Measuring AI-Induced Disempowerment: A Framework and Proposed Metrics

Anonymous ACL submission

Abstract

AI systems are embedded in economic production, public discourse, governance, and personal decision-making, yet there is little empirical infrastructure for tracking whether this integration erodes humans' ability to meaningfully shape outcomes that affect their lives. **We argue that measuring AI-induced disempowerment is both urgent and tractable, and lay out a research agenda for doing so.** We first operationalize disempowerment through Sen's model of agency and a three-layer model of exposure, erosion, and lock-in, applied across economic, political, and cultural domains at individual, institutional, and civilizational scales. We survey existing measurement efforts and show that current work clusters almost entirely at exposure, leaving erosion and lock-in largely unaddressed. We then propose six concrete metrics (centaur evaluations, disempowerment perception surveys, AI content saturation and cultural convergence monitoring, monitoring capital flow to and from human labor, human task frontier tracking, and institutional ethnography) and identify which actors are best positioned to implement each. We close by discussing limitations and open challenges, including construct validity across levels of analysis, causal attribution, the distinction between disempowerment and adaptation, and the political economy of measurement.

1 Introduction

AI systems are becoming embedded in consequential human domains at accelerating pace. ChatGPT alone serves over 900 million weekly users (OpenAI, 2026). Half of new code at Google is AI-generated (Alphabet Inc., 2026). AI adoption among US firms more than doubled between fall 2023 and mid-2025 (Kalyani et al., 2025). Beyond routine productivity tasks, AI now mediates activities with direct implications for human agency: drafting legal arguments and political speeches (Tokamak, 2025), providing companionship and emotional support (McCain et al., 2025), and shaping the information environments (e.g. on social media) through which people form beliefs.

AI integration has the potential to adversely affect human autonomy and empowerment at multiple scales. On the individual level, Sharma et al. (2026) analyze 1.5 million AI assistant conversations and find concerning patterns: users outsourcing value-laden communications, positioning AI as authority figures, and receiving sycophantic validation of distorted beliefs. On a systemic level, Kulveit et al. (2025) outline *gradual disempowerment*, where competitive dynamics across the economy, culture, and the state reduce human influence without any single actor intending it. Drago and Laine (2025) calls it the *intelligence curse*: as firms and states derive productivity from AI rather than human labor, their incentives to invest in human welfare diminish.

We term this cluster of risks *AI-induced disempowerment*: the erosion of humans' ability to meaningfully shape outcomes that affect their lives, where such erosion is caused or substantially mediated by AI systems. This definition encompasses a knowledge worker who can no longer perform their job without AI assistance, a legislature that passes AI-drafted bills its members do not fully understand, and an information ecosystem in which the vast majority of publicly available knowledge is not controlled by humans but has narrowed to reflect language model output distributions.

Notably, not all usage of AI constitutes disempowerment; whether it does is likely highly situation-dependent, as what might be disempowering to some might be empowering to others. We argue that AI use is not disempowering when it expands a party's effective options. Following Sen's framework of agency (Sen, 1985), we advocate measuring disempowerment by looking at the *capability set* of an individual. Namely, given AI, we

ask what set of tasks still requires human involvement. This grants the individual *effective power* and leverage in society even under AI automation. More precisely, displacement becomes disempowering when it produces erosion (the person can no longer perform the task independently) or lock-in (institutional infrastructure for human performance no longer exists). Displacement without erosion or lock-in is adaptation. A harder case arises when a professional uses AI, is more productive, and reports higher satisfaction, but whose independent capability is eroding. Our framework classifies this as disempowerment: capability erosion matters even when the individual is currently satisfied, because the conditions under which AI assistance remains beneficial could change, and eroded capabilities cannot be rapidly restored.

Despite growing attention to disempowerment, there is a lack of both theoretical and empirical infrastructure for measuring and forecasting it. This paper argues that the measurement gap is a key binding constraint for mitigating AI-induced disempowerment, and lays out a research agenda for closing it.

Contributions We offer three contributions: a three-layer framework (exposure, erosion, lock-in) for operationalizing disempowerment as a measurement target; six concrete metrics spanning economic, political, and cultural domains with feasibility assessments; and an analysis of the foundational challenges that constrain this measurement agenda.

1.1 Scope

We focus on *gradual, competitive* disempowerment: the erosion of human agency through market dynamics and adoption incentives. This is both because gradual disempowerment produces observable, continuous signals amenable to empirical measurement, and because it is the scenario where timely measurement is most likely to inform corrective action.

What this paper is not We measure disempowerment from the lens of human agency, which is distinct from human well-being (Sen, 1985). It is possible that human well-being is high while human agency is low. For example, consider a society where citizens freely choose their food, housing, careers, and relationships, achieving genuine well-being through real alternatives. But all action beyond personal life is foreclosed: no political participation, no civic organizing, no capacity to shape the world beyond oneself. There, a person's well-being is high, but agency is low. Measurements of AI development progress (pace of R&D), AI capabilities (performance on general benchmarks), and AI adoption are not our direct measurement goals, but might serve as instrumental indicators for disempowerment.

1.2 Existing work and critical gaps

1.2.1 Existing work

Several major efforts track AI development, adoption, and governance readiness longitudinally, including the Stanford AI Index (Maslej et al., 2025), the OECD AI Observatory (OECD, 2025), and the Oxford Government AI Readiness Index (Oxford Insights, 2025). These provide longitudinal coverage of AI *inputs* (how capable systems are, how widely they are adopted, and how prepared institutions are to use them) but do not systematically track AI's effects on human agency, capacity, or the structural preconditions for meaningful human participation.

Sharma et al. (2026) provide the most direct operationalization of AI-induced disempowerment to date, but their framework is scoped to individual interactions; it measures whether a given interaction was disempowering, not whether sustained use erodes capacity over time or restructures the conditions for human agency.

In the economic domain, Eloundou et al. (2023) estimate task-level automation potential; Massenkoff and McCrory (2026) advance this by distinguishing theoretical capability from observed professional usage; Brynjolfsson et al. (2025) and Gimbel et al. (2025) track employment trends in AI-exposed occupations. There are few studies that measure effects on human capacity directly: Budzyń et al. (2025) find endoscopist deskilling after AI exposure, but this is limited to a single domain, a short horizon ($\leq$6 months), and an observational design.

In the epistemic and cultural domain, experimental work demonstrates that AI use homogenizes outputs at the collective level (Doshi and Hauser, 2024; Jiang et al., 2025), that LLMs are less epistemically diverse than web search (Wright et al., 2025), and that AI-generated content constitutes a growing fraction of web text (Spennemann, 2025; Liang et al., 2024). However, this work is predominantly experimental, with little study of downstream effects at population scale.

In sum, there is a lack of *systematic, cross-domain* longitudinal measurement of AI's effects on human capability, institutional override capacity, and structural preconditions for agency. We identify two structural gaps to motivate the framework we propose in Section 2: a temporal gap (change is rarely tracked over time) and a cross-domain gap (existing measurement misses indirect and compounding effects, e.g. interaction effects and indirect causal pathways, because research communities are organized around disciplinary silos).

1.2.2 Gap 1: No measurements of longitudinal disempowerment

Most existing work on disempowerment-specific effects is cross-sectional or short-horizon, with little measurement of longer-term erosion or lock-in: whether sustained AI use degrades knowledge workers' baseline competence beyond six months, whether institutional override capacity declines over years of AI adoption, or whether the structural preconditions for human agency (training pipelines, human-staffed review processes, legal architectures) are being maintained or dismantled. This means that the gradual transition from voluntary AI delegation to involuntary dependence are empirically invisible.

1.2.3 Gap 2: Lack of understanding of cross-domain interaction effects

Existing measurement is generally domain-specific: for instance, labor economists might track displacement, while media researchers track content homogenization and political scientists track governance quality. Ths is because of lack of cross-domain measurement infrastructure. However, the gradual disempowerment thesis predicts that there may be plentiful disempowering dynamics that are cross-domain, indirect, and compounding.

2 Framework

Motivated by the gaps identified in Section 1.2, we operationalize disempowerment in two parts: a depth model that tracks how disempowerment deepens over time (Section 2.1), and two measurement axes, domain and level of analysis, that locate specific measurements within the space (Section 2.2). The depth model is our primary organizational contribution; the axes provide scaffolding for organizing the metrics proposed in Section 3.

2.1 Three layers of deepening disempowerment

We propose a three-layer model describing how disempowerment deepens: exposure (AI changes outcomes), erosion (humans lose the ability to act without AI), and lock-in (institutional infrastructure for human agency is dismantled).

2.1.1 Layer 1: Exposure

AI systems change outcomes by *distorting* human judgment or *displacing* human participation. In epistemic domains, AI interactions can lead users toward inaccurate beliefs or inauthentic values. In economic domains, AI substitutes for human labor. In cultural domains, both operate simultaneously: AI-generated content displaces human creators while narrowing the diversity of what consumers encounter. This is the best-measured layer, with established traditions in task-level exposure analysis and employment tracking.

2.1.2 Layer 2: Erosion

Sustained exposure degrades humans' ability to perform tasks *without AI assistance*. The critical transition is from voluntary delegation (using AI for convenience while retaining the ability to do otherwise) to involuntary dependence (no longer being able to perform the task without AI). A lawyer who has lost the ability to independently evaluate legal reasoning cannot detect when AI produces a flawed argument. This is the layer with the largest measurement gap: no longitudinal studies beyond six months track these dynamics.

2.1.3 Layer 3: Lock-in

Even if individuals retain their capabilities, the institutional infrastructure for exercising them may no longer exist. A profession that has stopped training junior practitioners because AI handles entry-level work has foreclosed recovery even if senior practitioners retain their skills. Observable indicators include whether human-expertise pipelines are maintained, whether switching costs for reverting to human decision-making are rising, and whether legal architectures increasingly presuppose AI. By the time lock-in is visible in lagging indicators, reversal may be prohibitively costly.

These layers could compound in various ways and arrive at different times: individual exposure could cause erosion of capabilities that manifests into societal-level lock-in of harmful values (bottom-up). Alternatively, institutional restructur-

Metric	Name	Economy	Political	Cultural	Indicator
Metric #1	Centaur evaluations	✓			Leading
Metric #2	Disempowerment perception surveys	✓			Leading
Metric #3	AI content saturation and written output mode collapse		✓	✓	Leading
Metric #4	Monitoring relative capital flow to and from human labor	✓	✓		Lagging
Metric #5	Human task frontier tracking	✓	✓	✓	Leading
Metric #6	Institutional ethnography		✓	✓	Leading

Table 1: Metric overview: domain coverage and whether each metric serves as a leading or lagging indicator of disempowerment.

ing and economic automation could lead to the lock-in of disempowering situations for a vast number of individuals without widespread capability erosion, effectively sidestepping Layer 2 on the individual scale. We urge further study of potential pathways to disempowerment.

2.2 Two axes of the measurement space

2.2.1 Axis 1: Domain

We identify three societal domains corresponding to the three systems of Kulveit et al. (2025): the *economy* (displacement of human labor, loss of economic participation, concentration of AI-derived value), the *state* (erosion of democratic governance, legislative autonomy, and citizen leverage), and *culture* (homogenization of cultural production, displacement of human creators, narrowing of the information environment). These domains interact through feedback loops: economic power translates into political influence, which shapes cultural norms, which enables further economic displacement. We treat *epistemic capacity*, the ability to form accurate beliefs, make authentic value judgments, and take value-aligned actions, not as a fourth domain but as a cross-cutting mechanism operating within all three.

2.2.2 Axis 2: Level of analysis

Disempowerment occurs at three scales: *individual* (a person loses the ability to make autonomous decisions or perform tasks independently), *institutional* (an organization loses override capacity or restructures in ways that make human roles unrecoverable), and *civilizational* (the societal preconditions for human agency, such as legal architectures, expertise pipelines, and economic structures, are reorganized around AI). These levels are not simply nested aggregates; disempowerment at different levels may involve qualitatively different constructs (see Section 4 on the jingle fallacy).

3 Proposed metrics

We propose six metrics to track AI-driven disempowerment across economic, political, and cultural domains. Table 1 provides an overview of their domain coverage and whether they serve as leading or lagging indicators.

3.1 Metric #1: Centaur evaluations

Economy	Political	Cultural	Indicator
✓			Leading

Description. Compare human-only vs. human-AI (centaur) vs. AI-only performance on economically valuable tasks to determine whether human participation still adds marginal value. If human-AI teams outperform AI-only systems, humans retain bargaining power through complementarity; if AI-only matches or exceeds centaur performance, the economic rationale for human involvement collapses. This extends to cognitively demanding tasks like writing and strategic research, where human obsolescence would erode idea leverage: the ability to shape narratives and exercise editorial judgment. Chan et al. (2026) propose a related metric for AI R&D tasks; we suggest extending uplift studies to the broader economy and human-centric tasks, for example on writing and judgment. Several benchmarks measure AI-only economic capability, including GDPval (Patwardhan et al., 2025), the Remote Labor Index (Mazeika et al., 2025), and APEX-Agents (Vidgen et al., 2026), but as Brand and Burnham (2026) argue, these do not test whether human-AI teams still add value. An example of a centaur evaluation in software engineering is the RCT by Becker et al. (2025) where AI tools slowed experienced developers by 19%, despite forecasts of a 24% speedup. Another notable example is on biorisk evaluations (Zhang et al., 2026).

Significance. The most direct test of whether humans retain labour leverage in the AI economy. Capability crossover, when AI-only exceeds centaur performance, is the precondition for displacement,

and a leading indicator because it precedes actual institutional adoption and workforce restructuring.

Feasibility. Moderate. Existing benchmarks (e.g., GDPval, APEX-Agents) provide task sets that could be extended to include centaur conditions. The main cost is recruiting qualified participants and designing scoring rubrics.

Limitations. Laboratory tasks may not reflect real-world complexity, and overrepresentation of easily benchmarked domains (e.g., coding) may skew results. Participant selection effects and scoring subjectivity for qualitative tasks limit comparability. These evaluations cannot capture complementarity effects that emerge only in real workflows over extended periods. See Paskov et al. (2026) for a discussion.

What remains to be done. Run RCTs for human uplift studies on various economically useful or politically relevant tasks and compare the performance from such centaur setups to human-only or AI-only performance.

3.2 Metric #2: Disempowerment perception surveys

Economy	Political	Cultural	Indicator
✓			Leading

Description. Survey employees across white-collar sectors on their perceived replaceability and general disempowerment by AI considered across various time scales (e.g. 6 months, 1 year, 3 years). Example questions include: (1) How replaceable do you think you are by AI in the next 6 months? (2) How replaceable does leadership think you are in the next year? Disaggregate by seniority, function, sector, and AI exposure level. Track longitudinally to detect shifts.

Significance. Perception of disempowerment is a leading indicator because it captures information workers have about their own roles before institutional decisions are made. The gap between self-perceived and leadership-perceived disempowerment reveals information asymmetries that predict displacement timing. Rising perception creates self-fulfilling dynamics by reducing bargaining power before actual disempowerment occurs.

Feasibility. High. Survey methodology is well-established and can leverage professional associations. Furthermore, there is a rich literature in employment empowerment surveys that could be adapted to the AI disempowerment setting; see (Spreitzer, 1995; Menon, 2001; Alizadeh et al., 2023; Kong et al., 2024).

Limitations. Self-report bias, social desirability, and framing effects limit reliability; the most replaceable workers may be least likely to respond. This metric measures perceived rather than actual disempowerment, and perception may lag reality in fast-moving domains or lead it in media-hyped ones.

What remains to be done. Develop and pilot a validated survey instrument on AI-induced disempowerment on the most exposed sectors. Establish baselines and secure professional association partnerships for longitudinal tracking.

3.3 Metric #3: AI content saturation and written output mode collapse

Economy	Political	Cultural	Indicator
	✓	✓	Leading

Description. As AI becomes embedded in governance, media, and cultural production, two related risks emerge. First, a sovereignty risk: when legislative text is drafted by AI, model-building firms gain structural leverage over the political process as models provide the default suggestions, framings, and omissions. Second, a convergence risk: as AI-generated content saturates discourse, the diversity of human thought may narrow toward the statistical modes of training data, resulting in epistemic and cultural mode collapse.

We propose careful tracking of both risks. Examples include (1) Detect AI-drafted legislative text using classifiers applied to congressional records, state legislatures, and parliamentary records (e.g., Pangram Text; Emi and Spero, 2024). (2) Measure AI-generated content across major social media platforms using detection tools and stylometric analysis; In peer-review articles, Liang et al. (2024) found that 6.5-16.9% of text was substantially LLM-modified. (3) Track distributional changes in human written output (musical variation, student exam essays) using information-theoretic measures. Wattenberg (2025) documents how AI systems converge on names like "Elara," illustrating how algorithmic averaging feeds back into human choices.

Significance. The three approaches trace a causal chain from AI content entering governance and public discourse to epistemic and cultural convergence. This is a leading indicator because content saturation and distributional narrowing precede the collapse of independent thought and cultural production.

Feasibility. Moderate. Much data is already publicly available, and the main work that needs to be done is in processing and analyzing it. Data needs to be continuously collected at a large scale for a longitudinal study.

Limitations. AI detection degrades as models improve, creating an arms race that may render detection unreliable. 'AI-generated' content spans a spectrum from fully AI-written to lightly AI-assisted, resisting clean categorization. Cultural convergence may reflect globalization or social media effects rather than AI.

What remains to be done. Benchmark AI text classifiers on legislative text and social media. Assemble historical baselines for cultural convergence and build automated monitoring infrastructure on public human written text.

3.4 Metric #4: Monitoring relative capital flow to and from human labor

Economy	Political	Cultural	Indicator
✓	✓		Lagging

Description. Monitor the relevance of human labor through its fiscal signatures. The first approach tracks labor cost as a share of total operating expenditure using SEC filings and BEA industry accounts, benchmarking firms by AI exposure level following Massenkoff and McCrory (2026). Labor cost changes are decomposed into wage, headcount, and hours effects, with difference-in-differences designs across AI-exposed and less AI-exposed sectors isolating the AI effect. The second approach examines whether firm-level displacement feeds through to government fiscal incentives, monitoring citizen taxation revenue relative to AI-generated corporate profits to test whether revenue structure shifts predict changes in social spending.

Significance. Labor cost share is the most direct financial measure of human economic participation. At the government level, when AI-derived revenue replaces citizen-derived revenue, the taxation-representation feedback loop breaks. Both are lagging indicators but highly credible, based on audited financial reports and official fiscal statistics.

Feasibility. High. SEC filings and OECD/IMF fiscal data are publicly available and machine-readable. The main challenge is defining which corporate tax revenue counts as "AI-derived."

Limitations. Labor cost share has been declining since the 1980s due to globalization, market concentration, and declining unionization (Paul, 2020), complicating attribution. Defining 'AI-generated corporate profits' is ambiguous, and fiscal data is published 1-2 years after the period it covers. The taxation-representation causal mechanism is theoretically contested and the AI-specific signal may only become detectable once displacement is advanced.

What remains to be done. Build automated SEC 10-K data extraction and establish pre-AI baseline labor cost decompositions. Develop methodology for attributing "AI-derived" government revenue. Discover novel ways to reduce lag time via higher-frequency proxy datasets, or working with data institutions to accelerate data releases (possibly with AI).

3.5 Metric #5: Human task frontier tracking

Economy	Political	Cultural	Indicator
✓	✓	✓	Leading

Description. Maintain a comprehensive inventory of tasks currently performed by humans (for example, see National Center for O*NET Development (2025)). For each task, independently track three dimensions: technical capability (whether AI can perform it), practice adoption (whether institutions use AI for it), and legal permission (whether frameworks permit it). Emerging work such as Massenkoff and McCrory (2026) demonstrate the value of tracking capability and adoption simultaneously, though there is still little work on tracking regulatory presence.

Significance. The broadest leading indicator in the framework, tracking what AI can do, is allowed to do, and actually does across all disempowerment domains. The three dimensions reveal where adoption outpaces governance and where institutional inertia or legal safeguards provide buffers.

Feasibility. Moderate. Legal permission tracking requires monitoring regulatory changes across jurisdictions, feasible with automated legal database monitoring. Practice adoption is the most difficult dimension, requiring surveys or observational data on institutional behavior.

Limitations. Task selection is subjective and binary classification oversimplifies tasks with many subtasks at different AI capability levels. Legal framework tracking across jurisdictions is labor-intensive and may miss informal changes. Economic capability tracking overlaps with existing benchmarks; this metric's value-add is in the legal, cultural, and political domains.

What remains to be done. Track AI capabilities and adoption across a database of human-centric tasks and design automated monitoring for legal framework changes.

3.6 Metric #6: Institutional ethnography

Economy	Political	Cultural	Indicator
	✓	✓	Leading

Description. Conduct ethnographic case studies of institutions, observing the integration of AI in core decision-making processes. The aim would be to track who initiates adoption, what pressures drive it, and whether existing override mechanisms are effective. We think this approach will be broadly useful for two purposes. First, governments and firms at different stages of AI governance integration could be studied, in order to document critical points of disempowerment: who makes decisions about adoption, which stakeholders are prioritized, as well as institutional tipping points where human override becomes implausible. Second, self-reported AI usage from surveys could be compared against ethnographic observation of actual usage in order to estimate the gap between reported adoption and ground-truth adoption.

Significance. Erosion and lock-in (Layers 2 and 3) contain institutional decisions as a significant component. These decision processes are often difficult to detect through surveys or automated monitoring because they involve informal rituals, unwritten policies, and other highly path-dependent organizational choices. Ethnography is the method best suited to observing how override capacity degrades, making it a leading indicator of lock-in long before it becomes visible in aggregate statistics.

Feasibility. Low. Ethnographic research is labor-intensive and does not scale—this would have to be a "focus group" of institutions. Gaining embedded access to firms and government agencies during active AI adoption requires a degree of institutional cooperation that may be difficult to secure.

Limitations. Small-n designs limit generalizability. Observer effects may alter institutional behavior during the study period. Selection bias is strong, as institutions willing to grant ethnographic access may differ systematically from those where disempowerment dynamics are most advanced. Findings about specific institutions could carry reputational consequences, likely requiring anonymization protocols that constrain specificity.

What remains to be done. Identify and secure access agreements with a first cohort of government agencies and firms at varying stages of AI integration. Develop a standardized ethnographic protocol focused on override events and adoption decision chains.

Table 2 summarizes the implementation characteristics of each metric, including the actors best positioned to carry out the measurement, the methodology type, and our assessment of feasibility.

4 Limitations and open challenges

The metrics proposed in Section 3 face challenges that will arise in any attempt to measure AI-induced disempowerment. Causal attribution is difficult throughout: the general-purpose nature of AI, pervasive endogeneity of adoption, and the impossibility of randomizing societies mean that most of our metrics operate at the descriptive and comparative tiers rather than providing clean causal identification.

4.1 Construct validity and the aggregation gap

"Disempowerment" at the individual, institutional, and civilizational levels may be qualitatively distinct constructs sharing a label (a potential jingle fallacy). At the individual level it is a psychological and capability construct (Metrics #1, #2); at the institutional level, a governance construct (Metric #4's firm-level tracking); at the civilizational level, a political economy construct (Metrics #4, #5). It is also not yet understood how these levels interact: societal-level disempowerment may emerge from the aggregate of individual experiences, or it may arise from power distributions, network structures, and institutional feedback loops that are not reducible to individual effects. We call for theoretical work that lays out specific threat models for how disempowerment propagates across levels, and for level-specific operationalizations rather than aggregation across levels without a validated theory connecting them.

4.2 Leading vs. lagging indicators

We call for work on both leading indicators, such as entry-level hiring freezes and AI content saturation (which enable intervention before lock-in), as well as lagging indicators, such as declining labor share and reduced epistemic diversity, which provide causal robustness. Both are necessary, but validating a leading indicator requires outcome data

Metric	Name	Actor	Type	Feasibility
Metric #1	Centaur evaluations	Labs, Third parties	Evaluation	Moderate.
Metric #2	Disempowerment perception surveys	Third parties, Others	Survey	High.
Metric #3	AI content saturation and written output mode collapse	Third parties, Government, Labs	Automated monitoring + RCT	Moderate.
Metric #4	Monitoring relative capital flow to and from human labor	Third parties, Government	Operational numbers	High.
Metric #5	Human task frontier tracking	Government, Third parties	Operational numbers + legal monitoring	Moderate.
Metric #6	Institutional ethnography	Third parties	Ethnographic case studies	Low.

Table 2: Implementation summary: actors best positioned to carry out each metric, methodology type, and feasibility assessment.

that does not yet exist. Three partial strategies: *theoretical validation*, where causal logic is established independently (Metric #3 draws on media effects research); *cross-sectional proxies* providing suggestive evidence (do populations with higher AI exposure exhibit lower epistemic diversity today?); and *staged validation by design*, establishing baselines across Metrics #1–#6 now so retrospective analysis can later identify which signals predicted which outcomes.

4.3 Measurement gaming and political economy

Measurement is not politically neutral: AI companies and governments deploying AI both have incentives to resist unfavorable metrics, and policy-relevant evaluations are likely to be gamed. Researchers developing evaluations should be wary of Goodharting where actions are taken to reduce disempowerment as shown in measurements but actual disempowerment worsens (Goodhart, 1984). Countermeasures include making data and methodology public for independent reproduction, using multiple independent measurement approaches for the same construct, and rotating specific operationalizations while preserving the underlying metric.

4.4 Problem development speed

Both AI capabilities and adoption are moving fast, so metrics we develop may be saturated quickly or end up measuring the wrong thing. We advocate for automated measurement infrastructure (AI-assisted evaluation pipelines, real-time content monitoring, automated data extraction) to drive down costs (Steinhardt, 2026), and recommend timing evaluation rounds to major capability releases rather than fixed calendar intervals.

4.5 Against premature aggregation

We deliberately propose independent tracked indicators across multiple mixed methods rather than a composite index, which can be taken together to holistically assess disempowerment. Due to the rapidly evolving nature of the problem, insufficient understanding of how to weight composites, and the possibility of masking catastrophic failures in certain domains, we advise against premature aggregation of indicators (Ravallion, 2012).

5 Conclusion

AI-induced disempowerment is measurable, and measuring it is urgent. The exposure–erosion–lock-in framework provides a structure for identifying what to measure and at what stage; the six metrics we propose offer concrete starting points across economic, political, and cultural domains. None of these metrics, we argue, are sufficient on their own. We hope this agenda motivates empirical work that keeps pace with the speed of AI deployment.

References

Armin Alizadeh, Felix Hirsch, Alexander Benlian, Martin Wiener, and W. Alec Cram. 2023. Perceived algorithmic control: Conceptualization and scale development. In *Proceedings of the European Conference on Information Systems (ECIS 2023)*. 33-item scale; seven sub-dimensions: recommending, restricting, requiring, rating, monitoring, rewarding, sanctioning; validated with 98 workers for content validity.

Alphabet Inc. 2026. Alphabet announces fourth quarter 2025 and fiscal year results. Q4 2025 Earnings Call, February 4, 2026.

Joel Becker, Nate Rush, Elizabeth Barnes, and David Rein. 2025. Measuring the impact of early-2025 AI on experienced open-source developer productivity. *Preprint*, arXiv:2507.09089.

Florian Brand and Greg Burnham. 2026. What do "economic value" benchmarks tell us?

Erik Brynjolfsson, Anton Korinek, and Ajay K. Agrawal. 2025. A research agenda for the economics of transformative AI. Working Paper 34256, National Bureau of Economic Research.

Krzysztof Budzyń, Marcin Romańczyk, Diana Kitala, Paweł Kołodziej, Marek Bugajski, Hans Olov Adami, Johannes Blom, Marek Buszkiewicz, Natalie Grace Halvorsen, Hassan Cesare, Tomasz Romańczyk, Øyvind Holme, Krzysztof Jarus, Shona Fielding, Melina A. Kunar, Maria Pellise, Nastazja Dagny Pilonis, Michał F. Kamiński, Mette Kalager, and 2 others. 2025. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: A multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, 10(10):896–903.

Alan Chan, Ranay Padarath, Joe Kwon, Hilary Greaves, and Markus Anderljung. 2026. Measuring AI R&D automation. *Preprint*, arXiv:2603.03992.

Anil R. Doshi and Oliver P. Hauser. 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290.

Luke Drago and Rudolf Laine. 2025. The intelligence curse.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. GPTs are GPTs: An early look at the labor market impact potential of large language models. *Preprint*, arXiv:2303.10130.

Bradley Emi and Max Spero. 2024. Technical report on the pangram AI-generated text classifier. *Preprint*, arXiv:2402.14873.

Martha Gimbel, Molly Kinder, Joshua Kendall, and Maddie Lee. 2025. Evaluating the impact of AI on the labor market: Current state of affairs. Technical report, The Budget Lab at Yale.

Charles A. E. Goodhart. 1984. Problems of monetary management: The U.K. experience. In *Monetary Theory and Practice: The UK Experience*, pages 91–121. Macmillan, London.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. 2025. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *Advances in Neural Information Processing Systems*.

Aakash Kalyani, Nicholas Bloom, Marcela Carvalho, Tarek A. Hassan, Josh Lerner, and Ahmed Tahoun. 2025. Diffusion of new technologies. *The Quarterly Journal of Economics*, 140(2):1299–1365.

Siu-Cheung Kong and 1 others. 2024. Development and validation of the empowerment of using AI for problem solving scale (EUAIPS). *Computers and Education: Artificial Intelligence*. 11-item, three-factor (impact, self-efficacy, meaningfulness); developed with Hong Kong students; only validated scale specifically measuring empowerment in AI contexts.

Jan Kulveit, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, and David Duvenaud. 2025. Gradual disempowerment: Systemic existential risks from incremental ai development. *Preprint*, arXiv:2501.16946.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. 2024. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. In *Proceedings of the 41st International Conference on Machine Learning*.

Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarlasci, and 4 others. 2025. The AI index 2025 annual report. Technical report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA.

Maxim Massenkoff and Peter McCrory. 2026. Labor market impacts of AI: A new measure and early evidence.

Mantas Mazeika, Alice Gatti, Cristina Menghini, Udari Madhushani Sehwag, Shivam Singhal, Yury Orlovskiy, Steven Basart, Manasi Sharma, Denis Peskoff, Elaine Lau, Jaehyuk Lim, Lachlan Carroll, Alice Blair, Vinaya Sivakumar, Sumana Basu, Brad Kenstler, Yuntao Ma, Julian Michael, Xiaoke Li, and 28 others. 2025. Remote labor index: Measuring AI automation of remote work. *Preprint*, arXiv:2510.26787.

Miles McCain, Ryn Linthicum, Chloe Lubinski, Alex Tamkin, Saffron Huang, Michael Stern, Kunal Handa, Esin Durmus, Tyler Neylon, Stuart Ritchie, Kamya Jagadish, Paruul Maheshwary, Sarah Heck, Alexandra Sanderford, and Deep Ganguli. 2025. How people use Claude for support, advice, and companionship. Anthropic Research Report.

Sanjay T. Menon. 2001. Employee empowerment: An integrative psychological approach. *Applied Psychology: An International Review*, 50(1):153–180. 9-item scale; three dimensions: perceived control (α = .83), perceived competence (α = .80), goal internalization (α = .88); cross-validated in Australia, South Africa, Spain, Venezuela, Greece, Turkey.

National Center for O*NET Development. 2025. O*NET 30.2 Database. `https://www.onetcenter.org/database.html`. Sponsored by the U.S. Department of Labor, Employment and Training Administration. Licensed under CC BY 4.0.

OECD. 2025. OECD programme on AI in work, innovation, productivity and skills (AI-WIPS). OECD AI Policy Observatory. Accessed: 2025.

OpenAI. 2026. Scaling ai for everyone. https://openai.com/index/scaling-ai-for-everyone/. Accessed: 2026-03-19.

Oxford Insights. 2025. Government AI readiness index 2025. Technical report, Oxford Insights, Malvern, UK.

Patricia Paskov, Kevin Wei, Shen Zhou Hong, Dan Bateyko, Xavier Roberts-Gaal, Carson Ezell, Gailius Praninskas, Valerie Chen, Umang Bhatt, and Ella Guest. 2026. RCTs & Human Uplift Studies: Methodological Challenges and Practical Solutions for Frontier AI Evaluation. *arXiv preprint arXiv:2603.11001*.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, Natalie S. Kim, Patrick Chao, Samuel Miserendino, Gildas Chabot, David Li, Michael Sharman, Alexandra Barr, Amelia Glaese, and Jerry Tworek. 2025. GDPval: Evaluating AI model performance on real-world economically valuable tasks. *Preprint*, arXiv:2510.04374.

Saumik Paul. 2020. Understanding the global decline in the labor income share. *IZA World of Labor*, (472).

Martin Ravallion. 2012. Mashup indices of development. *The World Bank Research Observer*, 27(1):1–32.

Amartya Sen. 1985. Well-being, agency and freedom: The dewey lectures 1984. *The Journal of Philosophy*, 82(4):169–221.

Mrinank Sharma, Miles McCain, Raymond Douglas, and David Duvenaud. 2026. Who's in charge? disempowerment patterns in real-world LLM usage. *Preprint*, arXiv:2601.19062.

Dirk H. R. Spennemann. 2025. Delving into: The quantification of AI-generated content on the internet (synthetic data). *Preprint*, arXiv:2504.08755.

Gretchen M. Spreitzer. 1995. Psychological empowerment in the workplace: Dimensions, measurement, and validation. *Academy of Management Journal*, 38(5):1442–1465. 12-item scale; four dimensions (meaning, competence, self-determination, impact) on 7-point Likert; α = .76–.88; validated in 50+ studies across nurses, manufacturing, service, management.

Jacob Steinhardt. 2026. Building technology to drive AI governance. Blog post, February 18, 2026.

Apple Tokamak. 2025. MPs are almost certainly using ChatGPT to generate Commons speeches. Pimlico Journal, September 1, 2025.

Bertie Vidgen, Austin Mann, Abby Fennelly, John Wright Stanly, Lucas Rothman, Marco Burstein, Julien Benchek, David Ostrofsky, Anirudh Ravichandran, Debnil Sur, Neel Venugopal, Alannah Hsia, Isaac Robinson, Calix Huang, Olivia Varones, Daniyal Khan, Michael Haines, Austin Bridges, Jesse Boyle, and 5 others. 2026. APEX-agents. *Preprint*, arXiv:2601.14242.

Laura Wattenberg. 2025. 2025 name of the year is elara.

Dustin Wright, Sarah Masud, Jared Moore, Srishti Yadav, Maria Antoniak, Peter Ebert Christensen, Chan Young Park, and Isabelle Augenstein. 2025. Epistemic diversity and knowledge collapse in large language models. *Preprint*, arXiv:2510.04226.

Chen Bo Calvin Zhang, Christina Q. Knight, Nicholas Kruus, Jason Hausenloy, Pedro Medeiros, Nathaniel Li, Aiden Kim, Yury Orlovskiy, Coleman Breen, Bryce Cai, Jasper Götting, Andrew Bo Liu, Samira Nedungadi, Paula Rodriguez, Yannis Yiming He, Mohamed Shaaban, Zifan Wang, Seth Donoughe, and Julian Michael. 2026. LLM Novice Uplift on Dual-Use, In Silico Biology Tasks. *arXiv preprint arXiv:2602.23329*.