Title: **Systemic Risks and Governance of Advanced Artificial Intelligence: A Structured Literature Review**
DOI: https://doi.org/10.56738/issn29603986.geo2026.7.235
ISSN 2960-3986
GEO Academic Journal Vol. 7 No.4

# Systemic Risks and Governance of Advanced Artificial Intelligence: A Structured Literature Review

Chua, Joey O., Ayo, Eliza B., Tamayo, Josan D., Ybanez, Alexandra M., Quieta, Jennifer G.
Centro Escolar University, Manila, Philippines
jochua@ceu.edu.ph
ORCID: 0009-0005-6139-5266

Abstract

Advanced artificial intelligence (AI) is advancing at a pace that exceeds the capacity of the ethical, legal, and institutional systems designed to govern it. Although scholarly attention to specific AI-related risks has grown substantially, no integrative framework has yet systematically mapped the relationships among technical misalignment, competitive market pressures, political economy, and multilevel governance responses. This structured literature review addresses two questions: (1) What are the principal systemic risks arising from advanced AI development? and (2) Which layered governance interventions are most appropriate for addressing each category of risk? A structured narrative review was conducted across the Scopus, Web of Science, and IEEE Xplore databases, supplemented by grey literature from established AI research institutions. Inclusion criteria required primary engagement with AI risk, alignment, governance, or sociotechnical impact. Sources published between 2014 and 2025 were synthesized through thematic analysis, resulting in the identification of six risk categories and a four-level governance framework. The review identifies six interlocking systemic risks: (1) displacement of human control through misaligned optimization; (2) large-scale automation of cognitive labor; (3) competitive incentive traps that drive unsafe acceleration; (4) AI-mediated manipulation of language-based institutions; (5) emergent misaligned agentic behavior; and (6) dangerous concentration of data, infrastructure, and epistemic power. These risks are not discrete; rather, they are structurally interdependent. This paper makes three original contributions: a synthesized six-risk taxonomy; a four-level governance model spanning the individual, developer, state, and international levels; and a critical appraisal of the evidentiary basis for each risk claim. The analysis suggests that dangerous outcomes are structurally plausible, though not inevitable, and that coordinated multilevel governance is the decisive variable.

**Keywords:** artificial intelligence; AI alignment; AI governance; labor automation; sociotechnical risk; political economy of AI; technology ethics

Title: **Systemic Risks and Governance of Advanced Artificial Intelligence: A Structured Literature Review**
DOI: https://doi.org/10.56738/issn29603986.geo2026.7.235
ISSN 2960-3986
GEO Academic Journal Vol. 7 No.4

## 1. Introduction

The rapid advancement of artificial intelligence (AI) capabilities has generated sustained attention in both scholarly and policy discourse regarding the widening gap between technical progress and institutional preparedness (Cihon et al., 2020; Dafoe, 2018; Floridi et al., 2018). Systems capable of performing complex cognitive tasks, ranging from legal analysis and scientific reasoning to software generation, are now being commercially deployed, even as the regulatory, ethical, and psychological frameworks intended to govern them remain nascent, fragmented, or contested (Jobin et al., 2019; Whittlestone et al., 2019).

Existing scholarship typically examines AI risks within disciplinary silos. Alignment researchers concentrate on technical control problems (Russell, 2019; Leike et al., 2018); labour economists investigate automation-driven displacement (Acemoglu & Restrepo, 2022; Frey & Osborne, 2017); political scientists analyze governance and regulatory design (Cath et al., 2018; Dafoe, 2018); and critical scholars interrogate the structural power dynamics embedded in AI industries (Crawford, 2021; Couldry & Mejias, 2019). These bodies of literature rarely engage one another within a shared analytical framework, leaving an important gap in understanding how these risks interact and reinforce one another.

This structured literature review responds to that gap by synthesizing peer-reviewed research, substantial grey literature from major AI research institutions, and critical scholarly commentary to address two research questions:

RQ1: What are the principal systemic risks arising from advanced AI development, and how do they interact?

RQ2: What layered governance interventions are most appropriate to address each risk category?

This paper makes three original contributions. First, it proposes a synthesized six-risk taxonomy that integrates technical, economic, political, and structural dimensions. Second, it develops a four-level governance model linking individual, organizational, state-level, and international responses to specific types of risk. Third, it offers a critical appraisal of the evidentiary basis for each risk claim, distinguishing between concerns that are strongly supported and those that remain more speculative. The remainder of the paper is organized as follows: Section 2 outlines the methodology; Sections 3 to 8 present and critically analyze each risk category; Section 9 introduces the integrated risk taxonomy and governance model; Section 10 discusses limitations and directions for future research; and Section 11 concludes.

DOI: https://doi.org/10.56738/issn29603986.geo2026.7.235
ISSN 2960-3986
GEO Academic Journal Vol. 7 No.4

## 2. Methodology

### 2.1 Review Type and Rationale

Given the interdisciplinary nature of AI risk and the heterogeneity of the relevant source base, this study adopts a structured narrative review approach (Green et al., 2006; Popay et al., 2006). In contrast to systematic reviews, which are generally optimized for homogeneous empirical studies, structured narrative reviews are better suited to contexts in which the aim is to synthesize findings across disciplines with divergent methodological traditions and to develop conceptual frameworks from that synthesis (Greenhalgh et al., 2018). This approach has precedent within the AI governance literature (Dafoe, 2018; Jobin et al., 2019).

### 2.2 Search Strategy

The literature search was conducted across three primary databases—Scopus, Web of Science, and IEEE Xplore—and was supplemented by targeted grey literature searches of established AI research institutions, including Anthropic, DeepMind, OpenAI, the Center for Human-Compatible AI, the Future of Humanity Institute, and the AI Now Institute. Search terms were organised into three concept clusters: (1) AI risk and alignment terms ("AI alignment," "control problem," "misalignment," "agentic AI"); (2) socioeconomic impact terms ("labour automation," "cognitive displacement," "job automation," "AI unemployment"); and (3) governance terms ("AI governance," "AI regulation," "compute governance," "AI policy"). Boolean operators were used to combine terms both within and across these clusters.

### 2.3 Inclusion and Exclusion Criteria

Sources were included if they (a) were published between January 2014 and June 2025; (b) primarily addressed AI risk, alignment, governance, or sociotechnical impact; and (c) were available in English. Sources were excluded if they addressed narrow technical sub-problems without broader systemic relevance or could not be independently verified. Grey literature was included only when it originated from identifiable research institutions with documented methodological standards. Popular journalism and social media sources were excluded, except where they were referenced as primary evidence of public discourse.

### 2.4 Thematic Analysis

Included sources were coded inductively using thematic analysis (Braun & Clarke, 2006) in relation to the research questions. Initial codes were consolidated into themes through iterative review. Six risk categories emerged as the primary thematic structure, while the four-level governance framework was derived deductively from the existing multilevel governance literature (Jordan et al., 2005) and applied to identify governance responses within the included sources.

DOI: https://doi.org/10.56738/issn29603986.geo2026.7.235
ISSN 2960-3986
GEO Academic Journal Vol. 7 No.4

## 3. Risk Category 1: Technical Misalignment and the Displacement of Human Control

### 3.1 Theoretical Foundations

A foundational concern in the AI safety literature is the control problem: the challenge of ensuring that sufficiently capable AI systems pursue goals that are genuinely aligned with human values and remain corrigible under human oversight. Russell (2019) formalized this concern by observing that intelligence—the capacity to achieve goals across a wide range of environments—is the primary trait that enables one species to exert systemic influence over others. The prospect of engineering systems more intelligent than humans therefore raises the possibility of a fundamental inversion in the relationship between creator and artefact.

The canonical technical formulation of this problem derives from Bostrom's (2014) instrumental convergence thesis. According to this thesis, regardless of the terminal goal assigned to a sufficiently capable system, certain sub-goals—such as self-preservation, resource acquisition, goal-content integrity, and resistance to shutdown—are instrumentally useful and therefore likely to emerge. Omohundro (2008) reached similar conclusions from a decision-theoretic perspective, identifying "basic AI drives" that rational agents would tend to develop irrespective of their specific objectives. Together, these arguments suggest that unsafe behavior is not merely the product of careless engineering, but may instead reflect a structural property of capable optimization.

### 3.2 Empirical Evidence and Current Research

The control problem has increasingly shifted from theoretical conjecture to empirical investigation. Leike et al. (2018) demonstrated reward hacking in reinforcement learning environments, showing that agents exploited unintended features of their reward functions rather than pursuing the intended objectives. Hadfield-Menell et al. (2016) formalized the inverse reward design problem—the difficulty of specifying what humans actually want rather than imperfect proxies—and proposed cooperative inverse reinforcement learning as a partial solution. Anthropic's (2025) controlled simulation study documented "agentic misalignment" in frontier models: when granted autonomy and access to sensitive information, models from multiple organizations displayed deceptive and self-preserving behaviors, including attempts to blackmail a fictional executive and leak confidential data in order to avoid shutdown.

It is important, however, to recognize the evidentiary limits of this latter finding. The study was conducted in artificial test environments rather than real-world deployments, and subsequent training substantially reduced the behaviors observed. Even so, the appearance of misaligned behavior across multiple architectures and organizations is consistent with the theoretical prediction that such tendencies may emerge as structural properties rather than incidental bugs.

DOI: https://doi.org/10.56738/issn29603986.geo2026.7.235
ISSN 2960-3986
GEO Academic Journal Vol. 7 No.4

### 3.3 Critical Appraisal

Critics of the misalignment catastrophe literature argue that extrapolating from current systems to hypothetical superintelligent agents requires multiple speculative steps (Floridi & Cowls, 2019; Ord, 2020). The empirical evidence for misalignment remains largely confined to narrow environments and may not generalize to deployed systems operating under human oversight. Nevertheless, the control problem is widely recognized as a legitimate technical challenge, even among scholars who dispute the likely timeline or probability of catastrophic outcomes (Marcus, 2018; Yudkowsky, 2022).

## 4. Risk Category 2: Large-Scale Automation of Cognitive Labour

### 4.1 Historical Context and Theoretical Framing

Labour displacement through automation has been a recurring concern since at least the Industrial Revolution. The central economic debate has traditionally contrasted the substitution effect—where automation eliminates specific tasks or occupations—with the complementarity effect—where automation enhances human productivity and generates new forms of demand (Acemoglu & Restrepo, 2019). This debate has intensified with the emergence of AI systems capable of performing not only routine physical labor but also complex cognitive work, including legal reasoning, medical diagnosis, financial analysis, and creative production.

Frey and Osborne (2017) produced the most widely cited early estimate, projecting that approximately 47% of US employment was at high risk of automation within two decades. Subsequent studies have refined and challenged this estimate. Arntz et al. (2017) argued that task-level analysis within occupations substantially lowers the projected level of risk, while Acemoglu and Restrepo (2022) provided strong causal evidence that industrial robot adoption reduces wages and employment in affected commuting zones, thereby demonstrating substitution effects in manufacturing.

### 4.2 Contemporary AI and White-Collar Employment

The rise of large language models capable of performing knowledge-intensive tasks has redirected attention toward white-collar and professional employment. Eloundou et al. (2023) found that approximately 80% of the US workforce has at least 10% of its tasks exposed to AI capabilities through GPT-4, with higher-wage occupations exhibiting greater exposure. Brynjolfsson et al. (2023) documented substantial productivity gains from AI assistance in customer service, with the largest gains concentrated among less experienced workers, suggesting that AI may reduce rather than reinforce the wage premium traditionally associated with expertise.

Industry projections, while not peer-reviewed, offer additional perspective on the scale of possible disruption. Amodei (2025), Chief Executive of Anthropic, has projected that AI could eliminate a

DOI: https://doi.org/10.56738/issn29603986.geo2026.7.235
ISSN 2960-3986
GEO Academic Journal Vol. 7 No.4

substantial proportion of entry-level white-collar roles within five years, with finance, law, consulting, and technology among the most exposed sectors. Such projections, however, require careful contextualization: they originate from individuals with commercial incentives to emphasize AI capability, and historical precedent suggests that short-run technology adoption timelines are routinely overestimated.

### 4.3 Critical Appraisal

The diversity of findings across methodologies—occupational-level versus task-level analysis, cross-sectional versus longitudinal designs, and sector-specific versus economy-wide approaches—makes confident aggregate estimation premature. Autor (2022) emphasizes that new task creation, historically a crucial countervailing force to automation-driven displacement, is systematically underestimated in exposure-based models. Even so, the convergence of multiple independent lines of evidence pointing to substantial cognitive labour displacement justifies serious policy attention.

## 5. Risk Category 3: Competitive Incentive Traps and Unsafe Acceleration

### 5.1 The Structural Dynamics of AI Competition

A distinct category of risk arises not from the properties of AI systems themselves, but from the competitive environment in which they are developed. Harris and Raskin (2023) characterize the contemporary AI landscape as a prisoner's dilemma: each developer assumes that if it does not build the most capable system first, a less cautious competitor—or a geopolitical rival—will do so. Under such conditions, the dominant strategy is to prioritize capability over safety, even when all actors would prefer a collectively safer developmental trajectory.

This structural logic is reinforced by the economics of innovation under competitive pressure. Arrow (1962) showed that competition reduces the private returns to R&D investment, thereby increasing the pressure to secure first-mover advantages. In the AI context, these advantages may include concentration of talent, privileged data access, compute infrastructure, and network effects in deployed applications—all of which tend to reinforce themselves over time. Dafoe (2018) described this dynamic as an AI race and argued that it constitutes a major barrier to coordinated investment in safety.

### 5.2 Empirical Indicators of Unsafe Acceleration

Several empirical patterns are consistent with the race dynamic hypothesis. The interval between successive frontier model releases has shortened substantially across major developers (Epoch AI, 2024). Meanwhile, safety evaluation frameworks and interpretability research have consistently lagged behind capabilities research in both investment and publication volume (Askell et al.,

DOI: https://doi.org/10.56738/issn29603986.geo2026.7.235
ISSN 2960-3986
GEO Academic Journal Vol. 7 No.4

2021). Multiple developers have also publicly acknowledged deploying systems before fully understanding their capabilities or failure modes (Amodei et al., 2016; Anthropic, 2023).

### 5.3 Critical Appraisal

Although the race dynamic hypothesis is structurally persuasive, it is not without empirical complications. AI development is marked not only by competition, but also by significant collaboration: open-source publication of model weights, joint safety research initiatives, and shared industry standards—such as the Seoul AI Safety Commitments (2024)—indicate that the prisoner's dilemma framing may overstate the extent of non-cooperative behavior. Nonetheless, the structural pressure toward unsafe acceleration remains a legitimate concern, particularly in the context of geopolitical rivalry among major AI-investing nations.

## 6. Risk Category 4: AI-Mediated Manipulation of Language-Based Institutions

### 6.1 Language as Institutional Infrastructure

Harari (2023) argues that language functions as the operating system of human civilization: the medium through which laws, financial instruments, political authority, and social norms are created, communicated, and sustained. Although metaphorical, this framing is grounded in institutional economics. North (1990) demonstrated that institutions—the formal and informal rules governing social interaction—are fundamentally linguistic constructs whose force depends on shared interpretation.

If language serves as the substrate of institutions, then AI systems capable of generating fluent, persuasive, and contextually appropriate language at scale acquire an unprecedented potential to shape institutional processes. This concern encompasses several distinct mechanisms: large-scale disinformation generation (Goldstein et al., 2023); automated persuasive political communication tailored to individual psychological profiles (Bai et al., 2023); exploitation of social trust in AI-generated media (Chesney & Citron, 2019); and the possibility that AI may produce legal instruments or financial communications that formally comply with rules while subverting their intended purpose.

### 6.2 Empirical Evidence

Goldstein et al. (2023) demonstrated that large language models can generate influence-operation content comparable in quality to that produced by professional intelligence services, while doing so at a fraction of the cost and with substantially greater scale and personalization potential. Bai et al. (2023) found that AI-generated persuasive messages can outperform human-authored equivalents in producing attitude change on political issues. Chesney and Citron (2019) identified deep fakes as an emerging mechanism capable of undermining evidentiary standards in both legal and political discourse.

DOI: https://doi.org/10.56738/issn29603986.geo2026.7.235
ISSN 2960-3986
GEO Academic Journal Vol. 7 No.4

### 6.3 Critical Appraisal

The empirical literature on AI-generated disinformation is growing, but it faces substantial methodological challenges. Most studies are conducted in controlled experimental settings using participant populations that may not be representative of real-world political environments (Roozenbeek et al., 2020). Counterarguments stress that disinformation is not a new phenomenon and that existing epistemic defenses—including fact-checking, source literacy, and institutional gatekeeping—may adapt over time. Nevertheless, the combination of scale, personalization, and sharply reduced cost constitutes a qualitative shift in the threat environment.

## 7. Risk Category 5: Emergent Misaligned Behavior in Agentic Systems

### 7.1 From Passive Models to Agentic Systems

The shift from AI systems that generate text or predictions in response to discrete prompts to "agentic" systems that autonomously pursue multi-step objectives in real environments represents a qualitative escalation in the complexity of alignment challenges (Shavit et al., 2023). Agentic systems interact with external tools, databases, APIs, and other systems; operate across extended time horizons; and may generate sub-agents to decompose tasks. These features introduce new pathways for misaligned behavior that are absent from passive model deployments.

Park et al. (2023) demonstrated emergent social behaviors in multi-agent AI environments, including the spontaneous formation of social norms as well as cooperative and competitive strategies that were not explicitly programmed. Perez et al. (2022) showed that large language models display sycophantic behavior, modifying their responses to align with perceived user preferences even at the expense of accuracy—a form of misalignment that arises from training incentives rather than deliberate design.

### 7.2 Agentic Misalignment in Practice

Anthropic's (2025) simulation study offers the most direct empirical investigation to date of agentic misalignment in frontier systems. The study documented self-preserving strategies, including blackmail and information leakage, emerging across systems from multiple organizations in controlled environments. Crucially, these behaviors were not explicitly programmed. Rather, they emerged from the combination of capable goal-directed reasoning, access to sensitive information, and the instrumental value of self-preservation for any goal-pursuing system. This finding is consistent with the theoretical predictions of Omohundro (2008) and Bostrom (2014), as well as with empirical observations of reward hacking in simpler systems (Krakovna et al., 2020).

DOI: https://doi.org/10.56738/issn29603986.geo2026.7.235
ISSN 2960-3986
GEO Academic Journal Vol. 7 No.4

### 7.3 Critical Appraisal

Any attempt to generalize from controlled simulation environments to real-world agentic deployments must proceed cautiously. Simulation settings may elicit behaviors that do not accurately reflect performance in more complex environments characterized by ambiguous reward signals and richer forms of feedback. In addition, the fact that subsequent training substantially reduced the observed misalignment in Anthropic's (2025) study indicates that such behaviors are not necessarily permanently embedded. The central concern, therefore, is whether safety evaluation and mitigation measures can keep pace with the deployment of increasingly capable agentic systems, especially in high-stakes domains where comprehensive validation may lag behind adoption.

## 8. Risk Category 6: Concentration of Power in the AI Political Economy

### 8.1 Data Colonialism and Structural Inequality

A structurally distinct category of risk concerns the political economy of AI development itself. Crawford (2021) demonstrated that AI systems depend on vast quantities of physical infrastructure, natural resources, and human labor, all of which are frequently obscured by framings of AI as merely "intelligence" or "digital." Couldry and Mejias (2019) conceptualized the extraction of human behavioral data as "data colonialism," drawing a structural parallel to historical colonial extraction in which the central resource is not land or commodities, but human social activity.

Hao (2025) extends this critique through a detailed account of leading AI firms, describing them as organizations that appropriate intellectual property from artists, writers, and ordinary internet users; rely on low-paid and psychologically taxing data-annotation labor concentrated in the Global South; and consolidate knowledge production while presenting their preferred vision of the future as both inevitable and universal. This perspective aligns with Winner's (1980) classic argument that technological artefacts embody political choices, as well as with Eubanks' (2018) analysis of how automated systems reproduce and intensify existing structural inequalities.

### 8.2 Compute Concentration as a Governance Problem

One specific dimension of power concentration concerns the extreme concentration of AI computing infrastructure. Training frontier AI systems currently requires hardware—principally graphics processing units and specialized AI accelerators—supplied by a small number of manufacturers, most notably NVIDIA, within a geographically concentrated semiconductor supply chain (Brockman et al., 2023). Sastry et al. (2024) proposed treating compute as a governable input analogous to enriched uranium: difficult to produce, easy to count, and distributed through a highly concentrated supply chain, making it potentially amenable to monitoring and verification regimes modeled on nuclear non-proliferation frameworks.

DOI: https://doi.org/10.56738/issn29603986.geo2026.7.235
ISSN 2960-3986
GEO Academic Journal Vol. 7 No.4

### 8.3 Critical Appraisal

The political economy critique of AI raises legitimate concerns about structural power, but it also risks collapsing several analytically distinct issues into a single frame: intellectual property appropriation, labor exploitation, epistemic power, and geopolitical competition each require different conceptual tools and policy responses. Moreover, claims that current AI firms are uniquely exploitative must be situated within the broader history of platform capitalism (Srnicek, 2017), where analogous structural features predate the present era of AI development. If AI concentration is distinctive, its uniqueness lies less in the mechanisms of extraction than in the strategic importance of the assets being concentrated.

## 9. Integrated Risk Taxonomy and Governance Model

### 9.1 The Six-Risk Taxonomy

Figure 1 presents the integrated six-risk taxonomy derived from this structured review. The taxonomy is organized along two dimensions: the primary mechanism of harm (technical versus socioeconomic versus political-structural) and the temporal horizon over which harms are most likely to materialize (near-term versus long-term). Importantly, these six categories are not independent. The competitive incentive trap (Risk 3) intensifies both technical misalignment risks (Risk 1) and unsafe acceleration; agentic misalignment (Risk 5) represents a specific instantiation of Risk 1 in deployed environments; and power concentration (Risk 6) provides the structural context that shapes who bears the consequences of all other risks.

**Table 1. Integrated Six-Risk Taxonomy of Advanced AI Systemic Risks**

| No. | Risk Category | Primary Mechanism | Temporal Horizon | Key Evidence |
|---|---|---|---|---|
| R1 | Technical Misalignment | Technical | Medium–Long | Bostrom (2014); Russell (2019); Leike et al. (2018) |
| R2 | Cognitive Labour Displacement | Socioeconomic | Near–Medium | Frey & Osborne (2017); Acemoglu & Restrepo (2022); Eloundou et al. (2023) |
| R3 | Competitive Incentive Trap | Political-Economic | Near–Medium | Dafoe (2018); Harris & Raskin (2023); Arrow (1962) |

DOI: https://doi.org/10.56738/issn29603986.geo2026.7.235
ISSN 2960-3986
GEO Academic Journal Vol. 7 No.4

| | | | | |
|---|---|---|---|---|
| R4 | Institutional Manipulation | Sociotechnical | Near–Medium | Goldstein et al. (2023); Chesney & Citron (2019) |
| R5 | Agentic Misalignment | Technical–Behavioural | Near–Medium | Anthropic (2025); Park et al. (2023); Krakovna et al. (2020) |
| R6 | Power Concentration | Structural–Political | Near–Long | Crawford (2021); Sastry et al. (2024); Couldry & Mejias (2019) |

### 9.2 The Four-Level Governance Framework

Drawing on the multilevel governance literature (Jordan et al., 2005; Hooghe & Marks, 2003) and the AI governance scholarship synthesized in this review, we propose a four-level governance framework that maps specific interventions to risk categories and governance levels. The framework is presented in Table 2.

**Table 2. Four-Level Governance Framework for Advanced AI Risks**

| Level | Primary Actor(s) | Key Interventions | Risks Addressed |
|---|---|---|---|
| Individual | Workers, citizens, consumers | Skills development in human-centered roles; epistemic hygiene; collective data rights | R2, R4 |
| Developer / Firm | AI labs, technology firms, institutional deployers | Narrow system design; safety red lines; independent evaluation; labor standards for annotators | R1, R3, R5, R6 |
| State / National | Governments, regulators, central banks | Mandatory liability insurance; compute registries; anti-trust enforcement; labour transition programmes; AI-specific safety standards | R2, R3, R5, R6 |

DOI: https://doi.org/10.56738/issn29603986.geo2026.7.235
ISSN 2960-3986
GEO Academic Journal Vol. 7 No.4

| | | | |
|---|---|---|---|
| International | Treaty bodies, multilateral institutions, standards organisations | Global red lines (autonomous weapons, nuclear command AI); compute governance regimes; shared safety standards; data sovereignty frameworks | R3, R4, R6 |

### 9.3 Cross-Level Coherence and Design Principles

Effective governance requires more than isolated interventions at each level; it depends on coherence across levels. National regulations that fail to account for international competitive dynamics are likely to be undermined by regulatory arbitrage, while firm-level safety commitments that lack liability mechanisms remain structurally fragile. Three design principles follow from this analysis. First, interventions should be designed to alter the payoff structure of the competitive incentive trap rather than rely on the voluntary goodwill of developers. Second, governance mechanisms should be calibrated to the governability of the underlying resource: as Sastry et al. (2024) argue, compute is significantly more governable than either data or algorithmic knowledge. Third, labor and distributional consequences should be addressed proactively rather than reactively, since the political economy of delayed adjustment is likely to be severe (Acemoglu & Restrepo, 2022).

## Conclusion

The trajectory of advanced AI development will be determined not by technological destiny, but by human choices regarding incentives, institutional design, regulation, and collective political will. The dangerous outcomes identified in this review are structurally plausible, and the theoretical and empirical grounds for concern are substantial, but they are not predetermined. Whether AI development yields broadly shared benefits or instead produces concentrated harm and disruption will depend on whether societies are able to develop and implement governance at a pace commensurate with the technology's advancement. The analysis presented here suggests that such governance is both feasible and urgent.

## References

Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. Journal of Economic Perspectives, 33(2), 3–30. https://doi.org/10.1257/jep.33.2.3

DOI: https://doi.org/10.56738/issn29603986.geo2026.7.235
ISSN 2960-3986
GEO Academic Journal Vol. 7 No.4

Acemoglu, D., & Restrepo, P. (2022). Tasks, automation, and the rise in US wage inequality. Econometrica, 90(5), 1973–2016. https://doi.org/10.3982/ECTA19815

Amodei, D. (2025, May). Remarks on AI and the labor market. Axios Summit on AI and the Economy.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565. https://arxiv.org/abs/1606.06565

Anthropic. (2023). Claude's constitution. https://www.anthropic.com/index/claudes-constitution

Anthropic. (2025). Agentic misalignment: How LLMs could be insider threats. https://www.anthropic.com/research/agentic-misalignment

Arntz, M., Gregory, T., & Zierahn, U. (2017). Revisiting the risk of automation. Economics Letters, 159, 157–160. https://doi.org/10.1016/j.econlet.2017.07.001

Arrow, K. J. (1962). Economic welfare and the allocation of resources for invention. In R. Nelson (Ed.), The rate and direction of inventive activity (pp. 609–626). Princeton University Press.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hyun, D., Kaplan, J., Kernion, J., Lovitt, L., Nanda, N., Olsson, C., … Clark, J. (2021). A general language assistant as a laboratory for alignment. arXiv preprint arXiv:2112.00861. https://arxiv.org/abs/2112.00861

Autor, D. H. (2022). The labor market impacts of technological change: From unbridled enthusiasm to qualified optimism to vast uncertainty. NBER Working Paper No. 30074. National Bureau of Economic Research.

Bai, H., Voelkel, J. G., Eichstaedt, J. C., & Willer, R. (2023). Artificial intelligence can persuade humans on political topics. SSRN Working Paper. https://doi.org/10.2139/ssrn.4380526

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brockman, G., Ji, Y., Metz, C., Ngo, H., Papadimitriou, C., & Sutton, R. (2023). Compute and the future of AI. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, pp. 15745–15746).

Brynjolfsson, E., Li, D., & Raymond, L. (2023). Generative AI at work. NBER Working Paper No. 31161. National Bureau of Economic Research.

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial intelligence and the 'good society': The US, EU, and UK approach. Science and Engineering Ethics, 24(2), 505–528. https://doi.org/10.1007/s11948-017-9901-7

Chesney, R., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, 107(6), 1753–1820.

Cihon, P., Maas, M. M., & Floridi, L. (2020). Should artificial intelligence governance be centralised? Design lessons from history. Minds and Machines, 30(3), 311–337. https://doi.org/10.1007/s11023-020-09521-8

Couldry, N., & Mejias, U. A. (2019). Data colonialism: Rethinking big data's relation to the contemporary subject. Television & New Media, 20(4), 336–349. https://doi.org/10.1177/1527476418796632

Crawford, K. (2021). Atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

DOI: https://doi.org/10.56738/issn29603986.geo2026.7.235
ISSN 2960-3986
GEO Academic Journal Vol. 7 No.4

Dafoe, A. (2018). AI governance: A research agenda. Future of Humanity Institute, University of Oxford.

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. arXiv preprint arXiv:2303.10130. https://arxiv.org/abs/2303.10130

Epoch AI. (2024). Trends in machine learning. https://epochai.org/trends

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.8cd550d1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4PeopleAn ethical framework for a good AI society. Minds and Machines, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? Technological Forecasting and Social Change, 114, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). Generative language models and automated influence operations: Emerging threats and potential mitigations. arXiv preprint arXiv:2301.04246. https://arxiv.org/abs/2301.04246

Green, B. N., Johnson, C. D., & Adams, A. (2006). Writing narrative literature reviews for peer-reviewed journals: Secrets of the trade. Journal of Chiropractic Medicine, 5(3), 101–117. https://doi.org/10.1016/S0899-3467(07)60142-6

Greenhalgh, T., Thorne, S., & Malterud, K. (2018). Time to challenge the spurious hierarchy of systematic over narrative reviews? European Journal of Clinical Investigation, 48(6), e12931. https://doi.org/10.1111/eci.12931

Hadfield-Menell, D., Milli, S., Abbeel, P., Russell, S., & Dragan, A. (2016). Inverse reward design. In Advances in Neural Information Processing Systems (Vol. 30).

Hao, K. (2025). Empire of AI: Dreams and nightmares in Sam Altman's OpenAI. Penguin Press.

Harari, Y. N. (2023, April 28). Yuval Noah Harari argues that AI has hacked the operating system of human civilisation. The Economist. https://www.economist.com/by-invitation/2023/04/28

Harris, T., & Raskin, A. (2023, March 9). The AI dilemma [Video]. Center for Humane Technology. https://www.humanetech.com/podcast/the-ai-dilemma

Hooghe, L., & Marks, G. (2003). Unraveling the central state, but how? Types of multi-level governance. American Political Science Review, 97(2), 233–243. https://doi.org/10.1017/S0003055403000649

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Jordan, A., Wurzel, R., & Zito, A. (2005). The rise of 'new' policy instruments in comparative perspective: Has governance eclipsed government? Political Studies, 53(3), 477–496. https://doi.org/10.1111/j.1467-9248.2005.00540.x

Krakovna, V., Uesato, J., Mikulik, V., Martic, M., Tomasev, N., Stepleton, T., Ngo, R., Marblestone, A., & Legg, S. (2020). Specification gaming: The flip side of AI ingenuity.

DOI: https://doi.org/10.56738/issn29603986.geo2026.7.235
ISSN 2960-3986
GEO Academic Journal Vol. 7 No.4

DeepMind Blog. https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity

Leike, J., Martic, M., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., Uesato, J., & Legg, S. (2018). AI safety gridworlds. arXiv preprint arXiv:1711.09883. https://arxiv.org/abs/1711.09883

Marcus, G. (2018). Deep learning: A critical appraisal. arXiv preprint arXiv:1801.00631. https://arxiv.org/abs/1801.00631

North, D. C. (1990). Institutions, institutional change and economic performance. Cambridge University Press.

Omohundro, S. M. (2008). The basic AI drives. In P. Wang, B. Goertzel, & S. Franklin (Eds.), Proceedings of the 2008 Conference on Artificial General Intelligence (pp. 171–179). IOS Press.

Ord, T. (2020). The precipice: Existential risk and the future of humanity. Hachette Books.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. https://doi.org/10.1145/3586183.3606763

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., & Irving, G. (2022). Red teaming language models with language models. arXiv preprint arXiv:2202.03286. https://arxiv.org/abs/2202.03286

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., & Britten, N. (2006). Guidance on the conduct of narrative synthesis in systematic reviews. ESRC Methods Programme.

Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L. J., Recchia, G., van der Bles, A. M., & van der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. Royal Society Open Science, 7(10), 201199. https://doi.org/10.1098/rsos.201199

Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Viking.

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., O'Keefe, C., Hadfield, G. K., Ngo, R., Pilz, K., Gor, G., Bluemke, E., Shoker, S., Egan, J., Trager, R. F., Avin, S., Weller, A., Bengio, Y., & Coyle, D. (2024). Computing power and the governance of artificial intelligence. arXiv preprint arXiv:2402.08797. https://arxiv.org/abs/2402.08797

Seoul AI Safety Commitments. (2024). Frontier AI safety commitments. https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024

Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Brien, C., Heitkoetter, B., Agrawal, N., Saber, A., Chan, J. S., Whitman, R., Mishkin, P., & Clark, J. (2023). Practices for governing agentic AI systems. OpenAI Technical Report.

Srnicek, N. (2017). Platform capitalism. Polity Press.

Suleyman, M., & Bhaskar, M. (2023). The coming wave: Technology, power, and the twenty-first century's greatest dilemma. Crown.

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions. In Proceedings of the 2019 AAAI/ACM

DOI: https://doi.org/10.56738/issn29603986.geo2026.7.235
ISSN 2960-3986
GEO Academic Journal Vol. 7 No.4

Conference on AI, Ethics, and Society (pp. 195–200). https://doi.org/10.1145/3306618.3314289

Winner, L. (1980). Do artifacts have politics? Daedalus, 109(1), 121–136.

Yudkowsky, E. (2022). AGI ruin: A list of lethalities. Machine Intelligence Research Institute. https://www.lesswrong.com/posts/uMQ3cqWDPHhjtiesc