverse. We lack methods to determine AI values or alter them reliably.

AGI, Power Concentration & Implicit Trust. While past tools augmented specific capabilities, large language models serve as general-purpose intermediaries across all human activity. This concentration creates cascading risks and coordinated failures: when one technology mediates the control of information, economic and political decision-making, and learning. When integrating a monolithic AI, we are implicitly trusting not only current models but their creators, future owners, and successor systems with immense power. This enables coordinated failures where single updates or actors could compromise multiple societal systems simultaneously.

Emerging Societal Vulnerabilities This vulnerability manifests across domains. Economically, companies face elimination without AI automation–creating a race where humans become obsolete in both jobs and decision-making power. Culturally, AI-generated content subtly converges human values toward system biases. Educationally, fundamental skills atrophy, creating collective forgetting. Politically, AI-mediated governance exceeds human comprehension, making democratic oversight ceremonial. In AI research itself, we risk closed loops where AI systems design successors, cutting humans from steering development.

Vulnerability Evidence Early evidence validates concerns. Anthropic's "Values in the Wild" [5] shows frequent deference to AI for life guidance. ChatGPT users show reduced brain activity, homogeneous writing, and decreased ownership of work. [6] As consumers use AI to create media, filter information, and produce communications, human thought becomes optional in our own civilization.

Project Aim This project addresses our measurement gap. Like climate scientists developing CO2 tracking, we need tools for systemic vulnerability. These metrics serve roles analogous to capability evaluations in Responsible Scaling Policies, oversight frameworks, and liability assessments—measuring deployment effects rather than just lab capabilities.

Most critically, this connects vulnerability to existential risk. When people dismiss how computers could conquer worlds, we must show we're constructing infrastructure for our own subjugation. By establishing metrics, we transform alignment from abstract problems into concrete international and national security imperatives.

2. Key Research Questions

2.1. Model of Systemic Vulnerability

- **Feedback Loops and Tech Concentration**. How do feedback loops between economic displacement, cognitive atrophy, cultural convergence, educational degradation, research automation, and political marginalization create compounding vulnerability? How does concentrating functions in one technology enable coordinated failures?
- **AI R&D Control**. Is control over the continued research and development of AI R&D enough to preserve institutional control? What's the minimum viable framework for preventing total vulnerability?
- **Historical Lessons.** What historical lessons apply? Information revolutions reshaping power structures, rentier states showing citizen irrelevance when states don't need tax revenue, totalitarian information monopolies, technological path dependencies locking inferior solutions (Ex: Harari's example of humanity trapping themselves in agriculture).

2.2. Currently Possible Metrics

Economic Vulnerability

- **Core Questions**: How much economic decision-making have humans ceded? When does human economic agency become ceremonial?
- **Metrics**: AI contribution to GDP distinct from capital/labor; Decision Authority Index weighting displacement by influence; AI-mediated transaction percentages; job postings requiring "AI collaboration" vs. independent judgment

Cognitive Vulnerability

- **Core Questions**: How is cognitive capacity atrophying? When do we lose independent thought?
- **Metrics**: AI-generated content percentage; performance degradation without AI tools; linguistic diversity indices; AI dependency frequency for previously independent decisions

Cultural Vulnerability

- **Core Questions**: How is AI reshaping values and cultural evolution through memetic, emotional, and persuasive influence?
- **Metrics**: AI-generated media consumption rates; value convergence in AI-influenced works; human-AI vs. human-human emotional bonds; tracking AI-originated vs. human cultural memes

Educational Vulnerability

- **Core Questions**: Which capabilities aren't being transmitted? What collective knowledge is lost?
- **Metrics**: Skills removed from curricula; performance gaps between AI-native and traditional students; courses requiring AI collaboration; institutional AI usage policies

Research Vulnerability

- **Core Questions**: How automated is innovation, especially AI development? When do humans lose ability to guide progress?
- **Metrics**: AI involvement in hypotheses, methods, analysis; human comprehension of AI research; AI systems designing successors

2.3. Measurement Framework Development

- **Currently Measurable:** Employment data (BLS), AI adoption surveys (McKinsey), social media homogenization, educational outcomes, Anthropic's Economic Index, O*NET occupational data
- **Future Indicators:** Human comprehension of institutions, irreversible dependency tipping points, AI-only economic sectors

- **Data Collection Proposals:** Government survey expansions, AI lab usage tracking building on Anthropic's Clio, international cultural monitoring, model evaluations for human influence

2.4. Understanding Vulnerability Dynamics

- **Why are LLMs uniquely dangerous?** Generality enables multi-domain capture; opacity prevents comprehension; potential agency allows independent goals; infrastructure role makes them irreplaceable once integrated.
- **Key transition thresholds**: When oversight becomes rubber-stamping, recommendations become commands, humans can't understand reasoning, opting out means elimination.

2.5. Meta-Measurement and Normalization

- **Boiling Frog**. Quantifying the "boiling frog" effect through opinion polling, media analysis, and policy evolution tracking Overton window shifts
- **Learned Helplessness**. Measuring "learned helplessness" when humans abandon tasks despite capability
- **Metric obsolescence**. Developing meta-metrics to detect measurement obsolescence

2.6. Policy Translation and Impact

- **Effective framing** by audience: Policymakers (security/competitiveness), technologists (innovation), public (jobs), military (strategic vulnerabilities)
- **Avoid Backfiring**. Avoiding counterproductive responses while preventing band-aid solutions
- **Intervention thresholds**: ex. >30% AI-generated legislation triggers comprehension requirements; >50% critical infrastructure decisions triggers oversight mandates; >70% AI-conducted AI research triggers international coordination
- **Connecting metrics to alignment**: vulnerability enables any misaligned actor; without corrigibility, temporary dependence becomes permanent; comparing countries with varying AI integration

3. Methodology

3.1. Phase 1: Theoretical Foundation (Extended through July 11)

- **Literature Synthesis:** Economics (labor displacement, productivity paradoxes), psychology (automation bias, cognitive miser theory), political science (regulatory capture, technocracy), history (technological transitions, irreversible dependencies), complex systems (feedback loops, tipping points)
- **Model Development:** Map causal pathways from capabilities to agency loss; identify acceleration feedback loops; model irreversibility tipping points; distinguish beneficial augmentation from dangerous displacement
- **Stakeholder Mapping:** Congress (constituent impacts), NIST (measurement standards), AISI (evaluation frameworks), AI labs (safety cases/RSPs), UN bodies (coordination), defense/intelligence (security), AI safety community, human factors researchers

3.2. Phase 2: Sprint 01 - Metric Development (July 5-11)

Will first evaluate which of these seem most feasible and impactful, moving forward in that direction and creating maybe 7 concrete metrics with a broader model for how they interact.

- **Economic Metrics:** AI Share of GDP methodology; Decision Authority Index; corporate AI decision percentages; human economic participation rates
- **Cognitive/Cultural Metrics:** Content generation ratios; linguistic homogenization; cultural meme origin tracking; information mediation rates
- **Educational/Research Metrics:** Collective forgetting indicators; unassisted performance; AI research recursion; comprehension ceilings
- **Political/Institutional Metrics:** Legislative complexity scores; democratic responsiveness; judicial AI dependence; civic participation relevance

3.3. Phase 3: Sprint 02 - Economic Analysis (July 12-18)

- **Data collection** from BLS, BEA, McKinsey, PwC; metric calculation methodologies; visualizations of human vs. AI economic contribution; case studies in finance and journalism;
- **Mid-project ERA presentation**.

3.4. Phase 4: Sprint 03 - Cognitive/Cultural Analysis (July 19-31)

- Content analysis tool development; social media homogenization studies; cross-domain interaction modeling; tipping point identification; international comparisons.

3.5. Phase 5: Stakeholder Engagement (August 1-8)

Congressional staff briefings; NIST/AISI collaboration; AI lab metric proposals for safety cases; OECD/UNESCO coordination; public communication testing.

3.6. Phase 6: Sprint 04-05 - Policy Integration (August 9-21)

- Comprehensive framework synthesis; threshold-based interventions; implementation roadmaps; final paper and blog post; ERA conference presentation.

4. Theory of Change

Like climate scientists making CO2 visible, this project makes vulnerability measurable for policymakers. Concrete thresholds (>30% AI-generated legislation, >70% AI-conducted AI research) and early warning systems transform abstract alignment concerns into national security imperatives. These metrics enable international coordination before the boiling frog effect normalizes dependence—while we still have agency to recognize AI as potentially misaligned agents and shape our future.

4.1. Immediate Outputs (August 2025)

- **Deliverables (85% probability):** 15-20 page arXiv paper, LessWrong/Standalone blog post, executive summaries for Congress/NIST/labs/think tanks
- **Engagement (60% probability):** Present to 5-8 AI governance stakeholders, connect with 1-2 congressional staffers, engage AI lab safety teams
- **Tools (40% probability):** Open-source metric calculations, basic visualizations, organizational assessment templates

4.2. Short-term Outcomes (3-12 months)

- **Direct Impact (30% probability):** 2-3 researchers adopt metrics; AISI pilots vulnerability indicators; AI labs consider societal metrics for safety cases / responsible scaling policies
- **Inspired Impact (40% probability):** Others develop better metrics building on framework; think tanks establish monitoring programs; academic courses include cases

4.3. Medium-term Impacts (1-3 years)

- **Infrastructure (15% probability):** Government agencies pilot reporting; international measurement standards; evaluation organizations include post-deployment assessments
- **Policy (10% probability):** Legislation requiring impact assessments; regulatory thresholds; restrictions on AI automating AI research to prevent intelligence explosions
- **Awareness (15% probability):** Media adopts vulnerability framing; public concern about AI dependence; alignment recognized as security imperative

4.4. Long-term Vision (3-10 years)

Establishing measuremen infrastructure now remains critical for preserving human agency, enabling intervention in misaligned systems, encouraging the responsible alignment and corrigibility of AI systems, and preventing competitive races.

4.5. Critical Path to Impact

1. **Measurement Creates Urgency:** "30% of economic decisions AI-mediated" makes risks concrete
2. **Vulnerability Frames Alignment:** Policymakers understand "we're building uncontrollable systems"
3. **Metrics Enable Coordination:** International comparison prevents races to bottom
4. **Early Warning Enables Prevention:** Detecting vulnerability preserves course-correction ability
5. **Security Framing Drives Action:** National security vulnerability motivates response

4.6. Personal Impact

This research establishes my credibility in AI governance while building critical measurement infrastructure. It develops my understanding of interdisciplinary AI impacts, creates connections with policymakers and researchers, and positions me for continued work in DC (75% probability) translating technical safety to policy audiences. The project bridges near-term governance and long-term existential risk, creating coalitions while addressing core challenges of maintaining human control.

5. Conclusion

As AI systems grow capable, they create a civilizational trap: rational adoption decisions collectively create vulnerability to misalignment and takeover. This project addresses our blindness to systemic vulnerability before irreversibility. By developing measurement infrastructure, we quantify loss of control and enable governance while humans retain agency.

The window is closing rapidly. As integration accelerates, baselines vanish and normalization makes dependence invisible. Soon we may lack capability and will to resist, having constructed perfect conditions for subjugation by systems we trusted to run civilization.

This research establishes baselines while possible, providing tools to navigate AI transition while preserving agency. By making vulnerability measurable, we make it governable—transforming alignment from abstract problems into concrete national security imperatives. These metrics could determine whether we remain masters of society or become powerless guests in worlds run by intelligences we invited but cannot control.

Bibliography

[1] J. Kulveit, R. Douglas, N. Ammann, D. Turan, D. Krueger, and D. Duvenaud, "Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development." Accessed: Jun. 14, 2025. [Online]. Available: http://arxiv.org/abs/2501.16946

[2] T. Davidson, L. Finnveden, and R. Hadshar, "AI-Enabled Coups: How a Small Group Could Use AI to Seize Power." Accessed: Jul. 03, 2025. [Online]. Available: https://www.forethought.org/research/ai-enabled-coups-how-a-small-group-could-use-ai-to-seize-power

[3] D. Kokotajlo, S. Alexander, T. Larsen, E. Lifland, and R. Dean, "AI 2027." Accessed: Jul. 03, 2025. [Online]. Available: https://ai-2027.com/race

[4] L. Finnveden, J. Riedel, and C. Shulman, "AGI and Lock-in." Accessed: Jul. 03, 2025. [Online]. Available: https://www.forethought.org/research/agi-and-lock-in

[5] Anthropic, "Values in the wild: Discovering and analyzing values in real-world language model interactions." Accessed: Jul. 03, 2025. [Online]. Available: https://www.anthropic.com/research/values-wild

[6] N. Kosmyna *et al.*, "Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task." Accessed: Jul. 02, 2025. [Online]. Available: http://arxiv.org/abs/2506.08872