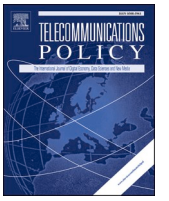
Telecommunications Policy 49 (2025) 102907

Contents lists available at ScienceDirect

Telecommunications Policy

journal homepage: www.elsevier.com/locate/telpol



Into the next generation of digital protection: AI resiliency as a public responsibility

Eli Noam [1]

Columbia University, Columbia Institute for Tele-Information, 645 W. 130th St., Geffen 694, New York, NY, 10027, USA

ARTICLE INFO

Keywords:
Artificial intelligence
AI resilience
AI regulation
Network security
Cybersecurity
Human-AI collaboration
Self-regulation
Transparency
Interoperability
Post-human

ABSTRACT

Even as the reliability of networks has risen, their control mechanisms up in the hierarchy of digital activities have become more vulnerable. Artificial intelligence algorithms are increasingly embedded in infrastructure and economic systems and their resiliency is essential for social and economic stability. This has led to widespread dystopic fears and defensive regulations, ignoring the considerable positives of AI-enhanced activities and institutions. AI resiliency problems include hardware failures, natural calamities, human error, software defects, and external attacks. AI networks of AI networks have emerged with high interdependence and complexity. Operations are often non-transparent 'black boxes' operating at lightning speeds, and hard to oversee or fix by humans. Most likely is a control of AI by other AI. This raises the question of human responsibility. The article examines various responses, including technology tools, managerial actions, self-regulation, and a role for government. The latter include rules evolving with technology and applications in a dynamic common law approach for liability, transparency, performance, market structure, interoperation, and more. Needed are principles for a 'shared intelligence' of humans with AI, with clear protocols for human overrides of AI. All this raises a new agenda for policymakers, managers, and researchers.

1. Introduction

As digital activities proliferate, societies become dependent on the reliability of the underlying infrastructures. Yet the attention given to such resiliency by academic research has not been strong or well-connected to regulatory policy. Such absence is particularly noteworthy as a new generation of network resiliency challenges is upon us, that of artificial intelligence functionalities. The objective of this article is to identify the problems of AI resiliency, to explore remedies and research implications, and to define the remaining role for humans in a complex system of interoperating AI systems. It analyzes AI's threats to resiliency, presents the potential managerial, governmental, and technological responses, and concludes that a practical remedy against bad-actor AI is good-actor AI, coupled with standards for operational redundancy and recovery and a human-controlled "kill-switch."

There has been huge progress in improving network reliability through the substitution of monopoly by a multiple-network system

This article is part of a special issue entitled: Resilience published in Telecommunications Policy.

[1] The author thanks Erik Bohlin for his leadership on the subject of resilience as well as for his comments on this article, and to the team at Ivey Business School and at the Lawrence National Center for Policy and Management, including Romel Mostafa. He also thanks three anonymous referees for their insights and helpful comments. Special thanks for their invaluable help go to Jason Buckweitz and Corey Spencer of the Columbia University Institute for Tele-Information. This article is part of a larger AI project, with Columbia's partner-affiliates Dumont Mediengruppe, Club de Investigation Tecnologica, Granahan McCourt Capital, National Broadband Ireland, Rise Global Foundation, and University College Dublin.

https://doi.org/10.1016/j.telpol.2025.102907
Received 26 August 2024; Received in revised form 9 December 2024; Accepted 13 January 2025
Available online 30 January 2025
0308-5961/

with parallel transmission capacities. The underlying technology of packet switching was specifically designed to revise its routing when failures occur. The remaining problems seem manageable (Serentschy, 2024). Yes, networks might experience brief breakdowns, as they did with AT&T's wireless operations in 2024, or when wildfires, earthquakes, or floods hit. In long-distance transmission, sub-sea cables are a major military target, and electro-magnetic pulse attacks can disable wireless and satellites. But on the whole, basic transmission has been more secure than ever. And yet, a much larger resiliency problem has emerged. With transmission now cheap and plentiful, it is also used for vastly more purposes, increasing societal dependence. With new types of failures and attacks, the next generation of resiliency issues will dwarf those of the past.

A hierarchy of resiliency challenges can be identified.

- Basic layer: transmission networks. They have been the focus of past resiliency research.
- Intermediate layer - platforms, in particular cloud platforms [2]: Here, outages have big footprints, exemplified by the failures of one of Google's regional clouds (Gienow, 2023) and of Cloudflare. The business of big customers was paralyzed for many hours.
- A further intermediate layer - applications: Examples are navigation service like Waze or a streaming service such as Netflix. These services often encounter problems.
- And there is a still higher layer - Artificial Intelligence (AI), as a software control mechanism for networks, platforms, applications, and further AI operations. [3]

Each of these network layers has its own operations management function, and these have become increasingly 'smart' in rapidly replicating human decision processes such as being able to understand and respond to various sensory inputs, to analyze data, make recommendations, and act. These are functions subsumed under the broad category of AI. Some AI systems control entire systems like telecom networks; others manage energy, transportation, financial services, law enforcement, etc. Beyond system-specific AI there must also be integrative and coordinating intelligent control systems for the manifold of AI systems to interact effectively. This system of system can be described as the AI of AI. It is the most recent manifestation of integrative tendencies in digital systems, identified years ago by the author as a "network of networks" with significant implications for stability, interoperation, pricing, and standards (Noam, 1988, 1994; 2001a). The AI of AIs is something that already exists in its early stages, and it will become central to the functioning of critical functions of society and economy. If such a system breaks down, there will be a society-wide disruption. Nothing will function, and sorting things out and fixing them will be a task that might take days – an eternity in the digital age.

This AI of AI system will become a desirable target for disruption, and preventing it becomes a priority. More innocently, things can simply go wrong, and the complexity of the moving parts will prevent rapid restoration. And therefore, the resiliency of the AI becomes a central issue of concern.

This fear does not stand in isolation. AI has already received much negative coverage in the idea chain of academia, media, and politicians (Wach et al., 2023). The general tone has been one of alarm. Critics have raised many issues: inequality, bias, jobs, privacy, transparency, plagiarism, disinformation, fraud, accidental wars and economic crises. In that view, AI systems will become so powerful and autonomous that humans become subordinated.

Many of these problems with AI are real, even if they are being overblown. Being over-cautious about losing control is part of the evolutionary survival mechanism of our species. We used to fear predator animals. Once we learned to vanquish them, we fought imagined ones, such as dragons and cyclops. And then, we worried about artificial creatures displacing us: Frankenstein's monster; the Golem of Prague; HAL in *2001: A Space Odyssey*; the Terminator, and so on.

Amongst this pessimism one should also look for the positives. AI can be our friend (Noam, 2023). On the individual level, for example, AI offerings could become personally controlled and empowering, as our personal assistant, toiling 24/7 as our alter ego, assisting us as an agent, gatekeeper, warning system, information gatherer, shopper, and much more.

But optimism does not mean that the problems are not real. It has been said that AI will soon outstrip natural intelligence. This gets headlines but we should be skeptical, because AI's intelligence will often trail that of humans. That, however, is not good news, because an imperfect AI system is much worse than perfect one. Critics of AI often cannot make up their minds whether they are more concerned about AI being too smart – leaving humanity behind—or not smart enough and making laughable errors known as hallucinations that might affect people negatively. What is certain is that AI is lightning-fast and often a non-transparent "black box" and it is hard to figure out how it came up with its conclusion. In consequence, humanity could be subject to near-instant decision-making by a flawed black box – a bad combination. Problems and disruptions are inevitable, impacting resiliency in fundamental ways.

Most critics of AI focus on what harms AI might cause but only touch in passing the problem of harm being done to AI systems, and the consequences of various forms of AI failure. These are caused in several ways, exemplified below (Saeed et al., 2023). (Note that some of the older examples are those of broader cyber-security rather than AI-security, a sub-category of the concern of this article.)

1. Breakdowns: They include hardware issues or power failures. For example, Cloudflare is a major content delivery, web-services, and cyber-security company. It experienced a significant outage in 2023 caused mostly by an outside supplier of a data center whose power supply failed and was not monitored or restored properly. As a result, Cloudflare was out for two days (Prince, 2023).

[2] An analysis of the resiliency issues of the intermediate layers of platforms and applications is outside the scope of this article.

[3] This article attempts to capture the interaction between AI and resiliency, but not to provide a new, general definition of AI. OECD has among others provided extensive work on internationally acceptable and recognized definitions of AI – see for instance https://www.oecd.org/en/topics/artificial-intelligence.html.

Similarly, Google's Cloud Platform in the region Europe-West 9 experienced an outage in 2023 because of a broken water pump that caused havoc.

2. Natural calamities, such as floods, wildfires, earthquakes, and hurricanes can create havoc (Cutter et al., 2013).
3. Human error: When Cloudflare went down, a decision early in the outage to delay the gradual restart procedures led to a "thundering herd" problem when many services were allowed to come online at the same time, overwhelming other connected systems and causing them to fail, too (Prince, 2023). Similarly, the cyber-security platform company CrowdStrike failed to test its software update, overlooking an internal logic error. This then affected Microsoft operating system Windows, and that, in turn, crashed Windows computers worldwide (Kerner, 2024). The secondary effects may be huge, exemplified by how Delta Airlines was heavily hit by the outage, more so than its competitors, grounding flights for 5 days, with more than 5000 flights cancelled. The company had to manually reset over 40,000 servers, and the outage cost Delta $500 million (Breslow, 2024). A 2017 human error by an "authorized team member" caused Amazon's Simple Storage Service (S3) to go down for about 3 h. The cost to the company alone was roughly $150 million. That figure does not include the consequent damages due to cloud interoperability to companies that rely on S3 for their operations, for example the payment processing firm Venmo (Hersher, 2017).
4. Software defects: Software breakdown of AI systems due to strange and unanticipated workings of the software may cause considerable damage. AT&T suffered a massive outage in February 2024 that was caused by an error in coding, where a worker misapplied software to equipment during a network upgrade. The error took AT&T service down for millions of users around the country. The outage, which lasted less than a day, cost the company $140 million (Dano, 2024). In 2015, a "software glitch" stopped trading on the New York Stock Exchange for 4 h, halting the buying and selling of stocks whose trading volume would have been about $200 billion.
5. Outside offensive actions: Hostile governments, criminals, and mischief makers can cause havoc (Aggarwal, 2023; Akhtar & Mian, 2018; Chakraborty et al., 2021; Gongye et al., 2020). China-affiliated hackers attacked electric utilities, ports, and pipelines, and Russia-related hackers targeted the European rail system. Either Chinese or Russian operators hit the important supply-chain system SolarWinds in 2020. In 2023, the messaging system of OpenAI – the maker of ChatGPT – was hacked and the intruder gained access to specifics about the company's AI designs (Metz, 2024). Microsoft's system was compromised by Chinese hackers who gained access to federal government networks in the United States. These attackers increasingly use AI itself – weaponizing artificial intelligence to harm humans and their institutions, making use of the fact that to bring down a network, a cloud, or a large app system, it is enough to compromise the AI that controls its operations.

2. Fundamental question of AI

Government institutions around the world have taken initiatives in AI risk analysis (EU General Secretariat of the Council, 2021; UK Department for Science, Innovation & Technology, 2024). This includes, in alphabetical order: China's Artificial Intelligence Industry Alliance (Luong & Arnold, 2021); the European Union Agency for Cyber Security (ENISA) and its Cyber Security Strategy (European Union Agency for Cybersecurity, 2020; European Union Agency for Cybersecurity, 2023), as well as the Council of Europe (2024); Japan's AI Safety Institute (METI, 2024); the Korean Ministry of Science and ICT Strategy (MSIT, 2024); the OECD iLibrary (2024); the UK's Central AI Risk Function (CAIRF), the AI Safety Institutes and the UK National Cyber Security Centre (UK, 2023); the US AI Safety Institute (USAIST, 2023, November 2); the US Cybersecurity & Infrastructure Security Agency (2023); and the National Institute of Standards and Technology (NIST) with its Artificial Intelligence Risk Management Framework (NIST, 2023, 2024).

The list can go on: Think tank and consulting reports were issued by the Alan Turing Institute (Janjeva et al., 2023; the Carnegie Endowment for International Peace (Pendleton et al., 2024); CEPS (Pupillo et al., 2021); and McKinsey (Boehm et al., 2023). and others (McKendrick & Thurai, 2022). Private sector analyses of risk factors in AI were produced by major companies, user communities, and others (Apruzzese et al., 2023; Anthropic, 2023a, 2023b; AWS, 2024; Brundage et al., 2018; Dawson, 2023; Hansen & Venables, 2023; Marshall et al., 2024; Columbus, 2023; NVIDIA, 2024; OpenAI, 2023; OWASP, 2024; Rosiek, 2024; Shevlane et al., 2023.)

Vulnerabilities of AI to intrusions were identified and documented in articles and case studies in a UK government report (UK, 2024.) They include:

- Lack of robust security architecture – enabling unauthorized access and code injection (Bécue et al., 2021).
- Inadequate threat modelling - Insufficient identification of potential threats, vulnerabilities, and attack vectors. (Bradley, 2020; European Union Agency for Cybersecurity, 2020; Li et al., 2022).
- Insufficient data privacy safeguards (Majeed & Hwang, 2023).
- Insecure authentication and authorization (European Union Agency for Cybersecurity, 2020; 2023; Mirsky et al., 2023).
- Use of unreliable sources to label data (Chiang & Gairola, 2018).
- Bias injection into machine learning models (Ferrer et al., 2021).
- Unsecured data handling in AI systems (Silva & Alahakoon, 2022).
- Inadequate input validation and sanitization (Hu et al., 2021; Vassilev, 2024).
- Insecure API endpoints (Boulemtafes et al., 2020; Carlo et al., 2023).
- Infrastructure security (Silva & Alahakoon, 2022).
- Configuration vulnerabilities in cloud services (Boulemtafes et al., 2020).
- Insider threats (Mirsky et al., 2023).

- Malicious script insertion, such as transforming Bing Chat into a tool that searches for and steals personal data. (Greshake et al., 2023).
- Production of bogus reality, directing open-source, pre-trained large language model (LLM) to produce a bogus reality (Huang et al., 2023).
- Backdoor attack on deep learning models in mobile Apps (Li et al., 2022).

These articles and case studies mostly discuss the problem of AI resiliency as an advanced cyber-security issue. And indeed, protecting AI against cyber-attacks is a new chapter in an un-ending resiliency problem.

But the problem goes far beyond resiliency. There is the fundamental issue of allocation of control over complex systems in a future society (Quaquebeke & Gerpott, 2023; Usmani et al., 2023). Will this control be with users? With private companies? With governments? And on a still larger scale, will it be with humans, or with machines?

The question is how much authority humanity can safely cede to software systems (Benzinger et al., 2023; Lamanna & Byrne, 2018; Schneiderman, 2022). While declarations that humanity must remain in control are routine, making that happen is easier said than done. What exactly are the practical steps to assert control? And what are the basic principles that would govern humans' role?

The reflexive response to AI vs. humanity is to demand full control by real people. This humanist approach, while ultimately axiomatic ("believe the human") can be rationalized by pointing to the human capabilities of intuition and empathy which machines do not possess. In that narrative, machines err and hallucinate while humans maintain good judgement. However, humanity should not be over-romanticized. One person's intuition is another person's bias. To err, after all, is human. People make mistakes all the time, and since they do, should not an unemotional, unflappable, sober, and rational machine with reams of information, have the last word, just like a car's safety features should automatically activate braking when the driver comes too close to an obstacle? While nobody would advocate for an AI to have decision power on launching nuclear missiles, the question here is about more practical scenarios, such as restoring credit cards, rerouting a tanker, or injecting liquidity into a bank. To judge the quality of judgement of an AI system versus that of humans one needs to know what the "correct" decision is. AI and humans may well come to different conclusions, and where the stakes are not huge and irreversible, a mechanism of "dispute resolution" must be put in place in which the AI explains its recommendation to its human partner/boss but can be overridden (Gesser et al., 2022).

The emerging reality is one of *shared intelligence* ("SI"). An example is a navigation system which provides a highly specific behavioral guide. When it is overridden by the human driver, the AI provides tactful corrections, but the human makes the decision on the destination, time of departure, mode of transportation, and whether to follow the AI at all. Shared-intelligence systems abound, such as home climate controls; safety features in cars; pre-screening of resumes by human resource (HR) departments; pre-screening of skin moles by dermatologists; etc. Humans are always in the loop, and the more significant a decision, the more human-controlled the process should be. Resiliency would similarly be a divided task, with humans at the front end – development, parameter setting, and systems integration—as well as the top end – ultimate decision control—and AI occupying things in-between, like layers of mid-level management.

To keep control over the AI systems – and even more so, of the AI of AIs – when they go down or are attacked, is a fundamental requirement for humanity's rule over machines. This question – how to protect the AI system itself—will prove, in time, to be just as important as ensuring that AI causes no harm. Several basic approaches exist and will be described below. They are not alternatives but models likely to co-exist, each dealing with particular aspects of emerging risks to AI and by AI. Together, these approaches form a portfolio of potential remedies to be implemented by policy makers, technologists, and managers, each of whom has a somewhat different priority (Inouye et al., 2021; Prem, 2023). As experience is gained, the benefits and weaknesses can be quantified, leading to various key performance parameters.

3. Managerial actions

Since AI disruptions are inevitable, organizations must be prepared, making investments in duplicative facilities, stress testing, and deploying hiring qualified problem solvers on 24/7 duty (Besen, 2024). There must be a collaboration across companies and industries, and these actions have surfaced already for the more general cyber-security.

Organizations will need redundancy. For example, Cloudflare's control and analytics systems are based on three separate and independent data centers in Oregon, each with separate power suppliers and each with its own network connections (Haller, 2024). Organizations also need vendor diversity, for example a multi-cloud architecture that distributes traffic and processing across several clouds. Multi-clouds are deployed by companies such as Capital One, Netflix, GE, Financial Times, and HSBC (Abbas, 2023).

All of these actions will be costly. When companies take these steps they reduce the probability of things going wrong, but substantial risks remains. Because it is hard to quantify this residual risk in its probability and severity, and because a failure of an inter-system AI can have immense consequential damages, no one will insure it. Even if the likelihood is small, it is unknowable. And when companies offering digital technology or services assume no liability for consequential damage and do not absorb the negative externalities for problems, they will under-invest, leading inevitably to some form of regulation.

4. Self-regulation

In the US, the preferred approach, promoted by the industry, has been self-regulation. While Washington policymakers were debating whether a governmental AI regulator should be established, seven major AI firms launched an industry-led body to develop safety standards (Zakrzewski & Tiku, 2023). The industry group advocated a set of principles for safe AI, such as third-party security

checks and watermarking of AI-generated content, to reduce the spread of misinformation (Kang, 2023). Many of these practices have already been adopted by OpenAI, Google, and Microsoft. Advantages of self-regulation are that experts are managing a situation in a practical rather than legal-procedural fashion. However, self-regulation may lead to a cartel-like coordination among established competitors, excluding non-industry participants such as public interest advocates, limiting liability, and offering only weak due process rights.

Several digital leaders such as Sam Altman, Elon Musk, Steve Wozniak, Andrew Yang, and Demis Hassabis, have advocated to take a timeout from development, in order to create a breathing space for reflection and for the establishment of rules and controls over the emerging AI juggernauts (Metz & Schmidt, 2023). This perspective combines several motives. One is a genuine concern about the threat posed by an unfettered technology, and such calls for responsible self-regulation are laudable. A second and less public-spirited reason for some technology leaders might be to moderate costly competition among themselves, and to slow down rivals catching up. However, a freeze on AI technology development is unlikely to work. An informal understanding is apt to be violated by some firms, while a formal agreement will cause antitrust problems. One would therefore have to set governmental rules. But any regulation-based time-out, in turn, would have to be coordinated with other leading countries which seems unlikely beyond the sharing of high-minded communiques. And even on a national level, how would such restrictions work? Is it realistic to jail a software programmer for writing advanced code in her basement? To fine Apple for coming out with an un-approved upgrade for its iPhone?

Such high-minded disavowal by technology leaders of their own creation is similar to what happened after the development of the first atomic bombs, when serious scientists involved directly or indirectly in their development, such as Robert Oppenheimer, Leo Szilard, and Nils Bohr wanted their use limited, stopped, or internationally controlled. Part of this is explainable by their deeper understanding of the terrifying force of the new technology; and part might be an understandable reluctance to assume the responsibility for their own creation's reality beyond science. Today, the well-meant calls for a moratorium by AI leaders will be as ineffective as those of their nuclear predecessors eighty years earlier, indeed even less so, since at that time a handful of decision makers, mostly in the U.S. government, could have stopped things for a few years, and because access to basic materials such as uranium and heavy water could be controlled. But with AI, an environment of many thousands of aspiring engineers, hundreds of competing companies, and dozens of rival nations cannot be controlled. At best, one might draft national laws and international agreements to establish regulatory restrictions.

5. A governmental role

There are several dimensions to governmental action. The first question is timing. As described in an Alan Turing Institute study: "Any further delay will risk one of two undesirable outcomes: either a scenario where AI risks transition into widespread harms, directly impacting individuals and groups in society; or the converse scenario where widespread fear of AI risk results in a lack of adoption, meaning the UK does not benefit from the many societal benefits presented by these technologies" (Janjeva et al., 2023). Similarly, the counter-push against a regulation of AI in the U.S. is that it runs counter to U.S. economic and technology interests, while restrictive regulations would be shaped by laggard countries.

This article does not aim to explore how to regulate protectively against a panoply of societal problems caused by an unrestrained AI, some of which will be serious. The focus here is on the issues of AI resiliency only. That area is harder to reach by government than many other problems. Most AI issues are provider-based, such as privacy violations, algorithmic bias, or hallucinations leading to calamities. Regulations can then be imposed on typically large and visible provider companies, holding them accountable for compliance. True, there will always be some firms that operate under the radar, but most firms will be reached. But for resiliency the bad actors number hundreds of thousands of anonymous individuals spread around the globe, not to mention hostile governments. Regulations will not deter the former and in some cases even add to the thrill. Therefore, the approaches to strengthen AI resiliency are to impose obligations on the recipient end, the provider companies, to raise defensive barriers to hostile activities and to deal with their consequences faster. This will not be easy to establish by governmental fiat, considering that the companies themselves are incentivized to do much of that on their own and have not been overly successful. That said, there are a variety of policy options for governments if they seek to address AI resiliency.

5.1. Law enforcement and national security operations

The investigation and exposure of hackers and of foreign threats is an obvious governmental responsibility as part of more general law enforcement and national security efforts.

5.2. Establish the basic standards for human control

Perhaps the most important measure is to set rules that establish ultimate human supremacy. The system that will emerge will not be fully automatic. It will involve semi-automatic human/AI combinations, like an autopilot in an airplane or in self-driving cars. The problem is the handover. Under what circumstances does human take over? Autopilots are important because pilots make mistakes. Pilots are important because the AI makes mistakes. The key is to establish protocols for shared responsibilities.

There must be rules on those areas where a decision must be authorized by humans. This covers, in particular, direct decisions over the life and death, for instance whether an airplane would be shot down when it intrudes into restricted airspace. Or, whether the health of young people would be prioritized over that of seniors. Whatever the final decisions, there must be a human as the party taking ultimate accountability and responsibility.

Decisions of handing control over to humans must be made rapidly, and therefore mostly by AI functionalities themselves. To do so reliably, any control technology would have to be totally independent from the systems that it must control, otherwise it would not do the job securely.

All of these upper-level rules would have to be controlled by a human entity, probably run by governmental and private policy makers and technologists.

5.3. Set liability rules

When it comes to digital activities, private companies' standard of care is strongly affected by the liability that they assume for consequential damages. An example is Section 230 of the U.S. Communications Act which absolves digital platforms from most liability for the content provided by users (Telecommunications Act of 1996, 1996). This has accelerated the development of such platforms but has also caused a proliferation of socially problematic content.

An example for AI liability is Air Canada and the 'lying AI Chatbot,' Case 2024 (Moffatt v. Air Canada, 2024). The airline's chatbot told a customer that there would be a bereavement discount. That information was incorrect, and when the passenger asked for the promised refund. Air Canada refused, presumably to defend the principle that it was not responsible for whatever mistakes the computer software came up with. The passenger sued and Air Canada lost. The principle established was that the principal of an AI system is responsible for its AI agent's mistakes.

A liability rule must be a goldilocks scenario, a middle course between too much accountability and too little. For AI, this means a balancing of financial responsibility for the consequences when things go wrong, versus the incentives for innovation. Such liability rules can be dynamic, lenient in the early stages of development when incentives might be more important, and stricter when adoptions are widespread. Catastrophic damages might never be covered because such a rule would choke off innovation with unknown future impacts. But a lower liability exposure and more stringent test for fault will create incentives for improving resiliency. How would an optimal liability system be reached? Most likely through cases before courts and administrative agencies, in a quasi-common law system.

5.4. Protect A competitive market structure

As with most digital operations, there are large economies of scale and scope as well as network externalities, and they lead to highly concentrated market structures. There are well-known problems when there are only a handful of suppliers and operators, such as giving users fewer options, having users pay more, and reducing users' risk diversification. On their part, AI providers have lower incentives to innovate in protective features. Even more problematic is that in the absence of serious rivals they might acquire vast influence because their algorithms run so many social and economic activities.

At the same time it is easier for government regulators to be effective if the number of players is small and manageable. Adding up these considerations shows that if governments seek an optimal market structure, a complex balancing act is needed.

5.5. Set transparency rules

This would include reporting requirements on problems encountered, actions taken, testing results, etc. When OpenAI was hacked in 2023, the company withheld that information from the public or from law enforcement agencies, based on the spurious argument that society had not been directly affected. It took a whistleblower to disclose the problem (Metz, 2024). Some people advocate a public disclosure of the AI algorithms themselves, the idea to allow for the public to check on hidden biases in those algorithms. However, this measure would also help bad actors to figure out ways to attack and penetrate, as well as reduce innovation by enabling a copying by competitors.

5.6. Technical and quality standards

Setting safety standards is a worthwhile goal, but easier said than done. How to define AI safety? How to measure robustness? How to reverse-engineer a problem to its cause?

There has been some certification of more general security-certified modules and devices. In the U.S., some of this is done by the Department of Homeland Security (DHS), in coordination with its Cybersecurity and Infrastructure Security Agency (CISA) (Ribeiro, 2024). The EU has its ENISA cybersecurity certification system, voluntary so far. These certification efforts are being extended to AI, as awareness of its problems rises.

Advocates have argued for the licensing of AI software, along the model for drug approval. OpenAI's CEO Sam Altman has proposed licensing of the most powerful AI platforms, possibly by a new agency (Chatterjee & Kern, 2023). Licensees would have to conform to certain requirements to get approval. Support for this concept is wide but not deep, because it is a classic devil-is-in-the-details issue. The requirement for testing, applying, reviewing, and evaluating stands to delay the introduction of new products, and in the digital world with its rapid pace of new versions and improvements, licensing in advance is not a practical idea.

5.7. Provision of support services by governmental entities

A number of government support services can be envisaged:

- Backups to AI platforms, such as a national defense and national security infrastructure of networks and data centers.
- AI Insurance of last resort: Government-sponsored insurance exists for flood insurance as basic coverage. Private insurers could cover additional exposure and would have incentives to create a culture of risk-reduction.
- Restructuring of emergency communications: The traditional system of dealing with major emergencies and restoring normalcy has been based on top-down, hierarchical, military-style public safety organizations (Noam & Sato, 1995; Noam, 2001a,b). With AI, governmental entities would instead enable an ad-hoc integration of relevant systems, based on users and on needs, and different for each situation. Such a system would take actions in fixing, ordering, deploying, supplying, and reporting.
- Government support of detection mechanisms for AI problems.

5.8. Coordination

Government could be the convenor of major participants to coordinate AI protections (Gomez et al., 2023). They would include:

- Governmental departments
- Different levels of government.
- Private parties.
- Private/public collaborations.
- International collaborations.
- Information exchange about threats and problems.
- Promotion of best practices.

5.9. Set operational rules

Among governmental actions, potentially the most effective for AI resiliency would be to set minimum standards for its safe operation and recovery for critical systems. Government could set such guardrails for safety in a variety of ways:

- Set standards on interoperation so that in a crisis, different AI systems can help each other and talk to each other.
- Require periodic stress-testing and resiliency auditing by private AI systems.
- Require that every essential AI provider and major supplier connected to such an AI provider have a rigorous recovery plan.
- Require redundancy.
- Require the separation of AI network management from operational data centers.
- Conduct war games.

6. Examples of governmental regulatory initiatives

At one end of the spectrum, some have advocated the creation of a specialized agency, a governmental digital agency/authority (Cote et al., 2023). More commonly, countries and regions have edged towards a higher level of regulation without necessarily creating new agencies.

Heeding warning calls about AI in general, and the massive amount of fear scenarios that AI received, the EU parliament drafted an AI Act and passed it in March 2024, with an overwhelming vote of 523 to 46 against, and 49 abstentions (Harrington et al., 2024; Naughton, 2023). It aims, in a proactive way, to prevent "Unacceptable risks" which include.

- Manipulation of people or specific vulnerable groups, or example children
- Social scoring based on behavior, socio-economic status or personal characteristics
- Remote biometric identification systems such as facial recognition.

These concerns seem to be defensible, though few people expected this to be the end rather than the beginning of more controversial expansions that would reach other subjects. It would develop a cadre of EU administrators with experience in AI regulation. The EU also flagged high risk products such as toys, planes, cars, medical devices and elevators. It set transparency and disclosure requirements, designs of AI to prevent it from generating illegal content. And it required the disclosure of copyrighted data used for training (European Parliament, 2023). For all of its energetic push to deal with AI harm, the resiliency aspect received only a minor coverage, since protecting AI ran somewhat *against* the narrative of protecting society *from* AI. Paragraph 113 reads: "The robustness of high-risk AI systems may be achieved through technical redundancy solutions, which may include backup or fail-safe plans … High-risk AI systems shall be resilient against attempts by unauthorized third parties to alter their use, outputs or performance by exploiting system vulnerabilities" (Regulation 2024/1689).

In China, the three major regulations are the 2021 regulation on recommendation algorithms, the 2022 rules for synthetically generated content, and the 2023 rules on generative AI (Xu et al., 2023). They require companies engaging in generative AI to uphold core socialist values, engage in non-discrimination when selecting training data and models, respect IP rules, respect the privacy of others, and engage in as much transparency as possible. These regulations would create experience, expertise, and tools such as disclosure requirements, auditing, and performance standards, for a subsequent crafting of a more comprehensive national AI law (Sheehan, 2023).

In the U.S., the Federal Trade Commission (FTC) initiated an analysis of OpenAI in 2023, stating that it was investigating whether the company had engaged in unfair or deceptive privacy or data security aspects or unfair practices (Zakrzewski & Tiku, 2023). The FTC issued a report on AI, expressing a concern with harms such as inaccuracy, bias, discrimination, and creeping commercial surveillance (Federal Trade Commission, 2022), but not about the security of AI-based operations. In 2023, the Biden White House issued an Executive Order that applied to all federal agencies (Exec. Order No. 14110, 3 C.F.R. 12110, 2023). It required "standardized evaluations of AI systems, ….to test, understand, and mitigate risks." President Trump immediately replaced this Order by a new one, requiring an action plan to promote U.S. "global AI dominance" by revoking policies that "act as barriers to American AI innovation" (White House, 2025). This Order would not necessarily stop individual States from going forward with regulations, as they already have.

Another approach, more traditional in nature, is a treaty among countries. The aims are to deal with supranational threats to global AI operations. Other goals are to reduce a 'race-to-the-bottom' competition among countries and companies, and the emergence of permissive 'haven' countries. Thus, the Council of Europe drafted a legally binding treaty for artificial intelligence. The treaty requires signatories to ensure that AI is designed, developed, and applied in a way that protects human rights, democracy, and the rule of law. This would possibly include moratoriums on technologies that pose a risk to human rights, for example facial recognition (Heikkilä, 2023).

There are several additional initiatives on trans-national agreements and treaties. AI principles were also established by the OECD. Member states agreed in 2019 on a set of nonbinding principles for AI development. Another approach was the Global Partnership on AI (GPAI), a 2020 initiative by Canadian Prime Minister Justin Trudeau and French President Emmanuel Macron to create an international body to foster international research collaboration and develop policies for responsible AI. Still another approach is to set technical industry standards. The International Organization for Standardization (ISO) has standards for risk management and impact assessments in AI. UNESCO, similarly, adopted a voluntary AI ethics framework. Countries pledged to introduce AI's impact assessments for ethics, the environment, and gender equality, and to oppose its use for mass surveillance.

7. Technology tools

As we discussed above, governmental actions and industry self regulation will not be able to resolve the numerous challenges to the resiliency of AI systems. According to one estimate, about 600,000 unique malware files are launched every day (Pupillo et al., 2021), and such future attacks will target AI systems. In such a situation, might technology be part of the solution to the problems caused by technology (Zohuri & Rahmani, 2019)? For AI resiliency, the answer is yes.

AI companies build protections into their software that aim to prevent certain activities such as deep fakes or disinformation. The vendors, under competitive and PR pressures, are trying to remedy problems. They are, however, not especially effective. OpenAI, for example, appointed a safety committee that includes Paul Nakasone, a former Army general and head of the National Security Agency and of the Cyber Command (Metz, 2024). Other companies such as Meta have taken a different approach by, in effect, outsourcing system safety through an open-source platform. The user community is then engaged in fixing problems.

IT system vendors offer protective and analytical AI tools to users. Amazon's AWS offers failure simulations with a tool called the Fault Injection Service (FIS) (Sharwood, 2023). Users can stage fake faults so they can test their ability to recover from the unexpected. AI has also been used by vendors to test the resiliency of systems. Microsoft as well as Anthropic conducted red team exercises trying to cause a service disruption (UK Department for Science, Innovation & Technology, 2024; Torkura et al., 2020; Anthropic, 2023c).

Similar tools can do chaos testing and 'blast radius analysis' which tests who and what will be affected by an outage or breach. Netflix created for itself a tool called Chaos Monkey (subsequently upgraded to Chaos Gorilla and Chaos Kong) to randomly break its systems so its engineers could get better at creating fixes. Such defensive AI can help predict, identify, and remedy resiliency problems. It can be used in planning, scenario playing, and the testing of vulnerabilities. It can help identify deep-fakes and divert attacks to mirror sites. It can automate threat detection in real time and it can analyze large data sets and communicate with other systems around the world.

Realistically, stemming the flood of attacks by malicious actors is far too complex for humans to manage by themselves, and it is impossible to do so rapidly. AI systems are needed to help control the flood of attacks by malicious AI systems or of breakdowns of faulty AI, to speedily identify the nature of the problem, to come up with commands for remedies, to communicate them, and to monitor their execution. AI will be part of the good-actor defensive team for resiliency, fighting against the offensive teams of bad-actor AI. Paradoxically perhaps, the best solution to bad AI might well be protective AI. But it is not unique that the same technology which creates negative effects also contains solutions to the problems it has created. Defensive missiles protect against offensive missiles. Chemicals compounds counteract addictions caused by other chemical compounds. People use handguns to rob banks and to protect banks. Some antidotes mirror the poison they seek to excise. For AI resiliency, too, the problem is also part of the solution.

8. Outlook

This article has identified and analyzed the resiliency challenges of artificial intelligence. How to deal with AI resiliency is an important issue, because AI is a profound challenge for all digital activities. The key issue is to ensure that we do not move into a post-human environment where we lose control over complex systems upon which societies depend. We should instead think about *enhanced* humanity, about a symbiotic environment of individuals aided by their AI. The path is to establish arrangements of shared responsibilities. The resiliency of AI is such a combined task, shared by humans overall and by the AI itself in the actual operation.

How to structure such arrangements? Will technology and decentralized market forces be enough to assure resiliency? Or, at the

other extreme, will we need a national or even global control room to flip switches when things go wrong? Do we wait for real problems and learn from them, or do we set up controls ex-ante? And if we create safeguards, how much strictness is useful, and how much is retarding the development of important tools of innovation? These and other questions are topics for a forward-looking research agenda that would encompass more theoretical and empirical economic and operations research approaches, technology studies, as well as legal and policy analyses (Eigner et al., 2021; Fraccascia et al., 2018). Several topics are provided below.

Even where the probabilities of harm from malfunctioning or compromised AI are low, their impact might be so high as to require attention, preparedness, and action. The reliability of digital systems is increasingly a pillar for social and economic stability. Therefore, AI resiliency is a public responsibility for policy makers, an emerging management responsibility for practitioners, a task to technologists, and an important challenge to researchers.

Appendix. A Research Agenda for AI Resiliency

The impact of open-source vs proprietary AI software platforms.
The impact of free markets in AI on its resiliency.
How to balance necessary redundancy with inefficient duplication.
Creating incentives for investments in AI resiliency.
The impact of different liability rules.
Supra-national cooperation mechanisms for supra-national AI threats.
Optimal redundancy: what it is and how to achieve it.
Prioritizing resiliency in a hierarchy of digital activity layers.
Models of self-regulation for the AI sector and for sectors deploying AI.
Modeling the balance of shared human/AI responsibilities in a hierarchy of threats.
Interoperation requirements for AI systems—technology and law.
International dimensions of threats and remedies.
Licensing of AI algorithms and the experience of other industries.
Ex-ante vs ex-post rule setting for AI resilience regulation.
Balancing data localization and sovereignty with resilience.

References

Abbas, A. (2023). Why is multi-cloud the future of resilient enterprises? *Techopedia*. https://www.techopedia.com/why-is-multi-cloud-the-future-of-resilient-enterprises.

Aggarwal, G. (2023). Harnessing GenAI: Building cyber resilience against offensive AI. *Forbes*. https://www.forbes.com/sites/forbestechcouncil/2023/09/25/harnessing-genai-building-cyber-resilience-against-offensive-ai/.

Akhtar, N., & Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access, 6*, 14410–14430. https://doi.org/10.1109/ACCESS.2018.2807385

Anthropic. (2023a). Anthropic's responsible scaling policy. Version 1.0 https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf.

Anthropic. (2023b). Frontier model security. https://www.anthropic.com/news/frontier-model-security#entry:146893@1:url.

Anthropic. (2023c). Frontier threats red teaming for AI safety. https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety#entry:146918@1:url.

Apruzzese, G., et al. (2023). "Real attackers don't compute gradients": Bridging the gap between adversarial ML research and practice. In *2023 IEEE conference on secure and trustworthy machine learning (SaTML)* (pp. 339–364). IEEE. https://doi.org/10.1109/SaTML54575.2023.00031.

AWS. (2024). Cloud adoption Framework for artificial intelligence, machine learning, and generative AI - whitepaper. https://docs.aws.amazon.com/whitepapers/latest/aws-caf-for-ai/aws-caf-for-ai.html.

Bécue, A., Praça, I., & Gama, J. (2021). Artificial intelligence, cyber-threats and industry 4.0: Challenges and opportunities. *Artificial Intelligence Review, 54*(5), 3849–3886. https://doi.org/10.1007/s10462-020-09942-2

Benzinger, L., Ursin, F., Balke, W., Kacprowski, T., & Salloch, S. (2023). Should artificial intelligence be used to support clinical ethical decision-making? A systematic review of reasons. *BMC Medical Ethics, 24*(48). https://doi.org/10.1186/s12910-023-00929-6

Besen, S. (2024). The landscape of emerging AI agent architectures for reasoning, planning, and tool calling: A survey. *Data Science*. https://towardsdatascience.com/the-landscape-of-emerging-ai-agent-architectures-for-reasoning-planning-and-tool-calling-a-a95214b743c1.

Boehm, J., Salmanian, W., & Wallance, D. (2023). *A technology survival guide for resilience*. McKinsey & Company. https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/a-technology-survival-guide-for-resilience.

Boulemtafes, A., Derhab, A., & Challal, Y. (2020). A review of privacy-preserving techniques for deep learning. *Neurocomputing, 384*, 21–45. https://doi.org/10.1016/j.neucom.2019.11.041

Bradley, P. (2020). Risk management standards and the active management of malicious intent in artificial superintelligence. *AI & Society, 35*(2), 319–328. https://doi.org/10.1007/s00146-019-00890-2

Breslow, J. (2024). *Delta's CEO says the CrowdStrike outage cost the airline $500 million in 5 days*. NPR. https://www.npr.org/2024/07/31/nx-s1-5058652/delta-crowdstrike-outage-500-million-dollars.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Ó hÉigeartaigh, S., Beard, S., Belfield, H., Farquhar, S., … Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv https://doi.org/10.48550/arXiv.1802.07228.

Carlo, A., Manti, B. P., WA, M., BA, S., Casamassima, F., Boschetti, N., Breda, P., & Rahloff, T. (2023). The importance of cybersecurity frameworks to regulate emergent AI technologies for space applications. *Journal of Space Safety Engineering, 10*(4), 474–482. https://doi.org/10.1016/j.jsse.2023.08.002

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2021). A survey on adversarial attacks and defences. *CAAI Trans. Intell. Technol., 6*, 25–45. https://doi.org/10.1049/cit2.12028

Chatterjee, M., & Kern, R. (2023). *Washington confronts a new AI fight*. Politico. https://www.politico.com/news/2023/05/17/washington-confronts-a-new-ai-fight-00097425.

Chiang, F., & Gairola, D. (2018). InfoClean: Protecting sensitive information in data cleaning. *Journal of Data and Information Quality, 9*(4). https://doi.org/10.1145/3190577, 22:1-22:26.
Columbus, L. (2023). 5 ways CISOs can prepare for generative AI's security challenges and opportunities. *VentureBeat*. https://venturebeat.com/security/5-ways-cisos-can-prepare-for-generative-ai-security-challenges-and-opportunities/.
Cote, B. J., Vella Moeller, E., Finch, B. E., Fork, W. E., & Trozzo, A. (2023). *Congress contemplates creating a new federal AI regulatory agency*. Pillsbury. https://www.pillsburylaw.com/en/news-and-insights/congress-federal-ai-regulatory-agency.html.
Council of Europe. (2024). The Framework convention on artificial intelligence. https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence.
Cutter, S. L., Ahearn, J. A., Amadei, B., Crawford, P., Eide, E. A., Galloway, G. E., Goodchild, M. F., Kunreuther, H. C., Li-Vollmer, M., & Schoch-Spana, M. (2013). Disaster resilience: A national imperative. *Environ. Sci. Policy Sustain. Dev., 55*, 25–29. https://doi.org/10.1080/00139157.2013.768076
Dano, M. (2024). *Counting the cost of AT&T's outage*. LightReading. https://www.lightreading.com/network-automation/counting-the-cost-of-at-t-s-outage.
Dawson, A. (2023). *The state of AI security*. Cohere. https://txt.cohere.com/the-state-of-ai-security.
Eigner, O., Eresheim, S., Kieseberg, P., Klausner, L. D., Pirker, M., Priebe, T., Tjoa, S., Marulli, F., & Mercaldo, F. (2021). Towards resilient artificial intelligence: Survey and research issues. In *Proceedings of the 2021 IEEE international conference on cyber security and resilience (CSR), rhodes, Greece, 26-28 july 2021; IEEE: Piscataway township, NJ, USA, 2021*. https://doi.org/10.1109/CSR51186.2021.9527986
EU General Secretariat of the Council. (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.
European Parliament. (2023). EU AI act: First regulation on artificial intelligence. https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.
European Union Agency for Cybersecurity. (2020). AI cybersecurity challenges: Threat landscape for artificial intelligence. https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges.
European Union Agency for Cybersecurity. (2023). Cybersecurity and privacy in AI: Forecasting demand on electricity grids. https://data.europa.eu/doi/10.2824/92851.
Exec. Order No. 14110, 3 C.F.R. 12110. (2023). *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. https://www.govinfo.gov/app/details/CFR-2024-title3-vol1/CFR-2024-title3-vol1-eo14110.
Federal Trade Commission. (2022). FTC report warns about using artificial intelligence to combat online problems. https://www.ftc.gov/news-events/news/press-releases/2022/06/ftc-report-warns-about-using-artificial-intelligence-combat-online-problems.
Ferrer, X., van Neunen, T., Such, J. M., Coté, M., & Criado, N. (2021). Bias and discrimination in AI: A cross-disciplinary perspective. *IEEE Technology and Society Magazine, 40*(2), 72–80. https://doi.org/10.1109/MTS.2021.3056293
Fraccascia, L., Giannoccaro, I., & Albino, V. (2018). Resilience of complex systems: State of the art and directions for future research. *Complexity*. https://doi.org/10.1155/2018/3421529
Gesser, A., Gressel, A., Xu, M., & Allaman, S. J. (2022). *When humans and machines disagree – the myth of "AI errors" and unlocking the promise of AI through optimal decision making*. Debevoise & Plimpton. https://www.debevoisedatablog.com/2022/11/14/when-humans-and-machines-disagree-the-myth-of-ai-errors-and-unlocking-the-promise-of-ai-through-optimal-decision-making-adm-algorithm/.
Gienow, M. (2023). Paris is drowning: GCP's region failure in age of operational resilience. *The New Stack*. https://thenewstack.io/paris-is-drowning-gcps-region-failure-in-age-of-operational-resilience/.
Gomez, C., Unberath, M., & Huang, C.-M. (2023). Mitigating knowledge imbalance in AI-advised decision-making though collaborative user involvement. *International Journal of Human-Computer Studies, 172*. https://doi.org/10.1016/j.ijhcs.2022.102977
Gongye, C., Li, H., Zhang, X., Sabbagh, M., Yuan, G., Lin, X., Wahl, T., & Fei, Y. (2020). New passive and active attacks on deep neural networks in medical applications. In *Proceedings of the ICCAD '20: IEEE/ACM international Conference on computer-aided design, ACM digital library*. https://doi.org/10.1145/3400302.3418782
Greshake, K., et al. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM workshop on artificial intelligence and security* (pp. 79–90). Association for Computing Machinery (AISec í23). https://doi.org/10.1145/3605764.3623985.
Haller, K. (2024). *A Guide to cloud resilience: Maximize security, minimize downtime.* DataCenter knowledge. https://www.datacenterknowledge.com/cloud/guide-cloud-resilience-maximize-security-minimize-downtime.
Hansen, R., & Venables, P. (2023). *Introducing Google's secure AI Framework*. Google. https://blog.google/technology/safety-security/introducing-googles-secure-ai-framework/.
Harrington, M., Hansen, M., Peets, L., Drake, M., & Young, M. (2024). *EU parliament adopts AI act*. Covington. https://www.insideprivacy.com/artificial-intelligence/eu-parliament-adopts-ai-act/.
Heikkilä, M. (2023). Our quick guide to the 6 ways we can regulate AI. *MIT Technology Review*. https://www.technologyreview.com/2023/05/22/1073482/our-quick-guide-to-the-6-ways-we-can-regulate-ai/.
Hersher, R. (2017). *Amazon and the $150 million typo*. NPR. https://www.npr.org/sections/thetwo-way/2017/03/03/518322734/amazon-and-the-150-million-typo.
Hu, Y., Kuang, W., Qin, Z., Li, K., Zhang, J., Gao, Y., Li, W., & Li, K. (2021). Artificial intelligence security: Threats and countermeasures. *ACM Computing Surveys, 55* (1). https://doi.org/10.1145/3487890, 20:1-20:36.
Huang, C., Zhang, Z., Mao, B., & Yao, X. (2023). An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence, 4*(4), 799–819. https://doi.org/10.1109/TAI.2022.3194503
Inouye, B. D., Brosi, B. J., Le Sage, E. H., & Lerdau, M. T. (2021). Trade-offs among resilience, robustness, stability, and performance and how we might study them. *Integrative and Comparative Biology, 61*, 2180–2189. https://doi.org/10.1093/icb/icab178
Janjeva, A., Mulani, N., Powell, R., Whittlestone, J., & Avin, S. (2023). Strengthening resilience to AI risk. *CETaS Briefing Papers*. https://cetas.turing.ac.uk/publications/strengthening-resilience-ai-risk.
Kang, C. (2023). *U.S., regulating A.I. Is in its 'early days'*. New York Times. https://www.nytimes.com/2023/07/21/technology/ai-united-states-regulation.html.
Kerner, S. M. (2024). CrowdStrike outage explained: What caused it and what's next. *TechTarget*. https://www.techtarget.com/whatis/feature/Explaining-the-largest-IT-outage-in-history-and-whats-next.
Lamanna, C., & Byrne, L. (2018). Should artificial intelligence augment medical decision making? The case for an autonomy algorithm. *AMA Journal of Ethics, 20*(9). https://doi.org/10.1001/amajethics.2018.902
Li, C., Zhang, M., & He, Y. (2022). The stability-efficiency dilemma: Investigating sequence length warmup for training GPT models. *Advances in Neural Information Processing Systems, 35*, 26736–26750. https://dl.acm.org/doi/10.5555/3600270.3602209.
Luong, N., & Arnold, Z. (2021). *China's artificial intelligence industry alliance*. Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/chinas-artificial-intelligence-industry-alliance/.
Majeed, A., & Hwang, S. O. (2023). When AI meets information privacy: The adversarial role of AI in data sharing scenario. *IEEE Access, 11*, 76177–76195. https://doi.org/10.1109/ACCESS.2023.3297646
Marshall, A., Rojas, R., Stokes, J., & Brinkman, D. (2024). *Securing the Future of AI and ML at Microsoft.* Microsoft. https://learn.microsoft.com/en-us/security/engineering/securing-artificial-intelligence-machine-learning.
McKendrick, J., & Thurai, A. (2022). AI isn't ready to make unsupervised decisions. *Harvard Business Review*. https://hbr.org/2022/09/ai-isnt-ready-to-make-unsupervised-decisions.
METI. (2024). *Launch of AI safety Institute*. Trade and Industry: Ministry of Economy. https://www.meti.go.jp/english/press/2024/0214_001.html.

Metz, C. (2024). *A Hacker Stole OpenAI Secrets, Raising Fears That China Could, Too. New York Times*. https://www.nytimes.com/2024/07/04/technology/openai-hack.html.
Metz, C., & Schmidt, G. (2023). *Elon Musk and Others Call for Pause on A.I., Citing 'Profound Risks to Society'. New York Times*. https://www.nytimes.com/2023/03/29/technology/ai-artificial-intelligence-musk-risks.html.
Mirsky, Y., Demontis, A., Kotak, J., Shankar, R., Gelei, D., Yang, L., Zhang, X., Pintor, M., Lee, W., Elovici, Y., & Biggio, B. (2023). The threat of offensive AI to organizations. *Computers & Security, 124*(C). https://doi.org/10.1016/j.cose.2022.103006
Moffatt, v (2024). Air Canada, 2024 BCCRT 149. https://decisions.civilresolutionbc.ca/crt/crtd/en/item/525448/index.do.
MSIT. (2024). MSIT announce strategy to realize trustworthy artificial intelligence. https://www.msit.go.kr/eng/bbs/view.do?sCode=eng&mId=4&mPid=2&pageIndex=&bbsSeqNo=42&nttSeqNo=509&searchOpt=ALL&searchTxt.
Naughton, J. (2023). Europe's AI crackdown looks doomed to be felled by Silicon Valley lobbying power. *The Guardian*. https://www.theguardian.com/commentisfree/2023/dec/02/eu-artificial-intelligence-safety-bill-silicon-valley-lobbying.
NIST. (2023). Artificial intelligence risk management Framework (AI RMF 1.0). https://doi.org/10.6028/NIST.AI.100-1.
NIST. (2024). AI risk management Framework. https://www.nist.gov/itl/ai-risk-management-framework.
Noam, E. (1988). The next stage in telecommunications evolution: The pluralistic network. In *Paper presentation at the international telecom society conference, Boston, MA*.
Noam, E. (1994). Beyond liberalization: From the network of networks to the system of systems. *Telecommunications Policy, 18*, 286–294.
Noam, E. (2001a). *Interconnecting the network of networks*. The MIT Press.
Noam, E. (2001b). Straining communications systems to the limit. *New York Times*. https://www.nytimes.com/2001/09/24/business/new-economy-straining-communications-systems-to-the-limit.html.
Noam, E. (2023). AI as citizen empowerment and game changer for regulation. In *Paper presentation at TPRC conference, Washington DC*.
Noam, E., & Sato, H. (1995). Kobe's lesson: Dial 711 for 'open' emergency communications. *Telecommunications Policy, 19*(8), 595–599.
NVIDIA. (2024). NVIDIA AI cybersecurity. https://www.nvidia.com/en-gb/industries/cybersecurity/.
OECD iLibrary. (2024). OECD artificial intelligence papers. https://www.oecd-ilibrary.org/science-and-technology/oecd-artificial-intelligence-papers_dee339a8-en.
OpenAI. (2023). OpenAI's approach to frontier risk. https://openai.com/global-affairs/our-approach-to-frontier-risk.
OWASP. (2024). AI security and privacy guide. https://owasp.org/www-project-ai-security-and-privacy-guide/.
Pendleton, J., Levite, A., & Kolasky, B. (2024). *Cloud reassurance: A Framework to engance resilience and trust*. Carnegie Endowment for International Peace. https://carnegieendowment.org/research/2024/01/cloud-reassurance-a-framework-to-enhance-resilience-and-trust.
Prem, E. (2023). From ethical AI frameworks to tools: A review of approaches. *AI Ethics, 3*, 699–716. https://doi.org/10.1007/s43681-023-00258-9
Prince, M. (2023). *Post mortem on the Cloudflare control plane and analytics outage*. CloudFlare. https://blog.cloudflare.com/post-mortem-on-cloudflare-control-plane-and-analytics-outage.
Pupillo, L., Fantin, S., Ferreira, A., & Polito, C. (2021). Artificial intelligence and cybersecurity. *CEPS*. https://www.ceps.eu/ceps-publications/artificial-intelligence-and-cybersecurity-2/.
Quaquebeke, N., & Gerpott, F. (2023). The now, new, and next of digital leadership: How artificial intelligence (AI) will take over and change leadership as we know it. *Journal of Leadership & Organizational Studies, 30*(3). https://doi.org/10.1177/15480518231181731
Regulation 2024/1689. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). http://data.europa.eu/eli/reg/2024/1689/oj.
Ribeiro, A. (2024). US DHS delivers safety and security guidelines to secure critical infrastructure from AI-related threats. Industrial Cyber https://industrialcyber.co/ai/us-dhs-delivers-safety-and-security-guidelines-to-secure-critical-infrastructure-from-ai-related-threats/.
Rosiek, T. (2024). How to build data resilience by leveraging backup, zero trust and AI. *C4ISRNET*. https://www.c4isrnet.com/opinion/2024/02/26/how-to-build-data-resilience-by-leveraging-backup-zero-trust-and-ai/.
Saeed, S., Suayyid, S., Al-Ghamdi, M., Al-Muhaisen, H., & Almuhaideb, A. (2023). A systematic literature review on cyber threat intelligence for organizational cybersecurity resilience. *Sensors, 23*(16). https://doi.org/10.3390/s23167273
Serentschy, G. (2024). *Digital infrastructure resilience and security policy implications and mitigation measures*. Serentschy Advisory Services.
Sharwood, S. (2023). You're so worried about AWS reliability, the cloud giant now lets you simulate major outages. *The Register*. https://www.theregister.com/2023/12/01/aws_az_fault_injection_service/.
Sheehan, M. (2023). *China's AI regulations and how they get made*. Carnegie Endowment for International Peace. https://carnegieendowment.org/2023/07/10/china-s-ai-regulations-and-how-they-get-made-pub-90117.
Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S., Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., … Dafoe, A. (2023). Model evaluation for extreme risks. *arXiv*. https://doi.org/10.48550/arXiv.2305.15324
Shneiderman, B. (2022). Ensuring human control over AI-infused systems. *NAE Perspectives*. https://www.nationalacademies.org/news/2022/04/ensuring-human-control-over-ai-infused-systems.
Silva, D. D., & Alahakoon, D. (2022). An artificial intelligence life cycle: From conception to production. *Patterns, 3*(6). https://doi.org/10.1016/j.patter.2022.100489
Telecommunications Act of 1996. (1996). https://www.congress.gov/104/plaws/publ104/PLAW-104publ104.htm.
Torkura, K. A., Sukmana, M. I. H., Cheng, F., & Meinel, C. (2020). CloudStrike: Chaos engineering for security and resiliency in cloud infrastructure. *IEEE Access, 8*, 123044–123060. https://doi.org/10.1109/ACCESS.2020.3007338
UK Department for Science, Innovation & Technology. (2024). Cyber security risks to artificial intelligence. https://www.gov.uk/government/publications/research-on-the-cyber-security-of-ai/cyber-security-risks-to-artificial-intelligence.
UK National Cyber Security Centre and US Cybersecurity and Infrastructure Security Agency. (2023). *Guidelines for secure AI system development*. UK Government. https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development.
US Cybersecurity & Infrastructure Security Agency. (2023). 2023 Year in review. https://www.cisa.gov/about/2023YIR.
USAIST. (2023). *NIST seeks collaborators for consortium supporting artificial intelligence safety*. NIST. https://www.nist.gov/news-events/news/2023/11/nist-seeks-collaborators-consortium-supporting-artificial-intelligence.
Usmani, U., Happonen, A., & Watada, J. (2023). Human-centered artificial intelligence: Designing for user empowerment and ethical considerations. In *2023 5th international congress on human-computer interaction, optimization and robotic applications (HORA)*. https://doi.org/10.1109/HORA58378.2023.10156761
Vassilev, A. (2024). *Adversarial machine learning: A taxonomy and terminology of attacks and mitigations*. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-2e2023. NIST AI NIST AI 100-2e2023.
Wach, K., Cong, D., Ejdys, J., Kazlauskaite, R., Korzynski, P., Mazurek, G., Paliszkiewicz, J., & Ziemba, E. (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review*, (2), 7–30. https://www.ceeol.com/search/article-detail?id=1205845.
White, House.. Executive Order: Removing Barriers to American Leadership in Artificial Intelligence. https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence.
Xu, H., Donovan, K., Franks, E. C., Lee, B. H., & Wong, M. (2023). *China's new AI regulations* (Vol. 3110). Latham & Watkins. https://www.lw.com/en/admin/upload/SiteAttachments/Chinas-New-AI-Regulations.pdf.
Zakrzewski, C., & Tiku, N. (2023). AI companies form new safety body, while Congress plays catch-up. *Washington Post*. https://www.washingtonpost.com/technology/2023/07/26/ai-regulation-created-google-openai-microsoft/.
Zohuri, B., & Rahmani, F. (2019). Artificial intelligence driven resiliency with machine learning and deep learning components. *Japan Journal of Research, 1*(5). https://journals.sciencexcel.com/index.php/jjr/article/download/8/5.