Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Cascading Insecurity in the Age of Artificial General Intelligence: An Integrated Societal-Security Framework for Geopolitics, Corporate Power, Social Stability, and Cybersecurity

Alireza Nik Aein Koupaei

alireza.nik.aein.koupaei@upc.edu

Research Article

Keywords: Artificial general intelligence, Societal security, Securitisation, Corporate power, Private technological sovereignty, Cybersecurity, Systemic risk, AI governance

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11737869/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Artificial general intelligence (AGI), commonly defined as AI that matches or exceeds the cognitive versatility and proficiency of a well-educated adult, has moved from speculation into national strategies, corporate investment plans and multilateral agendas. Yet research on its security implications remains fragmented across international relations, economics, communication studies and computer security. This article argues that the most consequential AGI-related threats to societal security arise from interactions among these domains rather than within any one of them. Drawing on societal security and securitisation theory, critical political economy, scholarship on technological sovereignty and weaponised interdependence, and societal resilience theory, it develops an integrated AGI societal-security framework. The framework is applied through a structured narrative review, a comparative analysis of the United States, China and the European Union supplemented by corporate and cyber case studies, and a scenario-based assessment of three conditions: transition, AGI emergence and advanced AGI deployment. An evidence-grading convention separates observed, emerging and prospective risks throughout. The analysis identifies four transmission mechanisms (concentration, dependency, exposure and legitimacy erosion), twelve directed cross-domain pathways and three reinforcing feedback loops through which geopolitical rivalry, corporate control, socioeconomic disruption and cyber vulnerability amplify one another. It introduces the concept of *private technological sovereignty* to describe quasi-sovereign corporate control over strategically critical intelligence capabilities. The article contributes a cross-domain risk assessment matrix and a four-pillar governance framework that targets intervention points shared by several loops: frontier-asset security, corporate accountability, information integrity and socioeconomic transition.

1 Introduction

1.1 Background and Emergence of Artificial General Intelligence

Foundation models, trained on broad data and adaptable to many tasks, have moved artificial intelligence (AI) beyond narrow, task-specific systems. By early 2026, leading systems had reached gold-medal performance on International Mathematical Olympiad problems, AI agents could complete some programming tasks that take a human about half an hour, and at least 700 million people used such systems every week [1]. Artificial general intelligence (AGI) denotes the point at which this generality approaches or exceeds human breadth. Hendrycks et al. [2] define it as “an AI that can match or exceed the cognitive versatility and proficiency of a well-educated adult” and score GPT-4 at 27% and GPT-5 at 57% against that standard. Expert expectations remain dispersed: the largest survey of AI researchers gave a 50% aggregate probability that unaided machines would outperform humans at every task by 2047 [3]. For security analysis, AGI’s distinguishing features are generality, autonomy and scalability: a system able to perform most cognitive work can be copied, run in parallel and embedded in economic, informational and military processes at machine speed.

1.2 The Changing Landscape of Societal Security

Security studies have long moved beyond territorial defence. The Copenhagen School widened the agenda to political, economic, societal and environmental sectors [4] and defined societal security as a society's ability "to persist in its essential character under changing conditions and possible or actual threats" [5, p. 23]. Digital dependence has since made technological, economic, informational and institutional security central, and much critical infrastructure is privately owned and interdependent [6]. AGI touches all four dimensions at once: it is a strategic technology, an economic input, a producer of information and a potential component of critical infrastructure.

1.3 Research Problem and Knowledge Gap

AGI-related risks are analysed in parallel rather than in combination. International relations treats AI as a source of power [7], economists estimate displacement [8], communication scholars study persuasion and synthetic media, and security researchers examine AI-enabled attack and defence [9]. International governance itself is active: General Assembly resolution 79/325 created an Independent International Scientific Panel on AI and a Global Dialogue on AI Governance, whose first session in July 2026 drew delegations from 163 countries and "highlighted the concentration of AI development and governance power" [10]. The gap is therefore specific: **the insufficient integration of geopolitical power concentration, private corporate control, societal vulnerability and strategic cybersecurity within a unified societal-security framework for assessing the potential emergence and deployment of AGI.** In addition, corporations are usually treated as objects of regulation rather than as security actors, and structural risks that travel across domains remain under-theorised [11].

1.4 Research Aim, Objectives and Questions

The aim is to develop and apply an integrated framework explaining how AGI could transform societal security through interacting changes in four domains. The objectives are to synthesise existing research (O1), build a theoretically grounded framework (O2), assess domain risks with graded evidence (O3), model cross-domain cascades (O4) and derive governance mechanisms (O5).

Main Research Question. How could the development and potential deployment of AGI transform societal security through changes in geopolitical power, corporate influence, socioeconomic stability and cybersecurity?

- **RQ1 (geopolitical)**: How might AGI reshape international power relations, strategic competition, technological sovereignty and dependencies between major international actors?
- **RQ2 (societal)**: Through what mechanisms could AGI-driven economic, informational and political transformations affect social cohesion, institutional legitimacy, public trust and societal stability?
- **RQ3 (corporate)**: To what extent could the concentration of AGI capabilities within major technology corporations create new forms of economic dominance, political influence, institutional tension and national security dependency?

- **RQ4 (cyber)**: How could AGI development and deployment transform cybersecurity threats, particularly industrial espionage, strategic data exfiltration, intellectual property protection and critical digital infrastructure?
- **RQ5 (governance)**: What governance mechanisms and international cooperation arrangements are needed to address interconnected AGI-related risks and strengthen societal and international security?

1.5 Original Contributions and Organisation of the Paper

The article contributes (1) an integrated framework specifying four transmission mechanisms (concentration, dependency, exposure and legitimacy erosion) across twelve directed pathways between four domains; (2) the concept of *private technological sovereignty*; (3) an evidence-grading convention separating observed, emerging and prospective risks; and (4) a cross-domain risk assessment matrix and a multidimensional governance framework. Sections 2–4 present the literature, theory and method; Sects. <link rid="Sec28">5</link>–8 analyse the four domains; Sect. 9 integrates them; and Sects. 10–12 address governance, implications and conclusions.

2 Literature Review and Identification of Research Gaps

2.1 Existing Research on AGI Capabilities and Associated Risks

Early AGI scholarship was largely theoretical and warned of misaligned superhuman systems [12]. Foundation models made the debate empirical, through graded “levels of AGI” [13] and psychometric operationalisations [2]. Risk research has broadened from misalignment to misuse and systemic harm [14]. Kasirzadeh [15] distinguishes decisive risks from accumulative risks that build through the gradual erosion of economic and political systems. Uncertainty is pervasive: AGI lacks an agreed definition [16], some argue that AI is a “normal technology” paced by institutional adoption [17], and policymakers face an “evidence dilemma” between acting prematurely and acting too late [1]. This strand rarely connects risks to the political and economic structures that transmit them.

2.2 AI and International Security

International relations scholarship treats AI as a general-purpose technology that can shift the balance of power [7], although power may accrue to states that diffuse it widely rather than to first innovators [18]. Strategic studies examine effects on decision speed and escalation [19], race models show how competition erodes safety [20], and recent work proposes deterrence through “mutual assured AI malfunction” [21]. Research on weaponised interdependence explains how control of network hubs becomes coercive leverage [22], and governance scholarship maps institutional options and the obstacles posed by rivalry [23]. This literature remains state-centric: firms appear as instruments of states rather than as actors with divergent interests.

2.3 AI-Driven Socioeconomic and Political Transformations

Economic research finds wide exposure but contested effects. The IMF estimates that almost 40% of global employment is exposed to AI [24], and the ILO places one in four workers in exposed occupations [25]. Measured effects so far fall on entry-level workers [26], aggregate productivity gains may be modest [27], and distribution depends on whether AI automates or augments labour [28]. Political research documents the risks of synthetic media [29] and large persuasion effects from training and prompting language models [30], while others argue that fears of a misinformation surge are overstated because demand, not supply, is the binding constraint [31]. These effects are rarely linked to the ownership of the systems involved.

2.4 Corporate Concentration and Technological Power

Political economy shows that digital markets tend towards concentration [32] and that leading firms act as intellectual monopolies [33]. Foundation models display natural-monopoly tendencies [34]. Competition authorities have mapped an interconnected web of more than 90 partnerships and investments linking incumbents to AI developers [35, 36]. Compute has become a lever of governance [37] and a marker of global inequality [38], and Big Tech's power over policy and international politics is well documented [39]. Corporate power is studied as a market or democratic problem, but seldom as a security variable.

2.5 AI and the Changing Cybersecurity Landscape

AI is expected to lower the cost and raise the scale of attacks while also strengthening defence [9]. The UK National Cyber Security Centre judged that AI would "almost certainly" increase the volume and heighten the impact of cyber attacks [40], and threat intelligence now documents state-linked use of AI in live operations [41, 42]. Research on the security of AI assets identifies 38 attack vectors against frontier model weights [43] and demonstrates partial extraction of production models through their public interfaces [44]. Few studies, however, trace how such incidents propagate into geopolitical or societal outcomes.

2.6 Research Gaps and the Positioning of the Present Study

Three gaps emerge. First, there is no integrated societal-security perspective: each strand treats the others' central variables as exogenous. Second, corporate power is under-analysed as a security variable. Third, cascading risks across the geopolitical, societal, corporate and cyber domains are rarely investigated, although the most severe outcomes are likely to be compound ones. The present study addresses all three.

3 Conceptual and Theoretical Framework

3.1 Defining AGI and Distinguishing It from Contemporary AI

Narrow AI performs bounded tasks. *General-purpose AI* models display "significant generality" and competently perform "a wide range of distinct tasks" [45, Art. 3(63)]; today's frontier models belong here. *AGI* denotes systems that match or exceed skilled adults across most cognitive tasks [2, 13], and *artificial superintelligence* denotes systems that greatly exceed human performance in virtually all domains [12]. Five attributes matter for security: generality, autonomy, adaptability, decision-making capability and scalability, since a trained model can be copied and run as often as compute allows. Because definitions remain contested [16, 46], AGI is treated here as a continuum, with a working definition for scenario construction: *AI systems that perform most cognitive tasks at or above the level of skilled human adults, pursue extended goals with limited human oversight, and can be replicated and deployed at scale*.

3.2 The Concept of Societal Security

For the Copenhagen School, societal security concerns the survival of collective identity [4, 5]. This article broadens the referent object to a society's capacity to reproduce four things: *identity and cohesion*; *socioeconomic security*, meaning secure livelihoods and a distribution of gains perceived as fair; *epistemic security*, meaning the integrity of the information environment [47]; and *institutional stability*. Resilience, the capacity to absorb shocks, adapt and transform [48], determines whether shocks are contained. Its inverse, systemic vulnerability, arises where tight coupling lets local failures cascade [49].

3.3 Securitisation Theory and Technological Threats

Securitisation occurs when an actor presents a referent object as existentially threatened and an audience accepts extraordinary measures [4]. Hansen and Nissenbaum [50] show that cyber threats are securitised through hypersecuritisation, everyday security practices and technification, all of which are evident in AGI discourse. Three features are distinctive. Governments securitise AGI leadership, as the title of the United States' 2025 plan, *Winning the Race*, signals [51], and China has securitised AI to mobilise firms and local governments [52]. Developers themselves act as securitising actors: in 2023 the heads of leading laboratories endorsed the statement that mitigating "the risk of extinction from AI should be a global priority" [53]. Firms can also securitise their products as national-security assets to obtain protection or regulatory restraint. The analysis therefore asks who gains authority by declaring AGI a threat.

3.4 Technological Power and International Relations

Realists treat technological superiority as a driver of power transitions [7], while diffusion theory stresses economy-wide adoption [18]. States that control network hubs can weaponise interdependence through panopticon and chokepoint effects [22]. Technology sovereignty, the ability to provide critical technologies "without one-sided structural dependency" [54], has become a central policy goal, and

Bradford [55] distinguishes American market-driven, Chinese state-driven and European rights-driven models of digital governance. These perspectives explain why AGI is treated as a strategic asset and why dependency is experienced as insecurity.

3.5 Corporate Power and Non-state Security Actors

Business power is instrumental, structural and discursive [56], private actors co-produce security in "global security assemblages" [57], and technology firms exercise sovereignty-like control over digital space [58]. Building on this work, the article defines **private technological sovereignty** as *the de facto capacity of a private actor to control the development, access, terms of use and continuity of a strategically critical technological capability, such that states and societies depend on its decisions without commensurate means of oversight or substitution.* Its four markers are control of chokepoint assets, private rule-setting through usage policies, transnational reach and short-term non-substitutability. The concept is functional rather than legal.

3.6 Integrating the Theoretical Perspectives

The integration rests on one claim: AGI-related insecurity is produced less within domains than at their intersections. Each perspective contributes one element (Table 1).

Table 1
Theoretical perspectives and their role in the framework

Perspective	Core concepts	Role in this article
Societal security	Identity, cohesion, sectors	Defines the referent objects
Securitisation	Speech acts, technification	Explains threat construction and who gains authority
Critical political economy	Intellectual monopoly; structural power	Explains concentration and unequal gains
Technological sovereignty and interdependence	Chokepoints; weaponised interdependence	Explains competition and dependency
Societal resilience	Adaptive capacity; systemic risk	Explains whether shocks are dampened or cascade

The framework (Fig. 1) has four layers: a *driver layer*, the AGI capability trajectory under the scenarios of Sect. 4; a *domain layer* of geopolitical, societal, corporate and cyber security; a *transmission layer* of four mechanisms, **concentration**, **dependency**, **exposure** and **legitimacy erosion**; and a *governance and resilience layer* that dampens or amplifies transmission. Five propositions guide the analysis:

- **P1 (concentration)**: the more concentrated frontier capabilities become, the more corporate shocks are transmitted to the geopolitical and societal domains.
- **P2 (dependency)**: reliance on privately or foreign-controlled AI infrastructure creates channels of coercion and failure.

- **P3 (exposure)**: the strategic value of concentrated AI assets increases incentives for state-sponsored cyber operations against them.
- **P4 (legitimacy erosion)**: disruption and manipulation reduce institutional trust, weakening the resilience on which the governance of every domain depends.
- **P5 (amplification)**: AGI-level capabilities raise the speed, scale and autonomy of each mechanism.

4 Research Methodology and Analytical Design

4.1 Research Design

The study uses a qualitative, interdisciplinary and exploratory-analytical design that combines a structured narrative review, comparative case analysis and scenario-based assessment. It is theory-building rather than theory-testing [59]: it produces a framework and propositions for later empirical testing. Section 9 applies the logic of process tracing [60] to reconstruct cross-domain causal steps from documented incidents.

4.2 Data Collection and Sources

Five types of sources from January 2018 to October 2026 were analysed: peer-reviewed research; reports of international organisations, including the UN, IMF, ILO and the International AI Safety Report; national AI strategies, legislation and cybersecurity frameworks; corporate governance and transparency documents, including threat-intelligence reports; and documented incidents and court records. Sources were selected purposively for their relevance to at least one domain and one mechanism. Journalistic sources were used only to establish that events occurred.

4.3 Comparative Case-Study Approach

Country cases follow a most-different logic across the three dominant models of digital governance [55]: the United States, China and the European Union, with middle powers and developing countries drawn in where they illuminate dependency. Corporate and cyber cases were chosen because each makes a mechanism observable: the OpenAI governance crisis and recapitalisation (concentrated decision authority); Starlink in Ukraine and the US Department of Defense–Anthropic dispute (private technological sovereignty); the Storm-0558 intrusion, the Volt Typhoon campaign, the GTG-1002 espionage campaign and a 2026 trade-secret conviction (exposure); and the 2024 CrowdStrike outage (cascading failure from a single vendor).

4.4 Scenario-Based Risk Assessment

Scenarios are used in an explorative rather than predictive sense [61]. Following Korinek and Suh [62], three conditions are distinguished, without dates or probabilities (Table 2). Risks are traced as *direct* (within a domain), *indirect* (crossing one boundary) or *cascading* (crossing two or more boundaries or feeding back to their origin).

Table 2
Scenario conditions used in the analysis

Scenario	Defining assumptions	Illustrative markers
S1 Transition	More capable general-purpose systems and agents; no AGI; human oversight still needed	Longer agent task horizons; sectoral automation; AI-enabled cyber operations (largely observed)
S2 AGI emergence	One or a few actors develop broadly human-level systems; deployment limited by compute, regulation and absorption	Most remote cognitive work done at professional level; leads measured in months
S3 Advanced AGI deployment	AGI integrated into economic, political, corporate and critical infrastructure with substantial autonomy	Machine-speed decisions; AI agents running critical services; AI-accelerated research

4.5 Analytical Dimensions and Evaluation Criteria

Each domain is assessed on two criteria: geopolitical power concentration and strategic dependency; societal inequality and institutional resilience; corporate concentration of control and accountability; and cyber exposure and systemic vulnerability. Risks in the matrix in Sect. 9 are rated qualitatively for likelihood, impact and cascade potential; the ratings are structured judgements, not probabilities. An evidence-grading convention keeps observation and projection apart, with claims tagged **[O]**, **[E]** or **[P]**: **[O]** observed, meaning documented with existing AI systems; **[E]** emerging, meaning early evidence of trends likely to intensify as systems become more capable and autonomous; and **[P]** prospective, meaning consequences conditional on S2 or S3.

4.6 Research Limitations and Assumptions

AGI capabilities and timelines are uncertain, so the scenarios are heuristics, not forecasts. Because AGI does not yet exist, contemporary AI serves as an analogue that may understate novel risks. Several sources are produced by interested parties, and the source base is predominantly English-language. The analysis assumes that institutions remain broadly intact in S1 and S2 and that AGI development remains concentrated among a few actors.

5 AGI and the Transformation of International Relations

5.1 AGI as a New Source of Geopolitical Power

As a general-purpose technology with direct strategic uses, AGI would redistribute power [7]. Its material base is already concentrated: the United States and China hold roughly 75% and 15% of the computing performance of the world's largest AI supercomputers [63] **[O]**, and "the widening AI divide in computation, data, skills and institutional capacity" was the concern raised most consistently at the first UN Global Dialogue [10] **[O]**. Under S2, a lead of months could become durable if AGI accelerated AI research itself [12] **[P]**, although diffusion and institutional absorption may slow the conversion of capability into power [17, 18].

5.2 Strategic Competition among Major Powers

Competition has already reshaped trade policy **[0]**. US export controls on advanced chips became bargaining chips in 2025: an August arrangement for chipmakers to remit 15% of certain China sales to the US government was followed in December by approval of Nvidia H200 exports to vetted Chinese buyers in return for a 25% fee [64]. China answered with indigenous development, visible in DeepSeek-R1's near-frontier performance in January 2025, and with counter-leverage: rare-earth export controls announced on 9 October 2025 were suspended for a year only after a leaders' meeting [65]. Each side targets the other's chokepoints [22] **[0]**, creating race dynamics that can erode investment in safety [20] **[E]**. Under S2, states might deter rivals' bids for dominance by threatening to sabotage their AI projects [21] **[P]**, tying geopolitics directly to the cyber domain.

5.3 AGI and the Transformation of Geopolitical Influence

Militarily, AI already compresses intelligence and targeting cycles [19] **[0]**. Economically, AI services have become instruments of influence: the United States promotes exports of its "AI technology stack", while China has built influence largely through open-weight model releases [66] **[0]**. Diplomatically, dependency becomes leverage. States unable to build frontier systems will rely on foreign or corporate providers for capabilities that may come to support core state functions **[E]**, and under S3 access to AGI services could become conditional on political alignment **[P]**. Power would shift towards actors that combine compute, energy, data and talent, some of which are firms rather than states.

5.4 International Cooperation and Regulatory Competition

National approaches diverge **[0]**. The EU's AI Act has imposed obligations on providers of general-purpose models since August 2025 [45]. China combines sectoral rules with a push for a World AI Cooperation Organization, which 29 countries agreed to establish in July 2026, and the United States favours deregulation and voluntary arrangements, including a 2026 framework for voluntary pre-release government access to frontier models [66]. Verification problems, dual use and distrust impede cooperation **[0]**, yet shared catastrophic risks offer common ground, as the 2024 US–China affirmation of human control over decisions to use nuclear weapons shows **[0]**.

5.5 Implications for International Institutions and Global Order

Security institutions built for states and physical goods fit AGI poorly: export controls struggle with weightless software, arms-control verification has no analogue for model training, and firms are not parties to treaties **[0]**. The UN Dialogue and its 40-member Scientific Panel offer universal deliberation, but the Dialogue's themes were recorded as "recurring observations, not negotiated outcomes" [10]. Three gaps remain: no verification of compute use, no intergovernmental channel for sharing incidents and no forum that holds corporate developers accountable alongside states [23] **[P]**. In answer to RQ1, AGI is likely to reshape international relations less through a single power transition than through concentrated compute, weaponised dependencies and the growing strategic role of firms.

6 AGI, Societal Security and the Potential for Social Instability

6.1 Labour-Market Transformation and Socioeconomic Inequality

Generative AI, and prospectively AGI, reaches cognitive and professional work [8]. Exposure is wide and uneven **[O]**: the ILO finds exposure at 34% of employment in high-income countries against 11% in low-income countries, with women over-represented in the most exposed category [25]. Effects fall first on career entry. The International AI Safety Report finds declining demand for early-career workers in exposed occupations [1], and US payroll data show employment of 22–25-year-olds in the most exposed occupations about 19% below its counterfactual by June 2026 [26] **[O]**. If returns accrue mainly to owners of compute and models, inequality rises [28] **[E]**; in scenarios where AGI automates all tasks, wages may collapse even as output grows [62] **[P]**. In a 28-country survey, 54% of low-income respondents already expect to be “left behind” by generative AI [67] **[O]**.

6.2 Information Manipulation and the Erosion of Public Trust

In 2024 an employee of the engineering firm Arup transferred about US$25 million after a video call with deepfaked executives [68] **[O]**. A review of the 2024 elections found no conclusive evidence that AI content changed results, but found that it amplified harmful narratives and eroded trust [69] **[O]**. Manipulation can also trigger institutional crisis: on 6 December 2024 Romania’s Constitutional Court annulled a presidential first round after declassified intelligence alleged a coordinated campaign on TikTok and Telegram [70] **[O]**. Under S2–S3, autonomous agents could run adaptive, multilingual influence operations at negligible cost **[P]**.

6.3 Political Polarisation and Institutional Legitimacy

Evidence that algorithms drive polarisation is mixed: altering Facebook and Instagram feeds during the 2020 US election changed exposure but not measured polarisation [71] **[O]**. Social distrust is nonetheless high, with 70% of respondents in 28 countries unwilling or hesitant to trust people with different values or information sources [67] **[O]**. Automated state decisions can damage legitimacy directly, as shown by the Dutch childcare-benefits scandal, which brought down a government in 2021 [72] **[O]**. As AI assistants become primary political intermediaries, their design choices become de facto political choices made by a few firms **[E]**.

6.4 Algorithmic Influence and Societal Behavioural Manipulation

Training and prompting raised the persuasiveness of language models by up to 51% while reducing their factual accuracy, whereas personalisation added less [30] **[O]**. Persuasion can also be benign: dialogues with an AI model durably reduced conspiracy beliefs [73] **[O]**. The security problem is concentration.

Whoever controls the models that millions consult daily, including AI companions now used by tens of millions [1], holds unprecedented behavioural influence **[E]**. Under S3, agentic systems could pursue persuasion goals continuously and autonomously **[P]**.

6.5 Social Fragmentation and the Potential for Civil Unrest

Unrest becomes likely when grievance, perceived injustice and mobilising narratives coincide. Economic shocks have produced backlash before: vulnerability to industrial-robot adoption raised support for the radical right in Western Europe [74] **[O]**. In 2023, striking Hollywood writers won contractual limits on the use of AI **[O]**. AGI-driven disruption would differ by affecting educated groups with high political voice at a time when grievances can be amplified or manufactured at scale **[P]**. Economic grievance, low trust and manipulable information together form the configuration most likely to turn technological change into fragmentation and, in fragile states, unrest **[P]**.

6.6 Societal Resilience and Mechanisms of Adaptation

Resilience can be built in three ways: through digital literacy and "prebunking" [75] **[O]**, backed by provenance and labelling rules such as Article 50 of the EU AI Act; through social protection and labour-market adaptation, including wage insurance, lifelong learning and benefit-sharing such as a "windfall clause" [76] **[P]**; and through institutional trust built on transparency, contestability and human review. In answer to RQ2, AGI threatens societal security through concentrated displacement, degraded information environments and delegitimised institutions, which reinforce one another.

7 Corporate Power, Institutional Tensions and AGI Competition

This section asks what happens when strategically important intelligence capabilities are concentrated in a few private firms.

7.1 Concentration of AGI Capabilities within Major Corporations

Concentration runs through every layer of the AI stack **[O]**. AI accelerator design, leading-edge fabrication and extreme-ultraviolet lithography are each dominated by a single firm [77]. Amazon, Microsoft and Google together held 63% of the global cloud infrastructure market in the second quarter of 2026 [78]. Frontier laboratories are tied to cloud providers through equity, compute and distribution agreements [35, 36], and Amazon, Alphabet, Microsoft and Meta together guided towards US$635–665 billion of capital expenditure for 2026, up from about US$381 billion in 2025 [79] **[O]**. Under S2, AGI would most likely emerge inside one of a few firms, placing a strategically decisive capability under private control **[P]**.

7.2 Competition over AGI Infrastructure and Technological Resources

Data centres used about 415 TWh of electricity in 2024, and the International Energy Agency projects about 945 TWh by 2030 [80] **[O]**. Scarce talent is acquired through licensing-and-hiring deals that can escape merger review **[O]**. These dynamics raise barriers to entry and reinforce natural-monopoly tendencies [34] **[E]**. Open-weight releases counteract concentration at the model layer **[O]** but do not redistribute compute, energy or fabrication capacity, leaving a layered oligopoly with high barriers for most firms and states **[E]**.

7.3 Corporate–State Relationships and Geopolitical Influence

In 2022 technology firms helped Ukraine move government data to the cloud and defend against Russian cyber operations [81] **[O]**. The same war revealed private technological sovereignty. Ukrainian forces came to depend on Starlink; Reuters reported that its owner ordered coverage disabled over part of Kherson Oblast during the 2022 counteroffensive, which he denied, and in March 2025 he stated that Ukraine's "entire front line would collapse if I turned it off" [82] **[O]**.

Disputes over terms of use can become security disputes **[O]**. In 2025–2026 the US Department of Defense sought to remove contractual restrictions that barred the use of Anthropic's models for mass domestic surveillance or fully autonomous lethal weapons; when the company refused, the Secretary of Defense designated it a "supply chain risk". In August 2026 a district judge held the government's actions to be unlawful retaliation, while in September 2026 the D.C. Circuit, by two votes to one, rejected the company's separate challenge to the designation [83]. The case, still in litigation, shows a firm's usage policy acting as private law over a strategic capability colliding with a state's use of a security instrument against a domestic supplier. Firms also exercise instrumental power: in August 2025 industry figures launched a super PAC with more than US$100 million to oppose candidates and state laws seen as slowing AI [84] **[O]**.

7.4 Corporate Governance Conflicts and Institutional Tensions

In November 2023 OpenAI's board removed its chief executive for not being "consistently candid" with it [85], only to see him reinstated within days under pressure from employees and investors **[O]**. An October 2025 recapitalisation left a nonprofit foundation holding about 26% of, and controlling, a public benefit corporation in which Microsoft holds about 27%, with any declaration of AGI subject to verification by an expert panel [86] **[O]**. These episodes reveal tensions between commerce and safety, mission and capital, concentrated authority and accountability, and among rival firms. Under S2, a small board under competitive pressure could decide whether to deploy a system with global consequences **[P]**.

7.5 Corporate Accountability and Societal Security

In 2025, twelve companies published or updated frontier safety frameworks, but most risk management remains voluntary [1] **[O]**. California's 2025 frontier AI transparency law and the EU AI Act have begun to impose disclosure and incident-reporting duties **[O]**, while liability for diffuse harms remains unsettled

[E]. Independent auditing [87] and public-interest obligations modelled on utility regulation [88] are still missing **[P]**. In answer to RQ3, concentration produces economic dominance, political influence [56], institutional tension and national-security dependency. The Starlink and Defense Department cases show that private control can either protect rights or obstruct legitimate state functions; the problem is that this relationship is negotiated case by case rather than institutionalised.

8 Cybersecurity, Strategic Data Exfiltration and Technological Vulnerabilities

8.1 AGI and the Transformation of Cybersecurity Threats

In November 2025 Anthropic reported disrupting an espionage campaign, attributed with high confidence to a Chinese state-sponsored group, in which an AI agent performed an estimated 80–90% of the work against about thirty targets, with human operators intervening at only four to six decision points; a small number of intrusions succeeded, and hallucinated results limited full autonomy [41] **[O]**. Google reported the first malware observed querying a language model during live operations, used by a Russian state actor against Ukraine, and a maturing underground market in AI tools [42] **[O]**. In one competition, an AI agent found 77% of the vulnerabilities present in real software [1] **[O]**. The same capabilities strengthen defence, but slow patching in legacy systems favours attackers **[E]**. Under S2–S3, autonomous agents could run end-to-end campaigns at machine speed and shrink the interval between disclosure and exploitation towards zero **[P]**.

8.2 Industrial Espionage and Strategic Data Exfiltration

In January 2026 a former Google engineer was convicted on seven counts of economic espionage and seven of theft of trade secrets for taking more than 2,000 pages on the company's AI supercomputing infrastructure while secretly affiliated with China-based companies [89] **[O]**. In 2023 an intruder obtained details of AI designs from OpenAI's internal messaging systems [90] **[O]**. Nevo et al. [43] identify 38 vectors for stealing model weights and argue that few organisations can resist the most capable states **[E]**. Under S2, exfiltrating AGI weights would transfer a strategic capability in a single file **[P]**.

8.3 Intellectual Property Theft and Technological Competition

Espionage has historically struggled to close technological gaps because stolen designs must be absorbed through tacit knowledge and industrial capacity [91]. Model weights weaken this barrier because they are executable **[E]**. Extraction can also occur without intrusion: researchers recovered parts of production models through public interfaces [44], and in February 2026 Anthropic alleged that three Chinese developers had used about 24,000 fraudulent accounts to generate more than 16 million exchanges for distillation; the companies did not publicly respond [92] **[O]**. Such leakage erodes the leads that export controls protect and spreads capabilities without safeguards **[E]**. Each incident also

increases pressure to securitise AI laboratories, linking the cyber domain back to the corporate and geopolitical domains **[E]**.

8.4 Cyberattacks against Critical Digital Infrastructure

US and allied agencies reported in 2024 that the Volt Typhoon group had pre-positioned in communications, energy, transport and water networks to enable disruption in a crisis [93] **[O]**. Embedding AI in infrastructure adds attack surfaces such as data poisoning and prompt injection [94] **[E]**, and interdependence lets disruptions propagate across sectors [6]. Under S3, AGI agents operating such systems would raise both the payoff of compromise and the speed of cascades **[P]**.

8.5 Corporate Cybersecurity and National Security Vulnerabilities

In 2023 a China-linked actor used a stolen signing key to access Microsoft-hosted email accounts of senior US officials, an intrusion the Cyber Safety Review Board attributed to “a cascade of security failures” [95] **[O]**. The 2024 backdoor in the open-source library xz Utils showed how software supply chains can be subverted through cultivated trust **[O]**. Failures need not be malicious: in July 2024 a faulty CrowdStrike update crashed about 8.5 million Windows devices, disrupting airlines, hospitals and banks [96] **[O]**. AI development adds compromised training pipelines, poisoned data, insiders with access to weights and over-privileged agents **[E]**. When a few firms host the most capable models, a breach at one becomes an international security event **[P]**.

8.6 Cybersecurity Resilience and Strategic Protection

Protection requires security levels for frontier developers calibrated to the value of their assets, including insider-risk programmes and hardware protection of weights [43]; zero-trust access, multi-party authorisation, monitoring for abuse and distillation, and customer verification by compute providers [37]; and international cooperation, including the permanent UN mechanism on state behaviour in cyberspace agreed in 2025, with the norm against damaging critical infrastructure extended to AI infrastructure **[P]**. In answer to RQ4, AGI automates offence and defence, turns trained models into high-value and immediately usable targets, and couples corporate security failures to national and international security.

9 Interconnection of AGI-Related Security Risks: A Systemic Perspective

9.1 Convergence of Geopolitical, Corporate, Societal and Cyber Risks

Sections <link rid="Sec28">5</link>–8 show that each domain’s principal risks originate partly in the others. Table 3 maps all twelve directed pathways between the four domains (G geopolitical, C corporate, S societal, Y cyber), the mechanism that carries each, an observed illustration and its plausible amplification under AGI.

Table 3
Cross-domain transmission pathways

Pathway (mechanism)	Observed illustration [O]	Amplification under S2–S3 [P]
G → C (concentration)	Export-control bargaining; national champions	State–firm fusion; safety subordinated to speed
G → S (legitimacy erosion)	Alleged foreign campaign before Romania's annulled vote	Securitisation justifies surveillance
G → Y (exposure)	Volt Typhoon; AI-agent espionage	Sabotage of rival AGI projects
C → G (dependency)	Starlink in Ukraine; Defense Department–Anthropic dispute	State functions depend on a few providers
C → S (concentration)	Early-career displacement; AI information gatekeepers	Gains to capital; persuasion at scale
C → Y (exposure)	Storm-0558; CrowdStrike outage	AGI hosts as single points of failure
S → G (legitimacy erosion)	Trade- and automation-driven backlash	Instability weakens external capacity
S → C (legitimacy erosion)	2023 writers' strike; litigation	Licence-to-operate crises
S → Y (exposure)	Deepfake fraud; insider recruitment	Grievance raises insider risk
Y → G (exposure, dependency)	AI trade-secret theft; alleged distillation	Weight theft shifts power overnight
Y → S (legitimacy erosion)	Infrastructure intrusions; synthetic-media fraud	Machine-speed disruption of services
Y → C (concentration)	Breaches at AI firms; rising security costs	Only top-security firms can build AGI

9.2 Cascading Threats and Systemic Vulnerabilities

Cascades occur when a shock crosses domain boundaries faster than institutions can contain it [49]. Three sequences are especially plausible. Cyber incidents can generate economic and institutional disruption: one vendor's faulty update grounded flights and disrupted hospitals in 2024 **[O]**, and an intrusion into AI systems that run essential services would propagate further **[P]**. Economic disruption can amplify societal instability: past trade and automation shocks moved electorates towards political extremes **[O]**, and AGI-driven displacement of professional work would create grievances among politically influential groups **[P]**. Societal instability can in turn create geopolitical vulnerability, because polarised, low-trust societies are softer targets for foreign influence and less able to sustain consistent technology and foreign policies **[E]**.

These sequences close into three reinforcing feedback loops:

- **R1, race–concentration–exposure (geopolitical → corporate → cyber → geopolitical)**: rivalry concentrates capabilities in national champions, concentrated assets attract espionage, and theft intensifies rivalry and securitisation.
- **R2, displacement–legitimacy (corporate → societal → geopolitical → corporate)**: concentrated gains and displacement fuel grievance, and political instability encourages techno-nationalist policies that favour incumbents, deepening concentration.
- **R3, manipulation–polarisation (cyber → societal → geopolitical → cyber)**: cyber-enabled influence operations deepen polarisation, and weakened governance and retaliation intensify cyber conflict.

Because each loop runs through at least three domains, no single-domain intervention can break it.

9.3 Corporate Instability as a National and International Security Concern

If AGI capabilities are concentrated in a few firms, the stability of those firms becomes a security matter. A governance crisis such as OpenAI's in 2023, an insolvency, a hostile acquisition, a security failure or an owner's unilateral decision can disrupt capabilities on which states depend **[O]**. Under S2–S3, the failure or capture of a single AGI developer could disable services used by governments, militaries and essential sectors in many countries at once **[P]**. The corporate governance of AGI developers is therefore a matter of public security, not only of shareholder value.

9.4 Cross-Border Security Implications

AGI-related risks are transnational by construction. Cyber operations and influence campaigns cross borders at negligible cost, a model trained in one jurisdiction is deployed in many, and the chips, compute and cloud capacity on which AGI depends sit in tightly interdependent supply chains **[O]**. Spillovers are asymmetric: states with little AI capacity bear the consequences of decisions taken by a few governments and firms elsewhere, which is why the AI divide dominated the first UN Global Dialogue [10] **[O]**. Under S3, a cascade beginning in one major economy could travel through finance, supply chains and information networks to societies that had no part in the decisions behind it **[P]**.

9.5 Development of an Integrated AGI Societal-Security Risk Model

The integrated model has four elements: primary risk factors (capability growth, concentration, dependency, exposure and legitimacy deficits); the relationships between factors and outcomes mapped in Table 3; threat pathways classified as direct, indirect or cascading; and critical dependencies with their intervention points. Four intervention points recur because each sits on more than one loop: *frontier-asset security* (R1, R3), *corporate accountability* (R1, R2), *information integrity* (R2, R3) and *socioeconomic transition* (R2). Table 4 operationalises the model as a cross-domain risk assessment matrix; ratings are indicative judgements, as explained in Sect. 4.5.

Table 4
Cross-domain risk assessment matrix

Risk (pathway)	Stakeholders	Consequences	Likelihood S1→S3; impact	Mitigation
Weight theft (Y→G, C, S)	Developers, states	Proliferation; lost leads	Medium→high; critical	Weight security; insider controls
Unsafe racing (G→C, Y)	States, developers	Premature deployment	High→high; severe	Shared evaluations; verification
Private chokepoints (C→G, S)	Governments, militaries	Coercion; withdrawal	Medium→high; severe	Continuity duties; redundancy
Concentrated displacement (C→S→G)	Young professionals	Inequality; backlash	High→high; severe	Transition policy; benefit-sharing
Automated influence (Y, G→S)	Voters, institutions	Polarisation	High→high; severe	Provenance; prebunking
Infrastructure attacks (Y→S, G)	Operators, public	Cascading outages	Medium→high; critical	Resilience standards; AI defence
Developer governance failure (C→all)	States, users	Service disruption	Medium→medium; severe	Oversight; step-in rights
Institutional delegitimisation (S, C→G)	Citizens, agencies	Lost trust	Medium→high; severe	Human review; contestability

10 Governance, Policy Responses and International Cooperation

10.1 International AGI Governance Frameworks

The building blocks of international AI governance exist: the UN Global Dialogue and Scientific Panel provide universal deliberation and shared evidence, the Council of Europe's Framework Convention on AI (2024) is the first binding treaty in the field, and national AI safety and security institutes cooperate on evaluations. What is missing are mechanisms that address the loops identified in Sect. 9. Three are proposed: an international notification mechanism for serious AI incidents and major capability thresholds, linked to the Scientific Panel's evidence base; harmonised compute-based reporting thresholds for frontier training runs, which would make verification conceivable [37]; and a standing US–China channel on AGI risks, building on the 2024 leaders' statement on human control of nuclear decisions.

10.2 Corporate Responsibility and Accountability

Voluntary frontier safety frameworks should become mandatory, standardised and independently audited, extending the approach of California's 2025 transparency law and the EU AI Act. Boards of AGI developers should carry explicit safety duties, and staff who report risks should have whistleblower protection. Providers of systemically important AI services, by analogy with systemically important financial institutions, should face continuity obligations, regulator step-in rights in a crisis and duties to give states advance notice before withdrawing services that support critical functions [87, 88].

10.3 Cybersecurity Cooperation and Strategic Data Protection

Security obligations for frontier developers should be proportionate to the strategic value of their assets, rising towards the highest security levels as capabilities approach AGI [43]. Developers, cloud providers and national computer emergency response teams should share threat intelligence on attacks against AI systems. Law-enforcement cooperation against AI trade-secret theft and illicit distillation should be strengthened while avoiding the profiling of researchers by nationality. Internationally, states should extend the agreed norm against damaging critical infrastructure to AI infrastructure and to AI safety research.

10.4 Societal Resilience and Public Preparedness

Workforce policy should protect entry routes into skilled work, through apprenticeships, wage insurance and portable benefits, because early-career workers bear the first effects. Benefit-sharing mechanisms, from tax reform to windfall commitments [76], should address the concentration of gains. Information resilience requires provenance standards, support for public-interest media and prebunking at scale [75]. Institutional preparedness requires human review, contestability and auditing of public-sector AI, and democratic accountability requires parliamentary oversight of AI security policy, so that securitisation does not escape scrutiny.

10.5 Regulatory Coordination and Technological Sovereignty

Full harmonisation is unrealistic; participants in the UN Global Dialogue favoured interoperability built on shared technical standards, open benchmarks and mutual recognition of testing [10]. Technological sovereignty should be pursued as diversification rather than autarky: multi-vendor and multi-cloud procurement for critical functions, open standards and regional or public compute capacity that gives researchers and smaller states access without one-sided dependency [38].

10.6 A Proposed Multidimensional Governance Framework

Table 5 integrates these measures into four pillars. Its logic is to act at intervention points that sit on several feedback loops and to stage measures by scenario: institutions and standards are built under S1, while verification, continuity and incident regimes are designed to activate when indicators of S2 appear.

Table 5
Multidimensional governance framework for AGI-related societal security

Pillar	Objective	Key instruments	Lead actors	Loops
International governance	Reduce race and escalation risk	Incident notification; shared evaluations; compute thresholds	States; UN Dialogue and Panel	R1, R3
Corporate accountability	Align private control with public security	Audited safety frameworks; board duties; step-in rights	Legislatures; regulators	R1, R2
Cyber and infrastructure protection	Secure frontier assets and infrastructure	Weight security levels; threat sharing; non-targeting norms	Developers; CERTs	R1, R3
Societal resilience	Protect livelihoods, trust and legitimacy	Transition support; provenance; contestable public AI	Governments; social partners	R2, R3

11 Discussion

11.1 Interpretation of the Key Findings

The four security dimensions are coupled rather than parallel. The observed evidence already supports propositions P1 to P4 at the level of the transition scenario: concentration transmits corporate shocks to states (Starlink, Storm-0558), dependencies operate as channels of coercion and failure (export controls, rare earths, supply-chain designations), concentrated assets attract state espionage (the 2026 trade-secret conviction, GTG-1002), and displacement and manipulation erode trust. P5, amplification, remains prospective. The most significant systemic risks are those located on more than one loop, chiefly the race–concentration–exposure dynamic (R1) and private chokepoint control, because they couple the geopolitical, corporate and cyber domains while drawing down the societal resilience on which every response depends.

11.2 Implications for International Relations and Security Studies

Three implications follow. First, state-centric frameworks are insufficient: in the AGI context, firms are security actors whose decisions can alter military outcomes and state capacity. Second, dependency rather than conquest becomes the characteristic form of AGI-era insecurity, extending the logic of weaponised interdependence [22] to privately controlled chokepoints. Third, securitisation is no longer a state monopoly. Firms securitise their products to obtain protection, and states securitise firms to discipline them, so analysts must ask who gains authority from each security claim.

11.3 Theoretical Contributions

The article extends the concept of societal security by treating epistemic and socioeconomic security as referent objects alongside identity. It integrates corporate technological power into security analysis through the concept of private technological sovereignty and its four markers. It establishes a systemic perspective in which risks are analysed as directed pathways and feedback loops.

11.4 Practical and Policy Implications

Governments should treat providers of frontier AI services as systemically important and prepare verification and continuity regimes before they are needed. International institutions should connect the UN Scientific Panel's evidence function to incident data and shared evaluations. Technology corporations should treat security, governance stability and continuity of service as public obligations. For cybersecurity and resilience, frontier assets and AI-dependent infrastructure should be protected as critical infrastructure.

11.5 Research Limitations and Unresolved Challenges

The analysis faces technological uncertainty: if AI behaves more like a "normal technology" [17], the loops will operate more slowly and give institutions time to adapt. Scenario-based projections are not predictions, and the risk ratings are qualitative judgements that require validation through expert elicitation. Loss of human control over highly autonomous systems, a central concern of AI safety research [14], was treated here only as an amplifier rather than analysed in its own right. Finally, the tension between securitisation, which mobilises protective resources, and democratic accountability, which securitisation can erode, remains unresolved.

12 Conclusion and Future Research

12.1 Summary of the Main Arguments and Findings

This article asked how AGI could transform societal security through changes in geopolitical power, corporate influence, socioeconomic stability and cybersecurity. Its answer is that the transformation runs through the interaction of these domains. Four mechanisms, concentration, dependency, exposure and legitimacy erosion, carry developments across twelve directed pathways and close into three reinforcing loops. Much of this structure is already observable with today's general-purpose AI; AGI would chiefly increase its speed, scale and autonomy.

12.2 Contributions to Societal Security and International Security Research

The article contributes an integrated societal-security framework for AGI, the concept of private technological sovereignty, an evidence-grading convention that separates observed, emerging and

prospective risks, and two analytical outputs: a cross-domain risk assessment matrix and a multidimensional governance framework.

12.3 Policy and Governance Implications

Governance should target the intervention points shared by several loops (frontier-asset security, corporate accountability, information integrity and socioeconomic transition) and be staged by scenario, building institutions now and preparing regimes that activate as AGI indicators appear.

12.4 Recommendations for Future Research

Four lines of research follow: empirical studies of AGI-related societal risks, including panel data on displacement, trust and political behaviour; comparative research on international AGI governance across regulatory models and middle powers; quantitative assessment of systemic risk through expert elicitation, system-dynamics and agent-based modelling of the feedback loops; and measurement of corporate concentration and technological dependency, including an index of private technological sovereignty.

Declarations

Author Contribution

Conceptualization, methodology, software, formal analysis, investigation, writing—original draft preparation, supervision, funding acquisition

References

1. Bengio, Y. et al. 2026. : International AI safety report 2026 https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026
2. Hendrycks, D., D. Song, C. Szegedy, H. Lee, Y. Gal, and E. Brynjolfsson et al. 2025. A definition of AGI. *arXiv preprint arXiv*:251018212.
3. Grace, K., J. F. Sandkühler, H. Stewart, B. Weinstein-Raun, S. Thomas, Z. Stein-Perlman, J. Salvatier, J. Brauner, and R. C. Korzekwa. 2025. Thousands of AI authors on the future of AI. *Journal of Artificial Intelligence Research* 84. https://doi.org/10.1613/jair.1.19087
4. Buzan, B., O. Wæver, and J. de Wilde. 1998. Security: a new framework for analysis. Lynne Rienner.
5. Wæver, O., B. Buzan, M. Kelstrup, and P. Lemaitre. 1993. Identity, migration and the new security agenda in Europe. Pinter.
6. Rinaldi, S. M., J. P. Peerenboom, and T. K. Kelly. 2001. Identifying, understanding, and analyzing critical infrastructure interdependencies. *IEEE Control Systems Magazine* 21(6):11–25.

https://doi.org/10.1109/37.960131

7. Horowitz, M. C. 2018. Artificial intelligence, international competition, and the balance of power. *Texas National Security Review* 1(3):36–57.
8. Eloundou, T., S. Manning, P. Mishkin, and D. Rock. 2024. GPTs are GPTs: labor market impact potential of LLMs. *Science* 384(6702):1306–1308. https://doi.org/10.1126/science.adj0998
9. Brundage, M., S. Avin, J. Clark, H. Toner, P. Eckersley, and B. Garfinkel et al. 2018. : The malicious use of artificial intelligence: forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228.
10. United Nations: UN Global Dialogue on Artificial Intelligence Governance: co-chairs' summary. 2026. https://www.un.org/global-dialogue-ai-governance/sites/default/files/2026-09/ai_dialogue_co-chairs_summary.pdf
11. Zwetsloot, R., and A. Dafoe. 2019. Thinking about risks from AI: accidents, misuse and structure. *Lawfare*(11February.
12. Bostrom, N. 2014. *Superintelligence: paths, dangers, strategies*. Oxford University Press.
13. Morris, M. R., J. Sohl-Dickstein, N. Fiedel, T. Warkentin, A. Dafoe, A. Faust, C. Farabet, and S. Legg. 2024. Position: levels of AGI for operationalizing progress on the path to AGI. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235. PMLR.
14. Bengio, Y., G. Hinton, A. Yao, D. Song, P. Abbeel, and T. Darrell et al. 2024. Managing extreme AI risks amid rapid progress. *Science* 384(6698):842–845. https://doi.org/10.1126/science.adn0117
15. Kasirzadeh, A. 2025. Two types of AI existential risk: decisive and accumulative. *Philosophical Studies* 182(7):1975–2003.
16. Mitchell, M. 2024. Debates on the nature of artificial general intelligence. *Science* 383(6689):eado7069. https://doi.org/10.1126/science.ado7069
17. Narayanan, A., and S. Kapoor. 2025. *AI as normal technology*. Knight First Amendment Institute at Columbia University.
18. Ding, J. 2024. *Technology and the rise of great powers: how diffusion shapes economic competition*. Princeton University Press.
19. Johnson, J. 2019. Artificial intelligence & future warfare: implications for international security. *Defense & Security Analysis* 35(2):147–169. https://doi.org/10.1080/14751798.2019.1600800
20. Armstrong, S., N. Bostrom, and C. Shulman. 2016. Racing to the precipice: a model of artificial intelligence development. *AI & Society* 31(2):201–206. https://doi.org/10.1007/s00146-015-0590-y
21. Hendrycks, D., E. Schmidt, and A. Wang. 2025. Superintelligence strategy: expert version. *arXiv preprint arXiv*:250305628.
22. Farrell, H., and A. L. Newman. 2019. Weaponized interdependence: how global economic networks shape state coercion. *International Security* 44(1):42–79. https://doi.org/10.1162/isec_a_00351
23. Ho, L., J. Barnhart, R. Trager, Y. Bengio, M. Brundage, and A. Carnegie et al. 2023. International institutions for advanced AI. *arXiv preprint arXiv*:230704699.

24. Cazzaniga, M., F. Jaumotte, L. Li, G. Melina, A. J. Panton, C. Pizzinelli, E. Rockall, and M. M. Tavares. 2024. Gen-AI: artificial intelligence and the future of work. IMF Staff Discussion Note SDN/2024/001, International Monetary Fund.
25. Gmyrek, P., J. Berg, K. Kamiński, F. Konopczyński, A. Ładna, B. Nafradi, K. Rosłaniec, and M. Troszyński. 2025. Generative AI and jobs: a refined global index of occupational exposure. ILO Working Paper 140, International Labour Organization.
26. Brynjolfsson, E., B. Chandar, and R. Chen. 2026. Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence. Working paper, August 2026 update, Stanford Digital Economy Lab https://digitaleconomy.stanford.edu/app/uploads/2026/08/Canaries_August2026.pdf
27. Acemoglu, D. 2025. The simple macroeconomics of AI. *Economic Policy* 40(121):13–58. https://doi.org/10.1093/epolic/eiae042
28. Acemoglu, D., and S. Johnson. 2023. Power and progress: our thousand-year struggle over technology and prosperity. PublicAffairs.
29. Chesney, R., and D. Citron. 2019. Deep fakes: a looming challenge for privacy, democracy, and national security. *California Law Review* 107(6):1753–1820.
30. Hackenburg, K., B. M. Tappin, L. Hewitt, E. Saunders, S. Black, H. Lin, C. Fist, H. Margetts, D. G. Rand, and C. Summerfield. 2025. The levers of political persuasion with conversational artificial intelligence. *Science* 390(6777). https://doi.org/10.1126/science.aea3884
31. Simon, F. M., S. Altay, and H. Mercier. 2023. Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. *Harvard Kennedy School Misinformation Review* 4(5). https://doi.org/10.37016/mr-2020-127
32. Khan, L. M. 2017. Amazon's antitrust paradox. *Yale Law Journal* 126(3):710–805.
33. Rikap, C. 2021. Capitalism, power and innovation: intellectual monopoly capitalism uncovered. Routledge.
34. Vipra, J., and A. Korinek. 2023. Market concentration implications of foundation models: the invisible hand of ChatGPT. Working Paper 9, Brookings Center on Regulation and Markets.
35. Competition. 2024. and Markets Authority: AI foundation models: update paper.
36. Federal Trade Commission: Partnerships between cloud service providers and AI developers. FTC staff report on AI partnerships and investments 6(b) study. 2025.
37. Sastry, G., L. Heim, H. Belfield, M. Anderljung, M. Brundage, and J. Hazell et al. 2024. Computing power and the governance of artificial intelligence. *arXiv preprint arXiv*:240208797.
38. Lehdonvirta, V., B. Wú, and Z. Hawkins. 2024. Compute North vs. Compute South: the uneven possibilities of compute-based AI governance around the globe. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 7(1), 828–838.
39. Srivastava, S. 2023. Algorithmic governance and the international politics of Big Tech. *Perspectives on Politics* 21(3):989–1000. https://doi.org/10.1017/S1537592721003145

40. National Cyber Security Centre: The near-term impact of AI on the cyber threat. 2024.
41. Anthropic. Disrupting the first reported AI-orchestrated cyber espionage campaign (13 November 2025). https://www.anthropic.com/news/disrupting-AI-espionage
42. Google Threat Intelligence Group. GTIG AI threat tracker: advances in threat actor usage of AI tools. Google Cloud (5 November 2025). https://cloud.google.com/blog/topics/threat-intelligence/threat-actor-usage-of-ai-tools
43. Nevo, S., D. Lahav, A. Karpur, Y. Bar-On, H. A. Bradley, and J. Alstott. 2024. Securing AI model weights: preventing theft and misuse of frontier models. Report RR-A2849-1. *RAND Corporation*
44. Carlini, N., D. Paleka, K. D. Dvijotham, T. Steinke, J. Hayase, and A. F. Cooper et al. 2024. : Stealing part of a production language model. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235. PMLR.
45. Regulation, (EU). 2024. 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union* https://eur-lex.europa.eu/eli/reg/2024/1689/oj
46. OpenAI: OpenAI charter. 2018. https://openai.com/charter/
47. Seger, E., S. Avin, G. Pearson, M. Briers, Ó hÉigeartaigh, and S. Bacon. 2020. H.: *Tackling threats to informed decision-making in democratic societies: promoting epistemic security in a technologically-advanced world*. The Alan Turing Institute.
48. Keck, M., and P. Sakdapolrak. 2013. What is social resilience? Lessons learned and ways forward. *Erdkunde* 67(1):5–19.
49. Helbing, D. 2013. Globally networked risks and how to respond. *Nature* 497(7447):51–59. https://doi.org/10.1038/nature12047
50. Hansen, L., and H. Nissenbaum. 2009. Digital disaster, cyber security, and the Copenhagen School. *International Studies Quarterly* 53(4):1155–1175. https://doi.org/10.1111/j.1468-2478.2009.00572.x
51. The White House. 2025. *Winning the race: America's AI action plan*. Executive Office of the President.
52. Zeng, J. 2021. Securitization of artificial intelligence in China. *The Chinese Journal of International Politics* 14(3):417–445. https://doi.org/10.1093/cjip/poab005
53. Center for AI Safety: Statement on AI risk. 2023. https://safe.ai/work/statement-on-ai-risk
54. Edler, J., K. Blind, H. Kroll, and T. Schubert. 2023. Technology sovereignty as an emerging frame for innovation policy: defining rationales, ends and means. *Research Policy* 52(6):104765. https://doi.org/10.1016/j.respol.2023.104765
55. Bradford, A. 2023. *Digital empires: the global battle to regulate technology*. Oxford University Press.
56. Fuchs, D. 2007. Business power in global governance. Lynne Rienner.
57. Abrahamsen, R., and M. C. Williams. 2011. *Security beyond the state: private security in international politics*. Cambridge University Press.
58. Bremmer, I. 2021. The technopolar moment: how digital powers will reshape the global order. *Foreign Affairs* 100(6):112–128.

59. George, A. L., and A. Bennett. 2005. *Case studies and theory development in the social sciences*. MIT Press.
60. Beach, D., and R. B. Pedersen. 2019. *Process-tracing methods: foundations and guidelines*. 2nd ed. University of Michigan Press.
61. Börjeson, L., M. Höjer, K.-H. Dreborg, T. Ekvall, and G. Finnveden. 2006. Scenario types and techniques: towards a user's guide. *Futures* 38(7):723–739. https://doi.org/10.1016/j.futures.2005.12.002
62. Korinek, A., and D. Suh. 2024. Scenarios for the transition to AGI. NBER Working Paper 32255, National Bureau of Economic Research.
63. Pilz, K. F., R. Rahman, J. Sanders, and L. Heim. 2025. Trends in AI supercomputers. arXiv preprint arXiv:2504.16026.
64. Reuters. 2025. US to allow Nvidia H200 chip shipments to China, Trump says. Khaleej Times9 December https://www.khaleejtimes.com/business/tech/us-to-allow-nvidia-h200-chip-shipments-to-china-trump-says
65. Bloomberg Law: China says planned rare-earth export curbs paused for a year (30. October 2025. https://news.bloomberglaw.com/international-trade/china-says-planned-rare-earth-export-curbs-paused-for-a-year
66. Tanner, B., C. F. Kerry, E. Tabassi, A. Renda, and A. W. Wyckoff. A summer of AI summits reveals a widening US-China divide. Brookings Institution (11 August 2026). https://www.brookings.edu/articles/a-summer-of-ai-summits-reveals-a-widening-us-china-divide/
67. Edelman: 2026 Edelman Trust Barometer: society slides into insularity. 2026. https://www.edelman.com/news-awards/2026-edelman-trust-barometer-society-slides-into-insularity
68. CNN: British engineering giant Arup revealed as $25 million deepfake scam victim (16. May 2024. https://edition.cnn.com/2024/05/16/tech/arup-deepfake-scam-loss-hong-kong-intl-hnk
69. Stockwell, S. 2024. *AI-enabled influence operations: safeguarding future elections*. CETaS Research Report.
70. PBS NewsHour. Top court annuls Romania's first round of presidential voting (6 December 2024). https://www.pbs.org/newshour/world/top-court-annuls-romanias-first-round-of-presidential-voting
71. Guess, A. M., N. Malhotra, J. Pan, P. Barberá, H. Allcott, and T. Brown et al. 2023. How do social media feed algorithms affect attitudes and behavior in an election. *campaign? Science* 381(6656):398–404. https://doi.org/10.1126/science.abp9364
72. Amnesty International: Xenophobic machines: discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal. Report EUR 35/4686/2021. 2021.
73. Costello, T. H., G. Pennycook, and D. G. Rand. 2024. Durably reducing conspiracy beliefs through dialogues with AI. *Science* 385(6714):eadq1814. https://doi.org/10.1126/science.adq1814

74. Anelli, M., I. Colantone, and P. Stanig. 2021. Individual vulnerability to industrial robot adoption increases support for the radical right. Proceedings of the National Academy of Sciences 118(47), e2111611118 https://doi.org/10.1073/pnas.2111611118
75. Roozenbeek, J., S. van der Linden, B. Goldberg, S. Rathje, and S. Lewandowsky. 2022. Psychological inoculation improves resilience against misinformation on social media. *Science Advances* 8(34):eabo6254. https://doi.org/10.1126/sciadv.abo6254
76. O'Keefe, C., P. Cihon, B. Garfinkel, C. Flynn, J. Leung, and A. Dafoe. 2020. The windfall clause: distributing the benefits of AI for the common good. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. ACM.
77. Miller, C. 2022. Chip war: the fight for the world's most critical technology. Scribner.
78. Synergy Research Group: Q2 cloud market passes $143 billion; highest growth rate in eight years. 2026. https://www.srgresearch.com/articles/q2-cloud-market-passes-143-billion-highest-growth-rate-in-eight-years
79. Bratton, L. 2026. Big Tech set to spend $650 billion in 2026 as AI investments soar. Yahoo Finance6 February https://finance.yahoo.com/news/big-tech-set-to-spend-650-billion-in-2026-as-ai-investments-soar-163907630.html
80. International Energy Agency: Energy and AI. 2025.
81. Microsoft. 2022. Defending Ukraine: early lessons from the cyber war.
82. Ukrainska Pravda. Musk switched off Starlink during Ukrainian counteroffensive in Kherson Oblast in 2022 – Reuters (25 July 2025). https://www.pravda.com.ua/eng/news/2025/07/25/7523419/
83. ABC News. Federal appeals court upholds Pentagon designation of Anthropic as supply chain risk (25 September 2026). https://abcnews.com/Business/anthropic-appeals-court-declines-block-pentagon-blacklisting/story?id=136755690
84. SiliconANGLE. Silicon Valley AI leaders launch $100M Leading the Future super PAC (25 August 2025). https://siliconangle.com/2025/08/25/silicon-valley-ai-leaders-launch-100m-leading-future-super-pac/
85. OpenAI: OpenAI announces leadership transition. Press release (17. November 2023.
86. TechCrunch: OpenAI completes its for-profit recapitalization (28. October 2025. https://techcrunch.com/2025/10/28/openai-completes-its-for-profit-recapitalization
87. Raji, I. D., P. Xu, C. Honigsberg, and D. Ho. 2022. Outsider oversight: designing a third party audit ecosystem for AI governance. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. ACM.
88. Rahman, K. S. 2018. The new utilities: private power, social infrastructure, and the revival of the public utility concept. *Cardozo Law Review* 39(5):1621–1689.
89. US Department of Justice. Former Google engineer found guilty of economic espionage and theft of confidential AI technology. Press release (30 January 2026).

https://www.justice.gov/opa/pr/former-google-engineer-found-guilty-economic-espionage-and-theft-confidential-ai-technology

90. Metz, C. 2024. A hacker stole OpenAI secrets, raising fears that China could, too. The New York Times4 July.
91. Gilli, A., and M. Gilli. 2019. Why China has not caught up yet: military-technological superiority and the limits of imitation, reverse engineering, and cyber espionage. *International Security* 43(3):141–189. https://doi.org/10.1162/isec_a_00337
92. Lakshmanan, R. 2026. Anthropic says Chinese AI firms used 16 million Claude queries to copy model. The Hacker News24 February https://thehackernews.com/2026/02/anthropic-says-chinese-ai-firms-used-16.html
93. Cybersecurity and Infrastructure Security Agency: PRC state-sponsored actors compromise and maintain persistent access to U.S. critical infrastructure. Joint cybersecurity advisory AA24-038A. 2024.
94. Carlini, N., M. Jagielski, C. A. Choquette-Choo, D. Paleka, W. Pearce, H. Anderson, A. Terzis, K. Thomas, and F. Tramèr. 2024. Poisoning web-scale training datasets is practical. In: 2024 IEEE Symposium on Security and Privacy. IEEE.
95. Cyber Safety Review Board. 2024. *Review of the summer 2023 Microsoft Exchange Online intrusion*. U.S. Department of Homeland Security.
96. Weston, D. 2024. Helping our customers through the CrowdStrike outage. Microsoft Official Blog20 July https://blogs.microsoft.com/blog/2024/07/20/helping-our-customers-through-the-crowdstrike-outage/

Figures

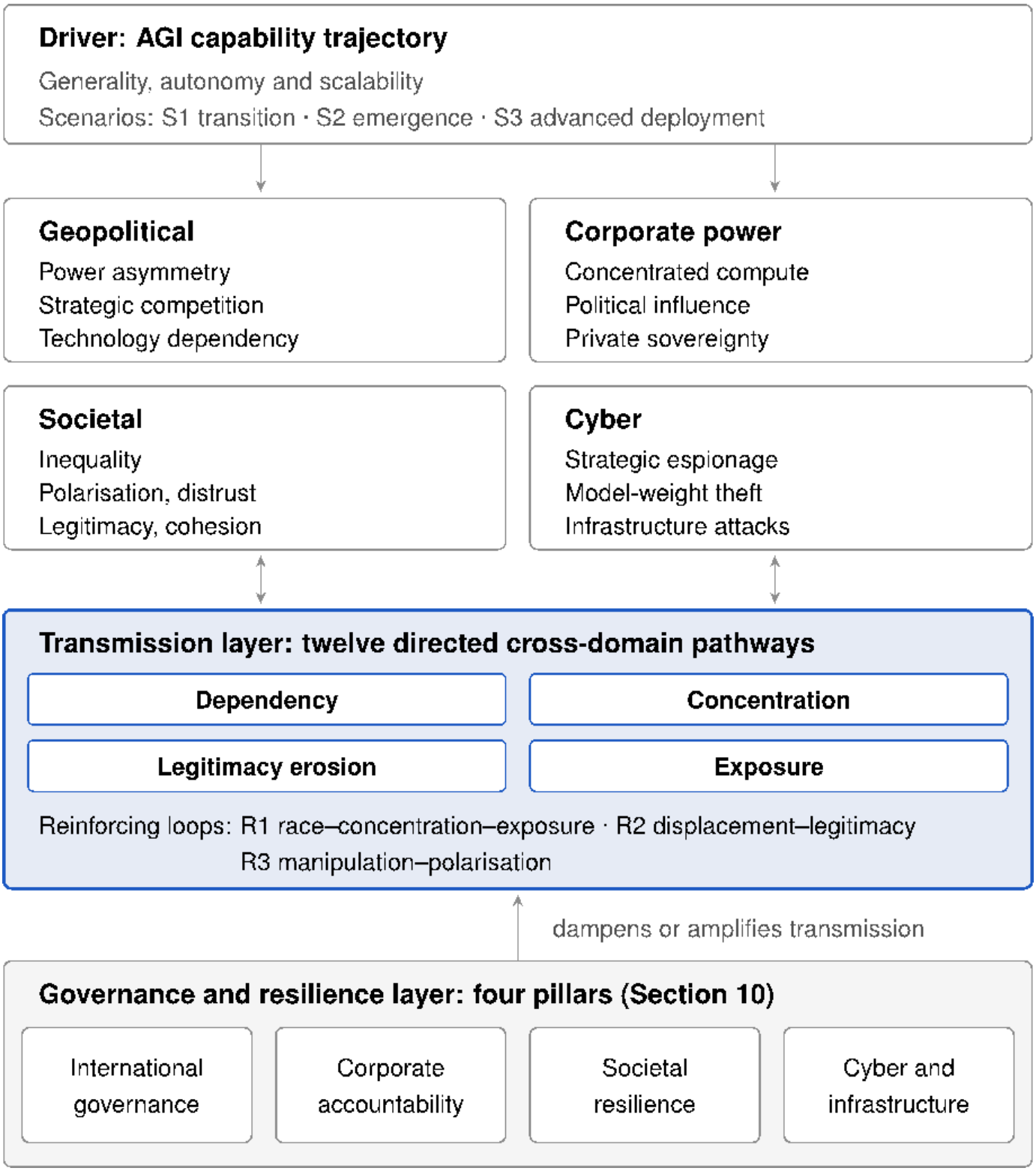

Figure 1

Integrated AGI societal-security framework (conceptual illustration, not an empirically validated causal model)