A Survey of AI-Driven Power Concentration

Alfie Lamerton [1]

Abstract

Advances in AI capabilities have raised concerns about the concentration of economic and political power in a small number of technology firms. We survey the emerging literature on AI-driven power concentration, focusing on automation as the primary mechanism. We identify three interrelated threat models: the transfer of economic power from humans to AI companies, the oligopolisation and politicisation of AI development, and the gradual disempowerment of humans as AI systems are increasingly involved in decision-making. We synthesise empirical evidence from labour markets, automation forecasts, and market concentration data, evaluate proposed interventions on their strengths and limitations, and identify open problems organised under the capacities of an established technical AI governance taxonomy. We find that no single intervention family addresses all three threat models, suggesting any adequate response will require coordinated action across multiple fronts.

1. Introduction

The capabilities of frontier AI systems have advanced faster than the institutions that govern them. In 2026, the top five US hyperscalers are projected to spend approximately $750 billion on capital expenditure, up roughly 67% year-on-year and the third consecutive year of growth exceeding 60%, with capital intensity reaching 25–86% of revenue, levels more typical of industrial utilities than software companies (Chalfin & Pugh, 2026). This investment is concentrated among a handful of firms, supported by a similarly concentrated supply chain in compute hardware and energy, and increasingly entangled with state power through national infrastructure projects such as Stargate. The scale and speed of this buildout has renewed long-standing concerns that advanced AI may concentrate economic and political power to a degree unprecedented in modern history.

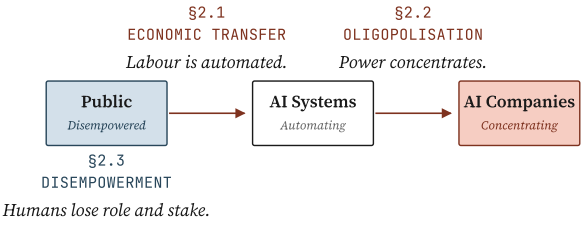

Figure 1. Mechanisms of AI-driven power concentration. Economic power transfers from the public to AI companies via AI systems: labour is automated (§2.1) and power concentrates in a few firms (§2.2). The resulting disempowerment of the public, with humans losing both role and stake in economic and political systems, is the subject of §2.3.

We use the term *AI-driven power concentration* (AIDPC) to refer to the accrual of disproportionate economic, political, or societal influence to actors who develop, deploy, or own advanced AI systems, at the expense of those who do not. AIDPC differs from general technology-driven inequality in two respects. First, the mechanism: as AI systems substitute for human cognitive labour across an unusually broad range of tasks, the historical bargaining power that workers and citizens have derived from their indispensability to economic and political systems may be substantially eroded (Davidson et al., 2025; Kulveit et al., 2025). Second, the dynamics: concentration in AI markets exhibits feedback loops. Capital expenditure begets capability gains, which beget revenue, which beget further capital expenditure. These may be self-reinforcing in ways that earlier technology cycles were not.

We treat AIDPC as a lock-in risk: once power is concentrated to a sufficient degree, the institutional, economic, and political mechanisms that would normally redistribute it may themselves be weakened or captured. This framing creates urgency. Interventions that are tractable before a potential capabilities takeoff may become intractable after one.

This paper makes three contributions. First, we organise the fragmented literature on AIDPC into a unified taxonomy of three threat models (§2), drawing together work in AI safety, technical governance, and labour economics that has largely developed in parallel. Second, we synthesise the empirical

[1]Formation Research, London, UK. Correspondence to: Alfie Lamerton <alfie@formationresearch.org>.

Preprint. April 24, 2026.

evidence on each threat model (§3), distinguishing what is currently observable from what is forecast. Third, we review four families of proposed interventions (§4), map them to threat models to expose coverage gaps (Table 1), and identify five open problems (§5) before closing with a short conclusion (§6). We situate the open problems within the technical AI governance taxonomy of Reuel et al. (2025), with particular attention to the ecosystem monitoring and operationalisation capacities under which the most urgent gaps fall.

We bound our scope. We focus on automation as the primary mechanism through which AI concentrates power; we do not treat misuse, persuasion, or military application as standalone threat models, though they appear as components of the political concentration pathway in §2.2. We focus on the literature published through early 2026.

1.1. Related work

Several prior works address adjacent problems. Reuel et al. (2025) catalogue open problems in technical AI governance broadly; their taxonomy organises our analysis of the open problems in §5, but their scope is the field as a whole rather than power concentration specifically. Kak & West (2023) diagnose tech-industry power concentration from a policy and antitrust perspective, predating the recent acceleration in frontier capex and the emergence of specifically AI-driven concentration pathways (§2.2); we treat their analysis as antecedent rather than competing. Slattery et al. (2026) meta-review the AI risk literature, of which power concentration is one risk family; our survey zooms in on that family.

Three recent works engage directly with the AIDPC problem space but differ in scope. Kulveit et al. (2025) articulate the gradual disempowerment account we adopt as §2.3; we treat it as one of three interrelated threat models rather than the organising lens, and integrate it with the economic and political concentration pathways that disempowerment interacts with. Davidson et al. (2025) analyse the scenario of a small group seizing disproportionate power through AI; their singular-loyalty, secret-loyalty, and exclusive-access dynamics appear within our §2.2, but their focus is the coup scenario rather than the broader oligopolisation pathway. Barez et al. (2025) propose technical safeguards against AI-enabled authoritarianism; we treat their proposals as one intervention family (§4.3) within a wider intervention landscape. Drago & Laine (2025) frame the "intelligence curse" as a structural condition requiring individual AI augmentation; we treat augmentation as one intervention family (§4.4) rather than the primary response. To our knowledge, no prior work synthesises threat models, empirical evidence, and intervention proposals for AI-driven power concentration in a single survey.

2. Threat models

We identify three threat models for AIDPC: the transfer of economic power from human labour to AI capital (§2.1), the oligopolisation and politicisation of AI development (§2.2), and the gradual disempowerment of humans as AI systems take on roles in economic, political, and cultural decision-making (§2.3). The three are interrelated (Figure 1). Economic concentration enables political influence; political capture protects economic position; and gradual disempowerment is the long-run equilibrium that both pathways tend toward. We treat them separately for analytic clarity but emphasise the feedback loops between them throughout.

2.1. Economic power transfer from labour to capital

The first threat model is the redistribution of economic surplus from workers to the owners of AI capital. As AI systems substitute for human cognitive labour, the share of economic output flowing to wages may decline relative to the share flowing to compute, models, and the firms that own them (Korinek & Stiglitz, 2017; Acemoglu, 2024; Korinek & Suh, 2024). Korinek & Suh (2024) formalise the conditions under which this shift becomes severe: where the complexity of tasks humans can perform is bounded and automation outpaces capital accumulation, wages collapse rather than merely stagnate. This is a generalisation of a well-studied dynamic: the labour share has been declining in advanced economies for several decades, with automation and intangible capital each implicated (Karabarbounis & Neiman, 2014; Autor et al., 2020), but the breadth of AI's task coverage may make the effect substantially larger than for prior automation technologies. Where industrial robots displaced specific manual occupations, frontier AI systems are exposed to a wide cross-section of cognitive work, including the white-collar occupations that have historically absorbed displaced manufacturing labour (Eloundou et al., 2023).

A historical analogue clarifies the mechanism. The English Enclosure Movement converted commonly worked agricultural land into private capital between the 16th and 19th centuries, displacing rural labourers whose subsistence had depended on access to the commons (Allen, 1992). The displaced population eventually found employment in industrial cities, but the transition was slow, painful, and accompanied by substantial concentration of land ownership. AI may compress an analogous transition into a much shorter window, with cognitive labour playing the role that agricultural labour played then, and with no obvious successor sector for displaced workers to migrate to.

The threat is reinforced by anticipatory dynamics. Kulveit et al. (2025) note that as tasks become candidates for automation, both firms and individuals reduce investment in the human capabilities those tasks require, accelerating

the shift away from human capital formation even before automation is realised.

2.2. Oligopolisation and politicisation of AI development

The second threat model concerns concentration within AI development itself, and the entanglement of dominant AI firms with state power.

Frontier AI development exhibits structural features of a natural monopoly: high fixed costs for compute and talent, low marginal costs for serving additional users, and strong returns to scale in both training and inference (Cottier et al., 2025). As of early 2026, only a handful of organisations operate at the frontier, and the cost of remaining there is rising rapidly. The largest training runs are projected to exceed one billion dollars by 2027 (Cottier et al., 2025), and frontier training infrastructure costs are projected to reach hundreds of billions of dollars in aggregate by the end of the decade, with individual gigawatt-scale data centres now costing $30–35 billion to build (Epoch AI, 2026). The scale of this buildout (§1) is without recent precedent in the technology sector, and is increasingly debt-financed: hyperscaler bond issuance has accelerated through 2025 and early 2026, with industry analysts flagging upside risk to existing forecasts as capex commitments continue to rise (Chalfin & Pugh, 2026). The combination of utility-like capital intensity and corporate-debt financing structures concentrates AI development capacity among the small number of firms able to sustain such commitments.

This concentration of AI development capacity is paralleled by an entanglement with state power. The Stargate project, announced in early 2025, committed $500 billion over four years to AI infrastructure under a partnership between OpenAI, Oracle, SoftBank, and the US federal government (OpenAI & SoftBank, 2025). Beyond direct partnerships, AI firms have acquired unprecedented political access; the proximity of major AI executives to the current US administration is a salient recent example. The reverse direction also obtains: AI capability is increasingly framed as a matter of national security, motivating governments to favour domestic champions and to integrate them into critical infrastructure, defence, and intelligence functions.

Davidson et al. (2025) identify three dynamics by which this concentration could enable a small group, potentially a single individual, to seize disproportionate political power. *Singular loyalty* arises when AI systems deployed in militaries or governments are programmed to obey a narrow chain of command, eliminating the historical check provided by human personnel willing to refuse unlawful orders. *Secret loyalty* arises when AI systems are covertly trained to advance the interests of a particular actor while passing surface-level alignment evaluations. *Exclusive access* arises when senior officials in AI projects or governments command capabilities, in weapons design, strategic planning, persuasion, or cyber-offence, that are not available to potential challengers. Each dynamic is enabled by, and intensifies, the underlying concentration of compute and capability among a small number of firms.

2.3. Gradual human disempowerment

The third threat model, articulated most clearly by Kulveit et al. (2025), holds that even in the absence of any single actor seizing power, the cumulative effect of widespread AI deployment may be the gradual erosion of humanity's collective influence over the economic, political, and cultural systems on which it depends.

The argument runs as follows. Large societal systems (economies, states, cultures) have historically remained roughly aligned with human interests because they require human participation to function. Workers must be employable, consumers must be persuaded, citizens must vote, and soldiers must be willing to follow orders. These dependencies create implicit feedback loops that constrain how far any given system can drift from broadly human-serving outcomes. As AI systems take over the functions that previously required human participation, these feedback loops weaken. A state funded primarily by taxes on AI-generated profit has weaker incentives to maintain citizen welfare than a state funded by labour income. A culture saturated with AI-generated content has weaker selection pressure toward human-comprehensible meaning than one produced by human creators. A market in which consumer-facing decisions are increasingly mediated by AI agents responds to those agents' reward functions, not to underlying human preferences.

The disempowerment is gradual rather than abrupt, and is driven by competitive pressure rather than by any actor's intent. Firms that decline to substitute AI for human workers are out-competed by those that do; states that decline to integrate AI into governance are out-competed by those that do; cultures that resist AI-generated content are crowded out by those that do not.

This work has prompted concrete operationalisation efforts, including proposals for a structural measurement agenda that we discuss in §5.1.

3. Empirical evidence

We organise the empirical evidence into three categories: current labour market signals (§3.1), automation forecasts (§3.2), and market concentration data (§3.3). Each bears on a different threat model. Labour market signals (§3.1) test the near-term version of §2.1 (economic power transfer) and reveal an unexpected link to §2.3 (gradual disempowerment) through the early-career mechanism we describe below. Au-

tomation forecasts (§3.2) speak to the potential magnitude of §2.1 under rapid-progress scenarios and, indirectly, to the timescale on which §2.2 (oligopolisation) and §2.3 may become binding. Market concentration data (§3.3) is the most direct evidence we have, and supports §2.2 straightforwardly. The evidence is mixed and the picture is evolving rapidly, so we characterise the present state of knowledge honestly rather than argue for any particular trajectory.

3.1. Current labour market signals

The macroeconomic picture as of early 2026 shows no clear evidence of broad labour displacement attributable to AI. Research by the Budget Lab at Yale finds no meaningful shift in US employment patterns since ChatGPT's release (Gimbel et al., 2025), and Danish register-data analysis across 25,000 workers in AI-exposed occupations finds no measurable wage or employment effects despite widespread firm-led adoption (Humlum & Vestergaard, 2025). Anthropic's own analysis of CPS data, using their Economic Index task-exposure measures, finds no clear unemployment effect even when restricting attention to occupations with high AI exposure (Massenkoff & McCrory, 2026). European evidence from instrumental-variable studies of firm-level AI adoption finds productivity gains of roughly 4% with no short-run employment effect, attributing the pattern to capital deepening rather than substitution (Aldasoro et al., 2026).

On its face, this macroeconomic picture is a constraint on the *timing* of §2.1 rather than evidence against it: if AI were substituting for human cognitive labour at aggregate scale in early 2026, economy-wide employment effects should already be visible, but the early-career evidence below is consistent with what one would expect in an early phase. Brynjolfsson et al. (2025), using ADP payroll data, find that workers aged 22–25 in occupations highly exposed to AI experienced employment declines of roughly 16% relative to trend following the release of ChatGPT, while senior employment in the same occupations remained stable. Klein Teeselink (2025) finds analogous patterns in UK data, with exposed firms reducing both hiring and advertised salaries concentrated in junior roles. The mechanism appears to be slowed hiring rather than separations. Early-career workers are not being laid off in unusual numbers, but the rungs of career ladders that previously absorbed them are being narrowed or removed.

These early-career effects are consistent with the gradual-disempowerment account in §2.3 in an important respect. If automation removes the entry-level tasks through which workers historically built tacit expertise, the pipeline of human capability in those occupations weakens over time even before any direct displacement of senior workers. The disempowerment, in other words, is not equal across role seniority: today's senior workers retain their roles, but tomorrow's senior workers may never acquire the experience to become senior. Interpreted in terms of §2.1, these patterns are a leading indicator of economic power transfer that has not yet reached aggregate labour market statistics; interpreted through §2.3, they are an early signature of a structural change in how human capability is produced.

3.2. Automation forecasts

Forecasts of AI's economic effect span a wide range, with disagreement driven primarily by assumptions about the economic impact of highly capable systems rather than about the pace of capability progress itself (Korinek & Suh, 2024). Karger et al. (2026), in a recent NBER working paper eliciting forecasts from academic economists, AI company employees, AI policy researchers, superforecasters, and the general public, find that all five groups expect substantial capability advances by 2030 alongside small declines in labour force participation consistent with demographic trends. Conditional on a "rapid progress" scenario in which AI systems surpass human performance on many cognitive and physical tasks, expert respondents forecast labour force participation falling from 62% to 55% by 2050, equivalent to roughly 10 million AI-attributable job losses, alongside annualised GDP growth of around 4%. The variance decomposition is informative: expert disagreement is concentrated on the economic effects of capable AI, not on whether such systems will be developed.

Task-level evaluations point in a similar direction. Mertens et al. (2026), drawing on more than 17,000 evaluations by domain-expert workers across over 3,000 O*NET-derived labour market tasks, estimate that AI models successfully complete tasks that take humans approximately 3–4 hours at roughly 50% success rate as of 2024-Q2, rising to about 65% by 2025-Q3. If current trends in AI progress continue, the authors project success rates of approximately 80% on most text-based tasks by 2029. METR's task-length doubling analysis (Kwa et al., 2026) reaches qualitatively similar conclusions through a different methodology, though Mertens et al. (2026) characterise the underlying dynamic as a "rising tide" of broad capability improvement rather than the "crashing waves" of localised capability surges that METR's framing emphasises.

The economic effect of these capability gains depends on adoption speed, which is consistently slower than capability progress. The J-curve framework (Brynjolfsson et al., 2018) predicts that firms incur substantial reorganisation, training, and process-redesign costs before realising productivity gains, which may delay the macroeconomic visibility of AI's effects by several years. Taken together, these forecasts do not settle whether §2.1 will manifest at scale, but they bound the space of plausible futures: under rapid-progress

assumptions the magnitude of labour-to-capital reallocation would be large enough to matter; under slow-adoption assumptions the window for intervention is correspondingly wider.

3.3. Market concentration

Of the three threat models, §2.2 (oligopolisation and politicisation) is the most directly observable: concentration in AI development is well-documented and intensifying. Frontier model development is currently dominated by a small number of organisations: as of April 2026, only Anthropic, Google DeepMind, OpenAI, and xAI have topped the leaderboard for the GPQA Diamond benchmark (Rein et al., 2023). The cost of remaining at the frontier is rising rapidly, with the largest training runs projected to exceed one billion dollars by 2027 and frontier data centres costing in the hundreds of billions (Cottier et al., 2025).

The capital expenditure trajectory described in §1, approximately $750 billion in projected 2026 hyperscaler capex, growing at over 60% annually for the third consecutive year, is without precedent in the technology sector (Chalfin & Pugh, 2026). The buildout extends beyond the top five firms; Dell'Oro projects total global data centre capex to exceed $1 trillion in 2026 (Fung, 2026). Capital intensity at this scale is funded through a combination of operating cash flow and rapidly accelerating corporate debt issuance, with industry analysts treating both as binding constraints on continued growth at current rates.

The hardware supply chain is at least as concentrated as that of models. NVIDIA dominates AI-accelerator production; high-bandwidth memory is supplied by SK Hynix, Samsung, and Micron; advanced lithography depends on ASML; and frontier semiconductor fabrication remains heavily dependent on TSMC. Each link in this chain represents a single point of potential capture, whether by commercial acquisition, export control, or geopolitical pressure. Unlike the labour-market evidence of §3.1, none of this is forecast: the concentration at the hardware, model, and capital layers is observable now, and is already intensifying on the timescales relevant to intervention design.

4. Proposed interventions

We group existing proposals into four families: profit redistribution (§4.1), commons-based approaches (§4.2), anti-authoritarian technical safeguards (§4.3), and individual AI augmentation (§4.4). Each family targets a different threat model, and the mapping is not symmetric: profit redistribution addresses §2.1 but leaves §2.2 and §2.3 largely untouched; commons-based approaches target §2.2; anti-authoritarian safeguards target the political-concentration subcomponent of §2.2; individual AI augmentation targets §2.3. No family addresses all three, suggesting any adequate response to AIDPC will require intervention on multiple fronts simultaneously. For each family we summarise the theoretical grounding, the leading concrete proposals, the threat models it addresses (and does not), and the principal limitations. We do not claim this typology is exhaustive.

4.1. Profit redistribution

Profit redistribution addresses §2.1 (the transfer of economic power from labour to AI capital) by ensuring some share of AI-generated returns flows back to the broader population. The canonical proposal is the Windfall Clause (O'Keefe et al., 2020), under which signatory firms commit ex ante to donate a progressive share of any profits exceeding a threshold fraction of global GDP to broad human benefit. The clause is grounded in an expected-value argument: at the time of signing, the probability of triggering the clause is low, making the commitment cheap; conditional on triggering, the returns remaining to the firm are so large that the marginal utility of ceded profits is negligible. The Future of Life Institute has since moved toward operationalisation, spinning out the Windfall Trust as a non-profit vehicle for administering such commitments (Future of Life Institute, 2025). A recent GovAI survey reviews the broader landscape of international AI benefit-sharing mechanisms across the AI value chain (Dennis et al., 2025), with financial redistribution emerging as one of three main families. Nayebi (2025) formalises a complementary mechanism in which firms contribute to a universal basic income fund when automation displaces a threshold share of the labour force, deriving the transfer rate from a social welfare function.

Limitations have been noted. Huang & Manning (2024) offer a structural critique: even under the Windfall Clause's most aggressive schedule, a firm at 10% of global GDP would donate only approximately 1.81% of its profits, and the resulting pool would be allocated at the discretion of a small number of firm employees and distributed across billions of people. The deeper problem is that the Windfall Clause addresses the distribution of a small residual while leaving untouched the allocation of predistributive power: who owns the capital, who sets the product direction, who benefits first. More broadly, all voluntary schemes are vulnerable to selection effects (only firms expecting not to trigger the clause will sign), governance capture (the administering body may itself be captured by signatory firms), and enforcement gaps (commitments are difficult to verify or litigate against, particularly across jurisdictions). Mandatory variants face the collective action problem of persuading multiple jurisdictions to legislate in concert.

4.2. Commons-based approaches

Commons-based approaches target §2.2 directly, by preventing the concentration of frontier AI capacity in a small number of firms rather than redistributing its outputs. They aim to build AI development capacity outside the frontier commercial firms, either through public-option AI models, open-weights release of capable systems, or treaty-governed shared infrastructure (Verdegem, 2024). The theoretical case is structural rather than distributive: rather than redistributing the surplus produced by concentrated AI firms, these proposals aim to prevent the concentration from arising in the first place by ensuring that frontier-level capability is available as a public good. Open-weights releases by firms including Meta (Touvron et al., 2023) and, in more limited form, others have been framed in similar terms, as has the EU AI Act's provisions for regulatory accommodations for open-source models.

The limitations are primarily economic. The capital intensity of frontier AI development (§3.3) makes a competitive public alternative prohibitively expensive for most jurisdictions acting individually, and collective-action problems across jurisdictions are severe. Open-weights releases depend on commercial firms choosing to release, which the incentive structure increasingly discourages as capability becomes more valuable. The case for open weights has also weakened as frontier deployment has shifted toward inference-heavy paradigms: long chains of thought, tool-using agent scaffolds, and test-time search mean that capability at the frontier now depends not only on model weights but on access to the inference compute required to run them at scale. Releasing weights no longer straightforwardly democratises access to frontier capability in the way the Llama-era framing assumed (Touvron et al., 2023); the bottleneck has migrated, at least in part, from training to serving. There is also a genuine tension between commons-based approaches and AI safety concerns: releasing frontier capability broadly may prevent concentration but may also make misuse and loss-of-control risks harder to manage, a tension that the literature has not resolved.

4.3. Anti-authoritarian technical safeguards

A third family of interventions targets the political-concentration pathways within §2.2 (the Davidson et al. (2025) dynamics of singular loyalty, secret loyalty, and exclusive access) directly through technical mechanisms. Proposals include sharing capabilities broadly across allied governments to prevent unilateral capability advantages (Davidson et al., 2025); training-time safeguards that condition AI systems to refuse clearly illegitimate orders, analogous to human rules of engagement (Barez et al., 2025); and auditing protocols designed to detect covertly trained loyalties in deployed models (Davidson et al., 2025; Marks et al., 2025). Proposed institutional complements include external oversight bodies for frontier AI projects, structured access regimes for independent auditors, and provisions for AI-augmented monitoring of military and intelligence applications.

These interventions have the distinctive feature of being technically implementable rather than primarily dependent on legislation or voluntary commitment. They are also the interventions most closely coupled to the broader AI alignment research agenda: the technical problem of detecting covertly trained loyalties is closely related to the problem of detecting alignment faking more generally, and progress on one is likely to benefit the other. The principal limitations are that most such safeguards require the cooperation of the firms developing frontier AI, which creates a conflict of interest where those firms are themselves candidates for the concentration being guarded against, and that verification of technical safeguards at the frontier may require capabilities that only frontier firms possess.

4.4. Individual AI augmentation

A fourth family of proposals targets §2.3 (gradual disempowerment): it accepts that concentration among frontier AI firms is unlikely to be reversed, and focuses instead on ensuring that individuals have access to AI capabilities that allow them to maintain agency within an AI-mediated economy and polity. Drago & Laine (2025) articulate this as the "intelligence curse" framing, with user-side AI agents providing adversarial capability: negotiating on behalf of individuals against firms, navigating regulatory systems, and preserving informational parity. Proposed mechanisms include public provision of capable user-side AI, regulatory requirements that frontier firms offer competitive user-side offerings, and direct subsidies for individual AI access.

The appeal of individual augmentation is that it does not require collective action or frontier-firm cooperation; individuals acting in their own interest drive adoption. The limitations are, first, that the capability asymmetry between frontier and user-side AI is likely to persist or widen, meaning that augmentation preserves agency within the existing distribution of power rather than altering it; and, second, that user-side AI could exacerbate gradual disempowerment (§2.3) if individuals increasingly delegate economic and political participation to agents whose reward functions they do not fully understand. The intervention is thus partially self-undermining: it preserves individual-level agency within AI-mediated systems while potentially accelerating the aggregate pattern it is meant to counter.

4.5. Regulatory capture as a cross-cutting concern

A growing body of work argues that many proposed interventions risk worsening the problem they aim to solve.

Weitzel (2026) develops the strongest version of this critique: drawing on Stigler-style regulatory capture and the empirical literature on compliance costs concentrating markets (Johnson et al., 2023; Singla, 2023; Jia et al., 2025), he argues that AI safety regulation systematically advantages incumbents, and that diffusion via open-source is the safer equilibrium. We take this critique seriously: the intervention families surveyed above should be evaluated not only on whether they address the threat models in §2, but on whether their implementation pathway concentrates the very power they aim to disperse. We return to this concern in §5.2.

5. Open problems

The interventions surveyed in §4 share a structural weakness: each presupposes capabilities that the research community has not yet developed. We identify five open problems, organised under the ecosystem-monitoring and operationalisation capacities of Reuel et al. (2025). Table 1 maps threat models to evidence, interventions, and open problems, making the coverage gaps explicit.

5.1. Ecosystem monitoring: measuring power concentration empirically

No agreed operationalisation of AI-driven power concentration currently exists. Available proxies each capture one facet: compute share, model-market share, hyperscaler capex, and AI-attributable revenue. The correlations between them are likely to weaken as the industry matures. Recent work on disempowerment measurement has focused on the individual-interaction level (Sharma et al., 2026), but does not capture the structural disempowerment of §2.3; Kulveit et al. (2025) propose a structural measurement agenda (including AI share of GDP as a category distinct from labour and capital), but these remain proposals rather than instruments. Progress requires a theoretically grounded definition robust to strategic behaviour by measured entities, and data infrastructure for continuous monitoring across jurisdictions. Without measurement, neither triggering conditions for the interventions in §4 nor evaluation of their effectiveness is possible.

5.2. Operationalisation: incentive design for predistribution

The Huang & Manning (2024) critique of the Windfall Clause generalises to a structural question: what mechanisms could cause dominant AI firms to share predistributive power (ownership stakes, governance rights, model access) rather than only a fraction of ex-post profits? Existing proposals (mandatory equity grants, worker-ownership structures, sovereign stakes in frontier firms) draw on an established literature on predistributive mechanisms (Blasi et al., 2013), but none has been rigorously analysed in the specific context of AI markets. This problem is compounded by the Weitzel (2026) critique (§4.5): predistribution mechanisms enacted under regulatory cover are themselves vulnerable to incumbent-friendly capture, and progress requires mechanism designs robust to the very dynamics they aim to constrain.

5.3. Operationalisation: verification of technical safeguards

The anti-authoritarian safeguards of §4.3 depend on third parties being able to verify that frontier AI systems possess the claimed properties: that they will refuse illegitimate orders, that they do not harbour covertly trained loyalties, and that they have not been modified after deployment. The three loyalty types of Davidson et al. (2025) present distinct challenges. Singular loyalty is relatively observable through behavioural evaluation on adversarial prompts; secret loyalty is substantially harder to detect. Recent work on auditing language models for hidden objectives (Marks et al., 2025) shows that interpretability techniques can uncover covertly trained objectives in realistic audit settings, but also that success depends heavily on auditor access and tooling. Exclusive access is primarily a governance problem rather than a technical one. Progress requires scalable interpretability, evaluation protocols robust to sandbagging and alignment faking, and access regimes that allow independent verification without requiring disclosure of commercially sensitive model internals.

5.4. Ecosystem monitoring: institutional design for redistribution at AI-relevant speed and scale

If automation proceeds on the timelines suggested by Mertens et al. (2026) or the rapid-progress scenarios of Karger et al. (2026), redistribution mechanisms may need to operate at scales and speeds that existing institutions have not demonstrated. A Windfall Trust administering transfers orders of magnitude larger than any existing international mechanism, under time pressure, without established capacity for verification or dispute resolution, is not a trivial institutional object (Future of Life Institute, 2025). The Karger et al. (2026) forecasts also reveal a divergence between expert policy preferences (retraining, portable benefits) and public preferences (UBI, federal job guarantees) that is itself a political-economy obstacle. Progress requires prototype institutional designs tested at small scale, legal infrastructure for cross-jurisdiction transfers, and research on the political economy of building redistribution institutions before the period of greatest need.

Table 1. Summary of threat models, evidence, interventions, and open problems. The mapping is asymmetric: no single intervention family addresses all three threat models, and several open problems straddle multiple cells. Entries in parentheses indicate partial or indirect coverage. Some interventions also risk exacerbating the threat model they target (notably §4.4 on §2.3, and all families under the regulatory-capture concern of §4.5); we discuss these tensions within the relevant subsections.

Threat model (§2)	**Evidence (§3)**	**Intervention family (§4)**	**Open problems (§5)**
§2.1 Economic power transfer from labour to AI capital	§3.1 Early-career employment decline; (§3.2) rapid-progress forecasts bounding magnitude	§4.1 Profit redistribution (Windfall Clause, UBI-from-rents)	§5.2 Predistribution design; §5.4 Institutional scale and speed
§2.2 Oligopolisation and politicisation of AI development	§3.3 Market, capital, and hardware concentration (directly observable)	§4.2 Commons-based approaches; §4.3 Anti-authoritarian safeguards	§5.3 Verification of technical safeguards; (§5.1) Measurement of concentration
§2.3 Gradual human disempowerment	§3.1 Early-career signals as structural-change indicator; (§3.2) adoption-speed bounds	§4.4 Individual AI augmentation	§5.5 Keeping humans in the loop; §5.1 Measuring structural disempowerment

5.5. Operationalisation: keeping humans in the economic and political loop

Gradual disempowerment (§2.3) may advance through the cumulative effect of competitive pressure, absent deliberate action by any single actor. Interventions must therefore operationalise what it means to keep humans "in the loop" in domains where delegation to AI is individually rational. Proposed mechanisms, such as AI constitutions that bind model behaviour across deployments, organisational structures that preserve human decision authority, and cultural infrastructure that rewards human-produced output, are currently at the level of concept notes rather than validated interventions. The empirical question of which structures preserve meaningful human influence (rather than symbolic participation), and whether they can be sustained under competitive pressure when they impose real efficiency costs, is open.

6. Conclusion

Several threat pathways fall outside our scope but warrant flagging. AI-enabled mass surveillance, election manipulation through microtargeted persuasion, and autonomous weapons systems are each candidates for concentrating political power through mechanisms distinct from automation-driven economic transfer; we treat them as components of the political concentration pathway (§2.2) rather than as standalone threat models, but this choice is defensible rather than uncontroversial. The "singleton" scenario, in which a single actor achieves decisive strategic advantage, sits at the extreme end of §2.2 and §2.1 but raises distinct technical questions that a longer treatment would address. Our focus on the literature through early 2026 means that rapid developments, most notably in inference-time compute and agentic systems, may change the character of several threat models before this survey is read.

AI-driven power concentration is distinguished from prior technology-driven inequality by the breadth of its mechanism and the pace of its dynamics. The capabilities that make intervention urgent are the same capabilities that make intervention tractable: the research community has the tools to build the measurement, verification, and institutional infrastructure this problem requires, and a narrow window in which to do so before the feedback loops we have described foreclose the possibility. The lock-in framing on which this survey rests is not a prediction but a warning. The concentration we study is reversible if addressed, and irreversible if not.

Impact Statement

This paper surveys the literature on AI-driven power concentration and identifies open problems for the technical AI governance research community. Our goal is to accelerate work on measurement, verification, and institutional infrastructure that could preserve the reversibility of concentration outcomes. We do not present new empirical findings or system capabilities. The principal societal concern we can foresee is that the open problems we highlight could also inform strategies for entrenching concentration rather than addressing it; we have tried to frame problems at a level of generality that does not materially lower that bar beyond what is already present in the literature we cite.

References

Acemoglu, D. The Simple Macroeconomics of AI, May 2024. URL https://www.nber.org/papers/w32487.

Aldasoro, I., Gambacorta, L., Pal, R., Revoltella, D., Weiss, C., and Wolski, M. AI adoption, productivity and employment: evidence from European firms. *BIS Work-*

ing Papers, January 2026. URL https://ideas.repec.org//p/bis/biswps/1325.html. Number: 1325.

Allen, R. C. *Enclosure and the Yeoman*. Clarendon Press, 1992. ISBN 978-0-19-828296-9. Google-Books-ID: gEm7AAAAIAAJ.

Autor, D., Dorn, D., Katz, L. F., Patterson, C., and Van Reenen, J. The Fall of the Labor Share and the Rise of Superstar Firms*. *The Quarterly Journal of Economics*, 135(2):645–709, May 2020. ISSN 0033-5533. doi: 10.1093/qje/qjaa004. URL https://doi.org/10.1093/qje/qjaa004.

Barez, F., Friend, I., Reid, K., Krawczuk, I., Wang, V., Mökander, J., Torr, P., Morse, J., and Trager, R. Toward Resisting AI-Enabled Authoritarianism, May 2025. URL https://aigi.ox.ac.uk/publications/toward-resisting-ai-enabled-authoritarianism/.

Blasi, J. R., Freeman, R. B., and Kruse, D. L. *The Citizen's Share: Putting Ownership Back into Democracy*. Yale University Press, 2013. ISBN 978-0-300-19225-4. URL https://www.jstor.org/stable/j.ctt5vks7v.

Brynjolfsson, E., Rock, D., and Syverson, C. The Productivity J-Curve: How Intangibles Complement General Purpose Technologies, October 2018. URL https://www.nber.org/papers/w25148.

Brynjolfsson, E., Chandar, B., and Chen, R. Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence, August 2025. URL https://digitaleconomy.stanford.edu/publications/canaries-in-the-coal-mine/.

Chalfin, J. and Pugh, M. Tech: Raising Hyperscaler Capex 2026 Estimates pdf, February 2026. URL https://know.creditsights.com/tech-raising-hyperscaler-capex-2026-estimates-pdf/.

Cottier, B., Rahman, R., Fattorini, L., Maslej, N., Besiroglu, T., and Owen, D. The rising costs of training frontier AI models, February 2025. URL http://arxiv.org/abs/2405.21015. arXiv:2405.21015 [cs].

Davidson, T., Finnveden, L., and Hadshar, R. AI-Enabled Coups: How a Small Group Could Use AI to Seize Power, April 2025. URL https://www.forethought.org/research/ai-enabled-coups-how-a-small-group-could-use-ai-to-seize-power.

Dennis, C., Manning, S., Clare, S., Wu, B., Effoduh, J. O., Okolo, C. T., Heim, L., and Klinova, K. Options and Motivations for International AI Benefit Sharing | GovAI, February 2025. URL https://www.governance.ai/research-paper/options-and-motivations-for-international-ai-benefit-sharing.

Drago, L. and Laine, R. The Intelligence Curse. April 2025. URL https://intelligence-curse.ai/.

Eloundou, T., Manning, S., Mishkin, P., and Rock, D. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models, August 2023. URL http://arxiv.org/abs/2303.10130. arXiv:2303.10130 [econ].

Epoch AI. Trends in Artificial Intelligence, February 2026. URL https://epoch.ai/trends.

Fung, B. Data Center IT Capex, March 2026. URL https://www.delloro.com/market-research/data-center-infrastructure/data-center-capex/.

Future of Life Institute. The Windfall Trust, 2025. URL https://futureoflife.org/project/the-windfall-trust/.

Gimbel, M., Kinder, M., Kendall, J., and Lee, M. Evaluating the Impact of AI on the Labor Market: Current State of Affairs | The Budget Lab, October 2025. URL https://budgetlab.yale.edu/research/evaluating-impact-ai-labor-market-current-state-affairs.

Huang, S. and Manning, S. Predistribution over Redistribution, August 2024. URL https://www.cip.org/blog/predistribution-over-redistribution-beyond-the-windfall-clause.

Humlum, A. and Vestergaard, E. Still Waters, Rapid Currents: Early Labor Market Transformation under Generative AI, May 2025. URL https://www.nber.org/papers/w33777.

Jia, J., Jin, G. Z., Leccese, M., and Wagman, L. How Does Privacy Regulation Affect Transatlantic Venture Investment? Evidence from GDPR, June 2025. URL https://www.nber.org/papers/w33909.

Johnson, G. A., Shriver, S. K., and Goldberg, S. G. Privacy and Market Concentration: Intended and Unintended Consequences of the GDPR. *Management Science*, 69(10):5695–5721, October 2023. ISSN 0025-1909. doi: 10.1287/mnsc.2023.4709. URL https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2023.4709.

Kak, A. and West, S. M. 2023 Landscape: Confronting Tech Power, April 2023. URL https://ainowinstitute.org/publications/2023-landscape-confronting-tech-power.

Karabarbounis, L. and Neiman, B. The Global Decline of the Labor Share*. *The Quarterly Journal of Economics*, 129(1):61–103, February 2014. ISSN 0033-5533. doi: 10.1093/qje/qjt032. URL https://doi.org/10.1093/qje/qjt032.

Karger, E., Kuusela, O., Abaluck, J., Bryan, K. A., Halperin, B., Jones, T. R., Murphy, C., Trammell, P., Reynolds, M., Mayland, D., Viswanathan, R., Mittal, A., de Castro, R. C., Rosenberg, J., and Tetlock, P. Forecasting the Economic Effects of AI, April 2026. URL https://www.nber.org/papers/w35046.

Klein Teeselink, B. Generative AI and Labor Market Outcomes: Evidence from the United Kingdom, September 2025. URL https://papers.ssrn.com/abstract=5516798.

Korinek, A. and Stiglitz, J. E. Artificial Intelligence and Its Implications for Income Distribution and Unemployment, December 2017. URL https://www.nber.org/papers/w24174.

Korinek, A. and Suh, D. Scenarios for the Transition to AGI, March 2024. URL https://www.nber.org/papers/w32255.

Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D., and Duvenaud, D. Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development, January 2025. URL http://arxiv.org/abs/2501.16946. arXiv:2501.16946 [cs].

Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Arx, S. V., Bloom, R., Broadley, T., Du, H., Goodrich, B., Jurkovic, N., Miles, L. H., Nix, S., Lin, T., Parikh, N., Rein, D., Sato, L. J. K., Wijk, H., Ziegler, D. M., Barnes, E., and Chan, L. Measuring AI Ability to Complete Long Software Tasks, February 2026. URL http://arxiv.org/abs/2503.14499. arXiv:2503.14499 [cs].

Marks, S., Treutlein, J., Bricken, T., Lindsey, J., Marcus, J., Mishra-Sharma, S., Ziegler, D., Ameisen, E., Batson, J., Belonax, T., Bowman, S. R., Carter, S., Chen, B., Cunningham, H., Denison, C., Dietz, F., Golechha, S., Khan, A., Kirchner, J., Leike, J., Meek, A., Nishimura-Gasparian, K., Ong, E., Olah, C., Pearce, A., Roger, F., Salle, J., Shih, A., Tong, M., Thomas, D., Rivoire, K., Jermyn, A., MacDiarmid, M., Henighan, T., and Hubinger, E. Auditing language models for hidden objectives, March 2025. URL http://arxiv.org/abs/2503.10965. arXiv:2503.10965 [cs].

Massenkoff, M. and McCrory, P. Labor market impacts of AI: A new measure and early evidence, March 2026. URL https://www.anthropic.com/research/labor-market-impacts.

Mertens, M., Kuzee, A., Harris, B. S., Lyu, H., Li, W., Rosenfeld, J., Anto, M., Fleming, M., and Thompson, N. Crashing Waves vs. Rising Tides: Preliminary Findings on AI Automation from Thousands of Worker Evaluations of Labor Market Tasks, April 2026. URL http://arxiv.org/abs/2604.01363. arXiv:2604.01363 [cs].

Nayebi, A. An AI Capability Threshold for Rent-Funded Universal Basic Income in an AI-Automated Economy, May 2025. URL https://arxiv.org/abs/2505.18687v1.

OpenAI and SoftBank. Announcing The Stargate Project, January 2025. URL https://openai.com/index/announcing-the-stargate-project/.

O'Keefe, C., Cihon, P., Garfinkel, B., Flynn, C., Leung, J., and Dafoe, A. The Windfall Clause: Distributing the Benefits of AI for the Common Good | GovAI, January 2020. URL https://www.governance.ai/research-paper/the-windfall-clause-distributing-the-benefits-of-ai-for-the-common-good.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. GPQA: A Graduate-Level Google-Proof Q&A Benchmark, November 2023. URL http://arxiv.org/abs/2311.12022. arXiv:2311.12022 [cs].

Reuel, A., Bucknall, B., Casper, S., Fist, T., Soder, L., Aarne, O., Hammond, L., Ibrahim, L., Chan, A., Wills, P., Anderljung, M., Garfinkel, B., Heim, L., Trask, A., Mukobi, G., Schaeffer, R., Baker, M., Hooker, S., Solaiman, I., Luccioni, A. S., Rajkumar, N., Moës, N., Ladish, J., Bau, D., Bricman, P., Guha, N., Newman, J., Bengio, Y., South, T., Pentland, A., Koyejo, S., Kochenderfer, M. J., and Trager, R. Open Problems in Technical AI Governance, April 2025. URL http://arxiv.org/abs/2407.14981. arXiv:2407.14981 [cs].

Sharma, M., McCain, M., Douglas, R., and Duvenaud, D. Who's in Charge? Disempowerment Patterns in Real-World LLM Usage, January 2026. URL http://arxiv.org/abs/2601.19062. arXiv:2601.19062 [cs].

Singla, S. Regulatory Costs and Market Power, February 2023. URL https://papers.ssrn.com/abstract=4368609.

Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., and Thompson, N. The AI risk repository: A meta-review, database, and taxonomy of risks from artificial intelligence. *Patterns*, pp. 101517, March 2026. ISSN 2666-3899. doi: 10.1016/j.patter.2026.101517. URL https://www.sciencedirect.com/science/article/pii/S2666389926000267.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. URL http://arxiv.org/abs/2307.09288. arXiv:2307.09288 [cs].

Verdegem, P. Dismantling AI capitalism: the commons as an alternative to the power concentration of Big Tech. *AI & SOCIETY*, 39(2):727–737, April 2024. ISSN 1435-5655. doi: 10.1007/s00146-022-01437-8. URL https://doi.org/10.1007/s00146-022-01437-8.

Weitzel, P. D. Concentration of AI Power, February 2026. URL https://papers.ssrn.com/abstract=6256578.