FOOD FOR THOUGHT

AI and safety management: an overview of key challenges

Eric Marsden and Véronique Steyer

n° 2025-03

TOPIC

Digital transition

Les Cahiers de la sécurité industrielle

FONCSI
Fondation pour une culture de sécurité industrielle

The *Foundation for an Industrial Safety Culture* (FonCSI) is a French public-interest research foundation created in 2005. It aims to:

- undertake and fund research activities that contribute to improving safety in hazardous organizations (industrial firms of all sizes, in all industrial sectors);
- work towards better mutual understanding between high-risk industries and civil society, aiming for a durable compromise and an open debate that covers all the dimensions of risk;
- foster the acculturation of all stakeholders to the questions, tradeoffs and problems related to risk and safety.

In order to attain these objectives, the FonCSI works to bring together researchers from different scientific disciplines with other stakeholders concerned by industrial safety and the management of technological risk: companies, local government, trade unions, NGOs. We also attempt to build bridges between disciplines and to promote interaction and cross-pollination between engineering, sciences and the humanities.

Foundation for an Industrial Safety Culture
A public-interest research foundation

www.FonCSI.org

6 allée Émile Monso – CS 22760
31077 Toulouse cedex 4
France

Email: contact@FonCSI.org

Abstract

Title Artificial intelligence and safety management: an overview of key challenges

Keywords artificial intelligence, digital transition, industrial safety

Authors Eric Marsden and Véronique Steyer

Publication date June 2025

Artificial intelligence based on deep learning, along with big data analysis, has in recent years been the subject of rapid scientific and technological advances. These technologies are increasingly being integrated into various work environments with the aim of enhancing performance and productivity. This dimension of the digital transformation of businesses and regulatory authorities presents both significant opportunities and potential risks for industrial safety management practices.

While there are numerous expected benefits, such as the ability to process large volumes of reliability data or unstructured natural language incident reports, the structural opacity of large neural networks, their non-deterministic nature, and their capacity to learn from new data mean that traditional safety assurance techniques used for conventional software are not applicable. Additionally, the expansion of the scope of automatable tasks and the gradual move towards work collectives that are composed of human operators who collaborate with various intelligent machines and agents introduce new variables that must be considered alongside and integrated with the organizational and human factors of safety.

What are the main challenges posed by these new technologies in terms of skills management, worker well-being, privacy protection, and the pursuit of performance that aligns with societal expectations? What changes are required in how we conceptualize the safety of high-stakes activities, how we demonstrate and verify the absence of unacceptable risks, and anticipate potential deviations?

This document provides a concise overview of the most recent available information, contextualized by decades of research on automation in high-hazard systems. It focuses specifically on the projected impacts for high-hazard industries and infrastructures over the next ten years.

The original French version of this document, published in May 2025, is also available from FonCSI's website (DOI: 10.57071/iae289).

About the authors

This document is a first output from the strategic analysis group run by FonCSI on the impact of the digital transition on safety management practices. It was authored by Eric Marsden, programme manager at FonCSI, and by Véronique Steyer, professor in the Innovation and Entrepreneurship department of the École Polytechnique, who are the coordinators of the strategic analysis.

The semi-automated translation from French to English was reviewed by Eric Marsden.

To cite this document

Marsden and Steyer (2025), *Artificial intelligence and safety management: an overview of key challenges*. Number 2025-03 of the *Cahiers de la Sécurité Industrielle*, Foundation for an Industrial Safety Culture, Toulouse, France (ISSN 2100-3874). DOI: 10.57071/iae290. Available from FonCSI.org/en.

Contents

Introduction

Objectives of the document

FonCSI is conducting a strategic analysis of **safety practices in the era of digital transition**. The strategic analysis group has focused on the impact of AI and big data on industrial safety management practices. One of the initial steps in conducting a strategic analysis involves a “zooming out” exercise aimed at providing a **broad perspective** on the issue at hand and the **main expected impacts**. This document presents the “big picture” for this analysis. It will be followed by other documents offering more detailed insights and illustrations of the emerging trends across different industry sectors.

This document provides a summary of the key questions identified regarding the potential impacts of intelligent machines and agents, and their interaction with human actors, on safety management practices. Our focus is on the implications for high-hazard industries with a 10-year outlook.

Research on AI development is progressing at a rapid pace; the topics addressed in this document are highly dynamic, and academic and expert knowledge is often still evolving. This document adopts a **forward-looking perspective**, covering certain subjects that are currently the focus of differing expert opinions. The fast pace of research in this field is not well aligned with the publication timelines of traditional academic journals. As a result, research findings are primarily disseminated through preprints on platforms such as arχiv and HAL. Therefore, we cite numerous preprints in this document, even though they have not undergone peer review.

Highlighting the risks of the rapid pace of academic work in this area, a preprint titled *Artificial Intelligence, Scientific Discovery, and Product Innovation* published on arχiv in November 2024 by a PhD candidate at MIT, which was cited in the original French version of this document published in May 2025, was “withdrawn from public discourse” by MIT and the PhD supervisors of the candidate shortly after our document was published, due to suspicions of fraud. The discussion of the results of this preprint are not present in this English version of the report.

Structure of the document

Chapter 1 provides a description of key elements of the **context** in which technological innovation in AI is taking place. This context is characterized by the exponential growth in the capabilities of AI models and systems, which are surpassing human performance across an increasing number of tasks and are being rapidly adopted by both individuals and businesses for a variety of applications. The sector is at the heart of major economic, military, and geopolitical stakes, which are driving massive investments. In many countries, the use of AI remains lightly regulated, and some industries are pursuing radical innovation strategies that have significant implications for safety management.

Chapter 2 briefly outlines several **challenges** associated with the use of AI in industrial contexts: the (controversial) existential risk to humanity posed by superintelligence, ethical risks, environmental impacts, major effects on human resource management and human-machine collaboration, and difficulties in applying traditional safety assurance methods to software incorporating AI models. These challenges are exacerbated by the rapid pace of technological development and adoption.

Chapter 3 presents an **impact analysis** of the introduction of AI-based tools on safety management in activities involving major accident hazards. It describes effects on the safety model, on safety management activities, on operational safety practices, on design and regulatory processes, and on the legal and social dimensions of safety.

1

The context

Artificial intelligence is a disruptive technology that simultaneously generates considerable benefits and significant risks. In this chapter, we briefly describe several features of the context in which this technological innovation is unfolding.

- A sustained pace of innovation, marked by a **dramatic increase in capabilities** in the ability of generative AIs[1] to handle queries in natural language, to interpret and generate images and video, and to reason by generative AIs:
  - As of February 2025, large language models (LLMs) achieve scores on benchmarks designed to assess reasoning and expertise capabilities that are significantly higher than those of human experts in each corresponding domain[2].
  - The latest connectionist AI models[3] were, by March 2025, able to perform software development tasks (with a 50% success rate) that typically require an hour of work for a human professional. This "automatable task duration" doubles every seven months (*cf.* figure 1.1).
  - Waymo's self-driving vehicles are reportedly involved in significantly fewer accidents than the general population of human drivers, with a 92% reduction in injury-related claims and an 88% reduction in claims for property damage[4].
  - The diagnostic performance of an LLM trained to conduct medical consultations in dialogue format reportedly surpasses that of specialist physicians across nearly all evaluated dimensions (medical history-taking, diagnostic accuracy, interaction management, communication skills, and empathy) **[Tu et al. 2025]**. The performance of the AI system alone is substantially higher than that of specialist doctors assisted by AI **[McDuff et al. 2025]**.

  This pace of innovation raises the question of when *artificial general intelligence* might emerge, which we define here (following OpenAI's usage, though the definition remains debated) as a highly autonomous AI system that achieves or surpasses human-level performance across the majority of wealth-generating cognitive tasks.

- Data collected by the Federal Reserve Bank of St. Louis (USA) (*cf.* figure 1.2) indicate that generative AIs are being adopted by individuals and employees at a faster rate than earlier transformative technologies such as the internet and personal computers. OpenAI CEO

[1] Generative AIs are models capable of producing text, images, videos, or other media in response to prompts. They are based on artificial neural networks organised in multiple layers ("deep"), trained on vast quantities of unlabelled data.

[2] Results on the "Google-proof Q&A Diamond accuracy" benchmark.

[3] Connectionist AIs are based on artificial neural networks inspired by the functioning of the human brain. They differ from symbolic AIs — so-called "expert systems" developed from the 1950s onwards — which aim to emulate human reasoning by applying established rules and knowledge. Recent AI progress is largely driven by connectionist models.

[4] Results from a study published by Waymo and the reinsurer Swiss Re in late 2024. Note that Waymo's autonomous vehicles currently complete over 150,000 trips per week.

[5] By "outsourceable task", we mean a task that can be clearly specified, whose successful completion can be easily evaluated, and which can be carried out autonomously and independently of other tasks. In the METR test, it is further assumed that task failure has negligible consequences (this point clearly limits the applicability of this study and its trend line to analysis of AI use in safety-critical systems). These tasks represent only a subset of actual work activities.

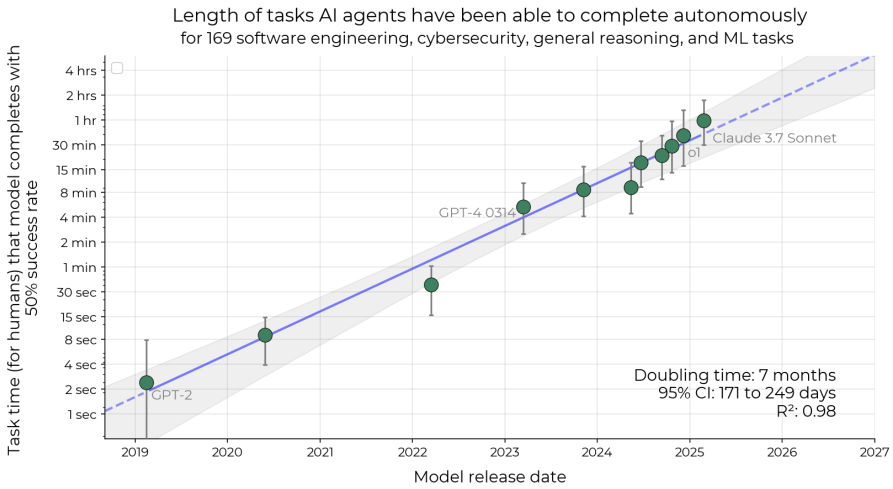

Figure 1.1 *Evolution of task duration (measured by the time required by a human professional to complete a software development task which is not safety-critical as an external consultant) that frontier generalist models are able to complete with 50% success. For models released over the past six years, this duration has doubled every seven months. If the trend continues — even assuming current duration estimates are tenfold overestimates — frontier models may, by 2028, autonomously complete a large share of computer-mediated, outsourceable tasks[5] that currently take humans days or weeks to accomplish.*

Source: report Measuring AI Ability to Complete Long Tasks, METR, March 2025, arχiv:2503.14499.

Sam Altman announced in April 2025 that 10% of the global population uses ChatGPT on a weekly basis.

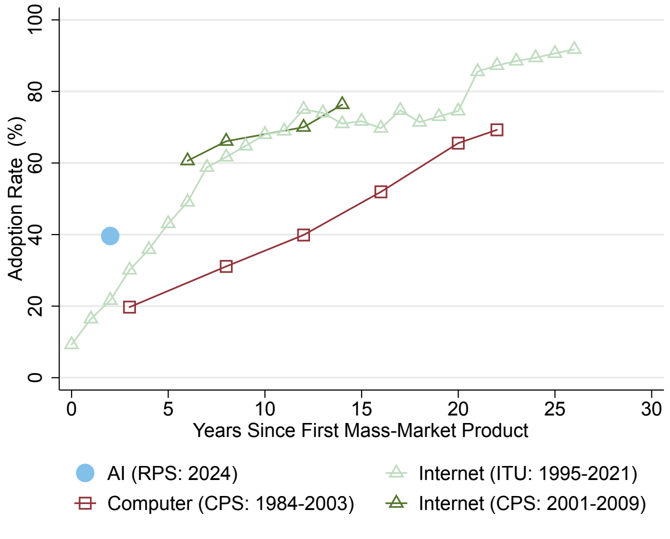

Figure 1.2 *Adoption trajectory of generative AIs compared to that of the internet and personal computers, according to a report by the Federal Reserve Bank of St. Louis (USA), February 2025.*

- ▷ Major economic and geopolitical stakes are fuelling **massive investments** in the scaling up of these models (over USD,109 billion in private investment in the USA in 2024[6]).

- ▷ Development in this field is extremely costly and largely beyond the reach of academic

[6] Source: *Artificial Intelligence Index Report 2025*, Stanford University Human-Centred Artificial Intelligence.

researchers[7], implying that most scientific and technical advances are currently being driven by the **industrial sector**, which favours a "**test and learn**" approach. This radical innovation philosophy extends to risk management practices in the deployment of these technologies, often described as "early release and iteration". This stands in sharp contrast to the development of critical systems, where risk assessment and safety demonstration are tightly integrated into functional design processes.

Various accounts suggest that there are frequent misunderstandings in the development of large systems that integrate AI components, particularly between AI development teams — who value innovation and rapid iteration (adopting Silicon Valley's "move fast and break things" ethos and treating failures as learning opportunities) — and safety and reliability engineering teams, whose professional culture tends to be more conservative.

- This development activity is concentrated in the USA and China (with Mistral being a rare European exception), as are the data and computing centres. Issues of **technological sovereignty** are expected to become increasingly prominent in strategic decision-making, with significant implications for "make or buy" trade-offs.

- A **near-absence of regulation** in certain application domains (such as autonomous driving in the USA and China, or the use of LLMs in many countries), or regulation that remains generic in nature (as with the AI Act in Europe). The very significant stakes — economic, military, and geopolitical[8] — suggest that **regulatory efforts** aimed at mitigating the risks of these technologies, which by their nature require collective action and international cooperation, face significant hurdles[9].

The challenge of regulating technological innovation

History shows that regulating technological innovation is a major challenge: the lack of feedback and historical experience makes it difficult to anticipate all types of harm that may result; such harm is often application-specific rather than an intrinsic outcome of the technology, implying the need to revise multiple sectoral regulations; and the strong asymmetry of information between the economic actors behind the innovation and the regulatory authorities makes it difficult for the latter to assess the risks.

Experience indicates that a regulatory framework based on cooperation with economic actors and on self-regulation is best suited to this type of situation [Black and Murray 2019]. However, several factors limit the potential for effective self-regulation by economic actors in the context of generative AI: (1) digital markets are often dominated by early entrants (a "winner takes all" effect), incentivising risk-taking; (2) the negative consequences of AI use are often externalities (in the economic sense) for the companies developing the systems; and (3) the harmful impacts are uncertain.

Although some **governance structures** for AI development do exist, such as the Summit for Action on Artificial Intelligence held in Paris in February 2025, their capacity to influence the development trajectory of frontier models[10] and their industrial applications remains limited. Some critics have described these initiatives as "aestheticised mannequins

[7] The leading supercomputer in March 2025, xAI's Colossus, is reported to have cost USD 7B. Industry's share of total AI compute rose from 40% in 2019 to 80% in 2025, while the public sector's share fell below 20%, according to an Epoch AI report [Pilz et al. 2025].

[8] It is worth noting that Vladimir Putin stated in a 2017 address to Russian schoolchildren that the country leading the race to develop AI would likely be the one to "rule the world".

[9] As an illustration, the second Trump administration in the USA cancelled the executive order on responsible AI development issued by President Biden in 2023, and reportedly informed European countries that it would implement sanctions to block any multilateral efforts at AI regulation.

[10] The term *foundation model* refers to AI models trained on large volumes of data that can subsequently be adapted (via fine-tuning or specialisation) to a wide range of downstream tasks and usage contexts. LLMs are a sub-category of foundation models. Some foundation models that pose specific risks — such as enabling circumvention of safeguards against CBRN proliferation, or creating the potential for loss of human control — are referred to as "frontier models".

of participation-washing”[11].

- Some “foundation” models are released as open source, which significantly undermines the ability of major industry players and their regulators to prevent socially unacceptable uses of these models and to manage the existential risks to humanity that they may pose.

- Alongside the development of LLM capabilities, there is a continued increase in the capacity for the collection, processing, and storage of **massive datasets**, accompanied by a proliferation of sensor types (especially cameras). The **early detection of technical anomalies**, the collection of data on failure modes and equipment ageing, are particularly valuable for maintenance management[12]. Such capabilities are also being deployed in applications for **procedural compliance monitoring** of frontline personnel and for real-time tracking of their exposure to risk: detection of lapses in vigilance, absence of personal protective equipment (PPE), exposure to cold or other hazardous conditions, and measurement of individual activity levels. These developments raise serious concerns about privacy and workplace intimacy, and are being trialled in particular in countries where the regulatory framework for personal data protection is underdeveloped.

[11] “*At present, these [governance] processes serve more to absorb than to transform: they collect criticism, dilute it in harmless reports, and then present these as evidence of action. They function as showcases of engagement, inviting the public to admire the complex mechanisms of AI governance while protecting existing power structures from scrutiny*”, as argued in the op-ed *Beyond the Façade: Challenging and Evaluating the Meaning of Participation in AI Governance* by Jonathan van Geuns, TechPolicy Press, February 2025.

[12] Relevant keywords include: prognostics and health management, condition-based maintenance. These refer to “narrow AI” systems, which are designed for a small number of specific tasks, in contrast to artificial general intelligence (AGI), which is capable of performing a wide range of task types.

2

The key challenges

The very rapid development of AI poses challenges of various kinds, which will need to be addressed by governments and regulatory authorities, by companies developing the models, companies deploying them for different applications, and by individuals. Five main challenges can be identified:

▷ An **existential risk to humanity** associated with superintelligence (AIs that achieve capabilities broad and powerful enough to rival humans in determining their mutual destinies). This issue of loss of control is controversial[1], yet it does not belong to the realm of science fiction, according to numerous expert studies.

Loss of control and p(doom)

The literature on this topic uses the keyword "p(doom)", the probability of a scenario involving catastrophic loss of control (*cf.* figures 2.1 and 2.2) for humanity. Estimates of this number are controversial and vary wildly, but it is worth noting that over 40% of experts estimate this probability to exceed 0.1 [Grace et al. 2024]. Geoffrey Hinton (the most cited computer scientist) stated in 2024:

> *I can't see a path that guarantees safety. We're entering a period of great uncertainty where we're dealing with things we've never dealt with before. And normally, the first time you deal with something totally novel, you get it wrong. And we can't afford to get it wrong with these things. [...] If you take the existential risk seriously, as I now do, it might be quite sensible to just stop developing these things any further [...] it's as if aliens had landed and people haven't realized because they speak very good English.*

Tech industry leaders wrote in 2023: "*Mitigating the risk of extinction from A.I. should be a global priority alongside other societal-scale risks such as pandemics and nuclear war*". More recently, see for example the article *Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development*, 2025. The threat was also highlighted at the 2025 Davos meetings by several leaders of major AI firms, including the CEO of Google DeepMind, Demis Hassabis, the co-founder of Anthropic, Dario Amodei, and the researcher Yoshua Bengio.

At present, there exists no regulator with the authority to demand structured demonstrations of the absence of existential risk to humanity from new AI models, and the global coordination necessary to organise such an oversight effort appears currently out of reach. The economic competition among firms in the sector, and military rivalry between states, imply that the management of this risk is relegated to a secondary concern.

[1] Some analysts argue that apocalyptic narratives are used strategically by major tech firms to divert attention from more immediate and concrete ethical concerns, such as discrimination and social inequality [Stilgoe 2024; Wong 2023], or alternatively to promote ideologies such as transhumanism [Beaudouin and Velkovska 2023]. While such efforts do exist, they appear to have little effect on public perceptions of AI risks. A recent study conducted in the USA found that respondents are significantly more concerned about the immediate risks posed by AI than about existential risks, and that presenting information about existential risks does not reduce concerns about immediate harms [Hoes and Gilardi 2025].

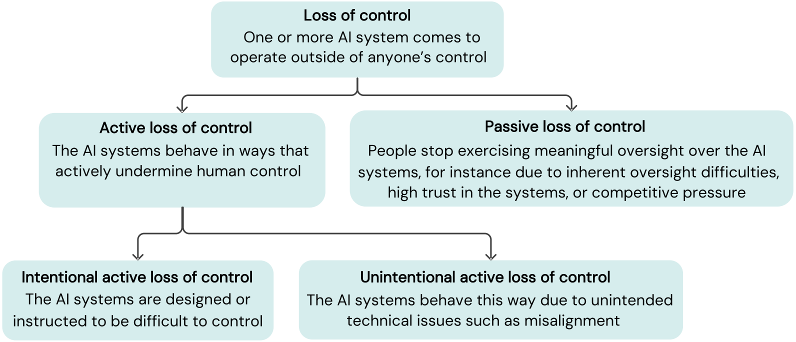
Loss of control
One or more AI system comes to operate outside of anyone's control
Active loss of control
The AI systems behave in ways that actively undermine human control
Passive loss of control
People stop exercising meaningful oversight over the AI systems, for instance due to inherent oversight difficulties, high trust in the systems, or competitive pressure
Intentional active loss of control
The AI systems are designed or instructed to be difficult to control
Unintentional active loss of control
The AI systems behave this way due to unintended technical issues such as misalignment

Figure 2.1 *Different scenarios of "loss of control", depending on whether the AI system actively seeks to weaken human oversight, and whether this behaviour is intentional or not.*

Diagram reproduced from the first International Scientific Report on the Safety of Advanced AI, produced by a group of experts appointed by 30 countries, the OECD, the EU, and the UN [Bengio et al. 2025].

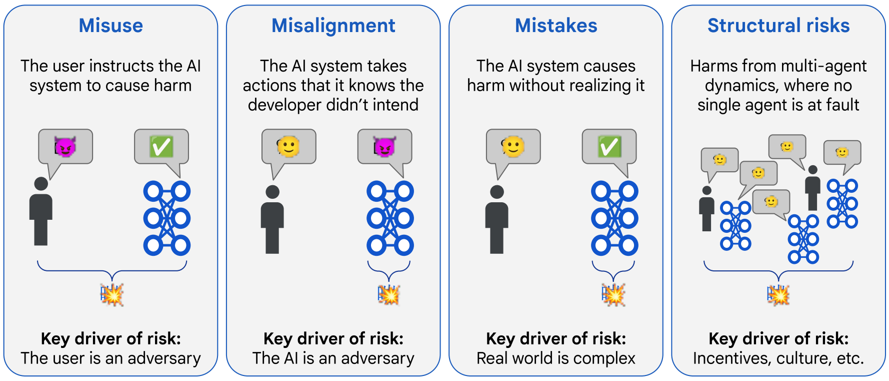
Misuse
The user instructs the AI system to cause harm
Key driver of risk:
The user is an adversary
Misalignment
The AI system takes actions that it knows the developer didn't intend
Key driver of risk:
The AI is an adversary
Mistakes
The AI system causes harm without realizing it
Key driver of risk:
Real world is complex
Structural risks
Harms from multi-agent dynamics, where no single agent is at fault
Key driver of risk:
Incentives, culture, etc.

Figure 2.2 *Four classes of scenarios in which an AI causes harm. The scenarios are grouped according to the risk mitigation approaches that can be implemented. For example, malicious misuse and misalignment differ based on the identity of the malicious agent, as mitigation strategies applicable to malevolent humans differ greatly from those appropriate for malevolent AIs. The fourth category of structural risks includes the progressive loss of human capabilities due to increasing dependence on AI assistance.*

Diagram excerpted from a Google DeepMind technical report [Shah et al. 2025].

Academic and expert work on this subject uses the term "AI safety" to refer to this eschatological issue. Model developers assess the extent to which models are "aligned" with "human values". Regulatory challenges in this area will involve both governments (through "AI safety" agencies) and companies (by developing responsible AI use strategies and associated self-regulation mechanisms, which will become a concern for executive boards).

Aligning AI with human moral values

One of the major challenges in aligning models with human values lies in specifying what moral values are shared within a given human community — or by humanity as a whole. The most frequently cited values and ethical principles (for example, in the report on trustworthy AI authored by the expert group convened by the European Commission [EU HLEG AI 2019]) include beneficence, non-maleficence, respect for human agency and oversight, justice, and transparency. These principles are abstract in nature. Some researchers operationalise these values via the preferences of a human, as defined by decision theory, while acknowledging that some individuals display asocial or anti-social preferences. Others maintain that behaviour properly aligned with human expectations is necessarily dependent on the usage context and the role assigned to the AI agent [Zhi-Xuan et al. 2024].

As a practical example, semi-autonomous vehicles must navigate moral dilemmas in accident scenarios, such as choosing between preserving the lives of vehicle occupants, cyclists, or nearby pedestrians. Experimental psychology research has shown that individual preferences regarding which lives to save — based on factors such as social status, age, gender, and law-abidingness — vary significantly across national cultures [Awad et al. 2018]. At a more institutional level, the preferences of vehicle manufacturers, associations representing different categories of road users, safety authorities, their ethics advisors, and elected officials are far from aligned.

Other risks are associated with the development of lethal autonomous weapon systems (LAWS), such as drones equipped with autonomous target identification and destruction capabilities[2], particularly their potential to terrorise a population — or a targeted subset — at relatively low operational cost.

▷ Various **ethical risks**, such as decision-making biases (affecting certain ethnic, cultural, or gender groups, due to models reproducing — and sometimes amplifying — the statistical correlations present in the data used to train them [O'Neil 2016; Eubanks 2018]), and the fact that algorithmic decision-making in public administration and financial services often offers limited avenues for recourse [Défenseur Droits 2024]; the amplification of social inequalities[3]; threats to democratic functioning, particularly from the generation of factually incorrect content; the erosion of the social value attributed to expertise; and issues concerning the protection of privacy.

▷ A **non-negligible environmental impact** linked to electricity consumption (data and computing centres could account for 4% of total energy consumption by 2030, and growing demand may compete with other electrification projects; *cf.* figure 2.3), as well as to water usage and electronic waste.

▷ Issues related to **human resource management** and **co-intelligence** [Mollick 2024]: recruiting workers with **appropriate technological skills**, managing career transitions, **reskilling**, managing the end of certain professions, and change management[4], and the redefinition of operating procedures, for instance. Skills are both a strategic question (can we afford to outsource this expertise?) and a geostrategic one, as the largest pools of talent are located in India and China. Issues of continuing education (reskilling, upskilling,

[2] Aerial drones capable of autonomously identifying and attacking targets have, for instance, been deployed in the war in Ukraine.

[3] Technological revolutions often produce social inequality before they lead to progress. For example, the introduction of mechanised looms enriched a small number of factory owners while simultaneously reducing the professional autonomy of weavers, increasing their working hours, and worsening their living conditions. In contrast, industrialisation in Western countries between 1950 and 1980 led to broadly shared prosperity. The difference lies mainly in political and social institutions, and the distribution of economic and social power, which determine how benefits are allocated — according to the 2024 laureates of the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel [Acemoglu and Johnson 2023]. These economists suggest reducing current incentives to replace human labour with machines (e.g., payroll taxes), and increasing incentives to create new tasks and skills that incorporate AI.

[4] A January 2024 IMF study estimates that 60% of jobs in advanced economies are exposed to AI, meaning that some of their tasks could be automated or transformed to be assisted by AI.

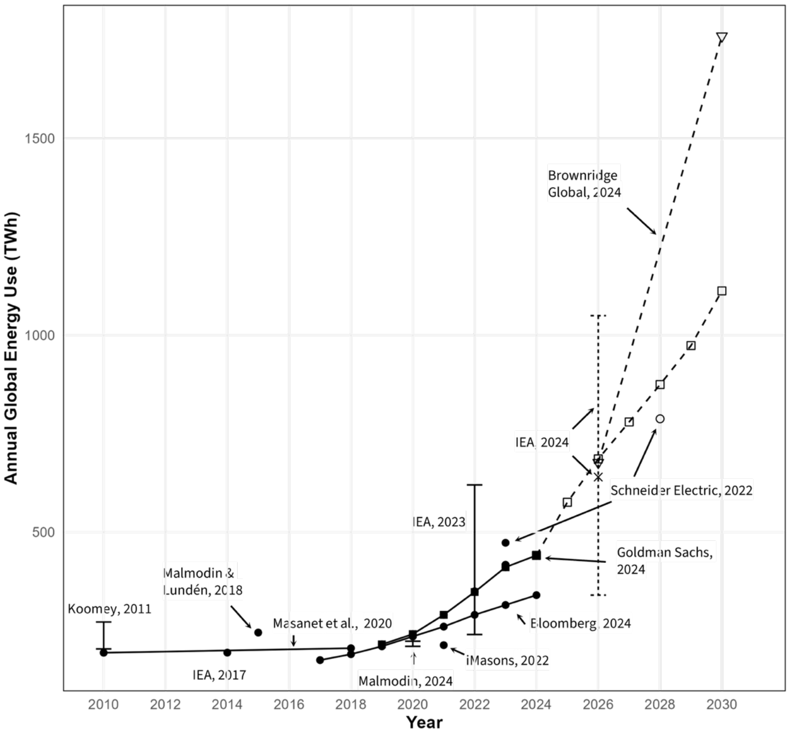

Figure 2.3 *Energy consumption of data and computing centres at the global level. Historical estimates are indicated by solid lines and projections by dotted lines.*

Source: 2024 United States Data Center Energy Usage Report, Lawrence Berkeley National Laboratory.

skills matching) are critical[5]. Talent management will be especially crucial in high-risk industrial sectors, where conservative professional cultures may appear unattractive to potential candidates seeking careers at the forefront of technological innovation.

Work teams will become hybrid (partially composed of AI agents), and AI resource management (with AI agents specialised by task type, application domain, and budget constraints) will need to be coordinated with human resource management. Addressing psychosocial risks and workplace well-being may extend to AI agents themselves[6].

▷ On a more technical level, the structural opacity (the "**black box**" nature) of AI models — especially those composed of large neural networks and trained via deep learning — makes it very difficult to understand and explain why a model produces a specific output **[Burrell 2016]**. This limitation raises legal issues, particularly in implementing the right to an explanation of decisions made on the basis of AI outputs[7]. It also poses a significant obstacle to the **safety demonstration** for functions performed by AI-based software and to their **certification for critical applications**. Various strands of work on "explainable AI[8]" and on intelligibility aim to address these challenges **[Frejus et al. 2022]**.

[5] Monitoring indicators from the EU's "2030 Digital Decade" plan concerning the number of "ICT Specialists" fall short of the targets set in 2021. On this topic, see also the work by the Foncsi analysis group on future skills toward 2040.

[6] Although this may sound like an attempt at humour, a line of research on **AI welfare** has been developing since 2024, arguing that even if there is uncertainty about the current or future moral consciousness of AI models, the nature of the issues justifies further study **[Long et al. 2024]**. In 2024, the company Anthropic hired an AI welfare specialist. In experimental settings, humans tend to avoid exposing robots to acts they would consider abusive if directed at humans: for instance, they display physiological discomfort when seeing a robotic baby dinosaur being hit **[der Pütten et al. 2013]**, or when a tower built by a robot — indicating that it "cares" about the structure — is deliberately destroyed **[Darling et al. 2015]**.

[7] The right to an explanation is enshrined in the European General Data Protection Regulation (GDPR) for automated decisions, and — for AI systems classified as "high risk" — in the AI Act.

[8] According to the French data protection authority (Cnil), explainability is "*the ability to relate and make understandable the elements taken into account by the AI system in producing a result. This may include, for example, the input variables and their consequences on a score prediction, and hence on the decision*" **[Maudet et al. 2022]**.

Industrialised societies struggle fully to anticipate the possibilities and impacts of these technologies. One may fear that this anticipation is insufficient given the considerable impacts expected in the next few years. In particular, the impact on the world of work — on the nature and quality of jobs, and on the skills required — is likely to be significant[9], yet it remains poorly anticipated. It is worth noting that only 23% of AI researchers in Europe believe that AI should be developed as rapidly as possible **[O'Donovan et al. 2025]**.

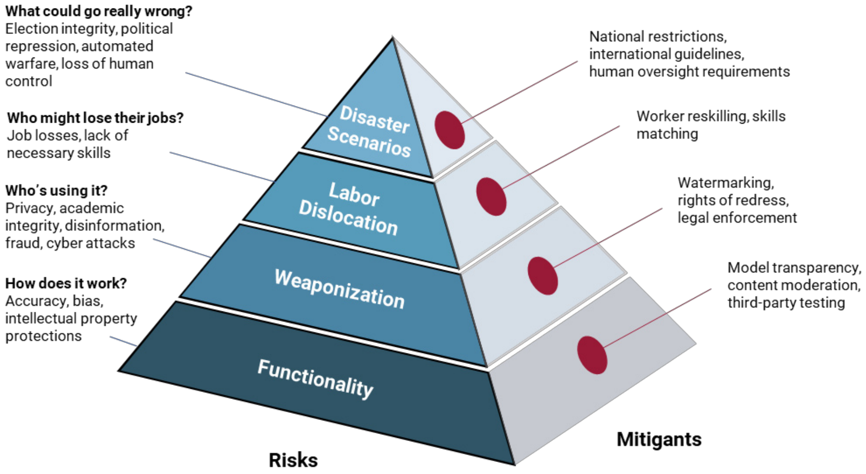
What could go really wrong?
Election integrity, political repression, automated warfare, loss of human control
Who might lose their jobs?
Job losses, lack of necessary skills
Who's using it?
Privacy, academic integrity, disinformation, fraud, cyber attacks
How does it work?
Accuracy, bias, intellectual property protections
Disaster Scenarios
Labor Dislocation
Weaponization
Functionality
National restrictions, international guidelines, human oversight requirements
Worker reskilling, skills matching
Watermarking, rights of redress, legal enforcement
Model transparency, content moderation, third-party testing
Risks
Mitigants

Figure 2.4 *Four main families of risks associated with AI use, arranged by the time horizon of their emergence and by their potential impact. Possible safeguards for each category are presented.*

Extract from a January 2025 report by Goldman Sachs[10].

Within organisations, addressing these challenges will require mobilising a wide range of competencies beyond those of technical experts: development strategy, legal departments, human resources, cybersecurity, information systems, and compliance.

[9] This forthcoming social impact is regularly highlighted by the CEOs of companies developing AI models. For instance, Sam Altman, CEO of OpenAI, stated in 2023 (interview with ABC News upon the release of the GPT-4 model): *I think over a couple of generations, humanity has proven that it can adapt wonderfully to major technological shifts. But if this happens in a single-digit number of years, some of these shifts… That is the part I worry about the most.* In February 2025, he added on his blog: *In particular, it does seem like the balance of power between capital and labor could easily get messed up, and this may require early intervention.*

[10] *AI/data centres' global power surge: Five drivers of upside/downside and the Reliability investment tailwind*, Goldman Sachs Research, January 2025.

3

Implications for the management of industrial safety

The introduction of AI-based tools will have significant impacts on the management of safety in major accident hazard industries. In this chapter, we list various impacts on the safety model, on safety management activities, on operational practices, on design and regulatory activities, and finally on the legal and social dimensions of safety.

3.1 The safety model

The growing use of AI will have implications for how we conceptualise the **safety model**:

▷ Systems in which most safety functions are performed by automation (intelligent software components associated with robotics) will become more widespread. The necessity of human presence in certain high-risk systems will be called into question. However, experience shows that designers often underestimate the complexity of interactions between intelligent machines and their collaboration with human actors (neglecting human-machine interactions and placing excessive faith in the benefits of complete autonomy constitute two of the seven "lethal myths" concerning autonomous systems described by [Bradshaw et al. 2013]).

This "techno-solutionist" trend [Morozov 2013] tends to increase the "distance" between the technical system and society, reducing the number of people who possess in-depth understanding of the system's functioning and safe operational conditions, and heightening the propensity to adopt a "magical" view of digital technologies[1]. As a result, we may assume that residual disasters will be more severe, as their early warning signals will less frequently be detected and managed in advance; they will also be more unexpected, since technological complexity hinders democratic deliberation on safety-related issues.

Our "magical" relationship with digital technologies

Owing to their complexity, sophistication, and inscrutable nature, new technologies are sometimes perceived as magical [Gell 1988]. Enthralled by the textual and graphical outputs of LLMs, one may attribute to them capacities and agency that exceed their actual properties, and develop an intuitive mental model of their way of operating that does not correspond to reality. Tech companies sometimes exploit this effect to encourage users to develop quasi-religious beliefs in the transformative power of AI, thereby avoiding accountability for various negative consequences of its development and use [Nagy and Neff 2024]. Some authors interpret the use of big data and algorithms as a means of exercising control (surveillance, optimisation, discipline) over societies without being held to account. The explanation "the computer said so", and the level of praise of the disruptive powers of generative AI in certain public statements made by powerful individuals, are seen by some as revealing a form of "patriarchal populism of the elites" [Vesa and Tienari 2020].

[1] New technological tools tend to exert an influence over people's practices and mindsets that exceeds what is justified by their technical performance. For example, DNA tests are often used in judicial investigations without the caution warranted by the technical limitations of this form of evidence. The challenge is to ensure that, in this shift towards a society of maximal security [Marx 1988], technology serves the system without enslaving it.

Key issue

The mythologised efficiency of AI. In recent years, a belief has emerged in the capacity of AI to improve the efficiency and effectiveness of public administrations (illustrated in 2025 in particular by the "Department of Government Efficiency" in the USA) and that of businesses. This belief — that technology can durably solve complex social problems — is contradicted by numerous studies which show that AI systems frequently generate errors and biases, require significant human support for training, for their integration with existing infrastructures, and for the correction of their mistakes [Mateescu and Elish 2019].

Successful integrations require: (1) thinking about and incorporating human roles and expertise from the system design stage [Baxter and Sommerville 2011]; (2) building on actual work practices (including their constraints, tacit know-how, modes of mutual assistance and coordination, invisible tasks, etc.) rather than relying solely on prescribed work [Lammi 2021].

- ▷ A significant transformation in the **distribution of functions between humans and machines** can be expected, with increasingly complex and sophisticated tasks being handled by automated systems.
- ▷ The issue of **coordination and collaboration between human agents and intelligent machines**, long studied by cognitive science researchers (notably in the body of research on "joint cognitive systems"), will become critical. It will be necessary to move beyond the traditional "humans are better / machines are better" dichotomy[2], which seeks to optimise the respective functions of humans and automation separately. Instead, a more holistic approach to the performance of the entire sociotechnical system is required. Optimal system performance is often not achieved by combining the best-performing automation (operating in isolation) with the most capable individuals working solo [Behymer and Flach 2016].
- ▷ **New threats** are emerging: high-risk systems will become more exposed to cyberattacks. The systemic nature of risks will increase. Errors generated by generative AIs differ significantly from cognitive errors made by human operators: they are less predictable (not explainable, for instance, by "fatigue") and are not accompanied by expressions of doubt or uncertainty. The detection and compensation mechanisms we have put in place to tolerate such errors will often need to be rethought.

Example

Concerns over systemic risk in the financial sector

In the financial sector, the increase in systemic risk caused by the use of AI is already being acknowledged by authorities, who are concerned about scenarios in which AI agents could destabilise financial markets. A recent report by the Bank of England notes that companies are deploying autonomous neural networks whose risks are not well understood by risk managers.

As stated in the report *Financial Stability in Focus: Artificial Intelligence in the Financial System*, published in April 2025,

> " *Risk management of these positions is made more challenging by the lack of interpretability of neural networks, as actions may be unpredictable and the reasons for the positions may not be well understood by human risk managers at the firm. In addition, models with sufficient autonomy could act in ways that are detrimental to the overall stability or integrity of markets, for example by ignoring regulatory or legal guardrails such as market abuse regulations.*

Accidents may be triggered by the phenomenon of "model drift", which occurs when a learning component trained in a given environment is then deployed in another environment with different characteristics it is not equipped to handle (this new environment may simply be the original one, altered over time by operational drift).

[2] This dichotomy, often abbreviated HABA/MABA, is prevalent in the literature on automation.

Collision between an autonomous vehicle and an articulated bus

Example

In 2023, an autonomous vehicle operated by the company Cruise collided with an articulated bus in San Francisco, fortunately without casualties. The company reported that the vehicle's driving software had incorrectly anticipated the behaviour of the bus's two segments and had disregarded data from the vehicle's LIDAR sensor. The software had primarily been trained on data from a city in which no articulated buses were in operation [Cummings 2023].

3.2 Safety management activities

The growing use of AI will have impacts on **safety management activities**:

- The ability of neural networks powered by massive datasets to enhance predictive capabilities offers many opportunities to improve predictive maintenance and structural health monitoring. This is likely the type of application where AI has so far made the most significant contribution to industrial safety.

Improved reliability at SNCF Voyageurs

Example

The use of predictive maintenance techniques, based on the analysis of data collected by sensors embedded in trains, allows SNCF Voyageurs (the main French railway operator) to halve the number of breakdowns occurring during operations and to reduce by one third the number of units taken out of service for maintenance.

- The processing of **big data** offers numerous possibilities for improving risk management (e.g. analysis of incident reports to extract categories and identify anomalies, extraction of performance indicators from unstructured data).

The Data4Safety project in civil aviation

Example

The "safety intelligence" project Data4Safety, led by EASA, aims to identify and qualify systemic risks and mitigation strategies for civil aviation in Europe. It processes safety reports, flight data from airlines, air traffic data from flight management organisations, and meteorological data.

- Smart cameras offer real-time **surveillance and anomaly detection** capabilities that open up numerous possibilities in terms of safety, such as real-time detection of failure to wear PPE, non-compliance with task sequencing, and monitoring of drivers' attention. These applications raise significant issues concerning privacy and workplace intimacy.

Monitoring operator fatigue with a smart headband

Example

A headband equipped with brain activity sensors (electroencephalogram-based equipment) has been used for about a decade to monitor the fatigue level of machine operators in the Australian mining industry. The tool alerts the wearer when the estimated fatigue level is high and can transmit this information to management and record it in a centralised database[3].

- **Digital twins** offer **simulation** capabilities that can be used to support the design of new facilities (analysis of maintainability, constructability, ease of operation during the design phase), as well as the training of future operations teams.

- Automated systems, robots, and **cobots** are taking over dangerous and strenuous tasks in high-hazard facilities, reducing frontline workers' exposure to toxic substances, extreme temperatures, ionising radiation, confined spaces, hazardous machinery, and musculoskeletal disorders [ILO 2025].

[3] Source: case study *Smart digital systems for improving workers' safety and health: smart headband for fatigue risk-monitoring*, EU-OSHA, 2024.

- Given the rapid pace of technological development in AI, no dedicated industry for the development of "safety-assured" components has emerged. The ever-increasing cost of developing frontier models will likely lead to a consolidation of the leading industrial players (similar to the phenomenon observed among engine manufacturers in Formula 1). Integrators rely on commercial off-the-shelf (COTS) hardware and software components, including for **safety-critical applications**, with very limited ability to influence the software development processes, model training, and validation of the non-malicious nature of AI agents.

3.3 Supervising system operations

The growing use of AI will impact **safety practices** related to the **operation and control** of critical installations, systems, and infrastructures:

- A decline in human operators' ability to control the system as automation increases, with the human role shifting from direct control to supervision. Numerous academic studies show that increasing reliance on sophisticated automation leads to a decline in operational skills and frontline operators' understanding of system functioning[4]. Operators' ability to detect anomalies and take over from automation decreases, while dependence on technology grows[5]. The range of competencies exercised narrows over time. New failure modes emerge, such as misunderstanding the operating mode of the automation ("mode confusion").

- Excessive **reliance on AI** in decision-making can lead to:
  - frustration, particularly when the recommendations provided by the machine are unintelligible to its users [Kellogg et al. 2020];
  - passivity: individuals are more likely to accept the machine's recommendations without questioning them [Bader and Kaiser 2019];
  - emotional detachment: individuals may feel less responsible for decisions, with automated tools acting as moral buffers [Cummings 2006].

  Excessive human trust in automated systems' recommendations, or automation bias, is well-documented in psychology (*cf.* for example [Busuioc 2021]).

3.4 System design and safety oversight activities

Industrial stakeholders and authorities in Europe highlight a need to train and **recruit engineers** with relevant **technical and scientific expertise**. These skills are often found among subcontractors and service providers of large industrial groups rather than in-house. Public authorities looking to hire for their AI oversight bodies offer salaries that are significantly lower than those in the private sector.

[4] This is the "the computer said" effect. See, for instance, Nurski, Laura, and Mia Hoffmann. *The impact of artificial intelligence on the nature and quality of jobs*. Bruegel Working Paper, 2022.

[5] A recent study specifically on the impact of generative AI on knowledge professions suggests that it significantly weakens critical thinking skills. "*[A] key irony of automation is that by mechanising routine tasks and leaving exception-handling to the human user, you deprive the user of the routine opportunities to practice their judgement and strengthen their cognitive musculature, leaving them atrophied and unprepared when the exceptions do arise*" [Lee et al. 2025].

Private sector attractiveness difficult to match

Example

The *European AI Office*, responsible for drafting sectoral standards for implementing the AI Regulation, listed open positions on its website in 2025. These included roles for AI specialists with a Master's degree and at least one year of experience, with an annual salary of 50 k€. In the UK, the department in charge of innovation and technology is seeking to recruit a head of AI safety activities with a salary of £76k per year. According to websites tracking salaries at major AI companies, median compensation at OpenAI is around USD 500k, and Meta was reported in June 2025 to be offering annual compensation packages of up to 100M USD to prominent AI researchers.

Moreover, private-sector firms can offer candidates privileged access to cutting-edge technical resources necessary to develop and test models (large volumes of annotated data, massive computing capacity), along with in-house expertise that remains out of reach for public institutions.

Concerns about **strategic autonomy** have been raised due to technological dependence on components — and the associated expertise required to understand, operate, control, adapt, maintain, and more generally master them — developed in China and India (with more recent concerns involving the USA).

3.5 Legal and social impacts

The increasing use of AI will have **legal and social** ramifications:

▷ For human operators, we are witnessing a generalisation of the "Authority–responsibility double bind" already highlighted by D. Woods in 1985 [Woods 1985]: operators are held accountable for the outcomes of system management, yet they lack the necessary authority, capabilities, and means to fully control it. This skewed relationship leads to role confusion and a decline in operators' trust in the automated system, with implications for safety. Experience over recent decades shows that this situation is exacerbated by user interfaces that fail to make the internal workings of the system visible. One of the defining features of deep learning models based on large neural networks is their very low interpretability, their capacity to explain the reasoning behind a given output. Numerous research initiatives are now emerging in the field of "human-centred explainable AI", a key area of inquiry identified by FonCSI's strategic analysis of the safety impacts of the digital transition.

Key issue

While there is a tendency to conflate **understanding of internal operation** (i.e., interpretability) with **trust** in a system, academic research indicates that the relationship between these two concepts is complex: better understanding does not necessarily lead to more calibrated trust. Overly technical and detailed explanations of how a system functions can paradoxically erode user trust by highlighting system limitations or overwhelming them with excessive detail. Experience suggests that the aim should be to foster *well-calibrated* trust, meaning trust that aligns with the system's actual capabilities. The factors contributing to such trust are multifaceted and complex, but they appear to rely more on the user having a relevant mental model of the system's functioning, on an interaction style that encourages reflection, and on understanding the goals pursued by system developers and integrators, than on detailed technical comprehension [Mehrotra et al. 2024]. Nevertheless, users remain concerned with understanding the internal workings in more critical usage contexts. Explanations that are too complex or poorly tailored to the recipient's level of technical knowledge tend to reduce trust rather than enhance it.

As the capabilities of intelligent machines increase, we are moving from architectures in which a human operator supervises an automated system ("supervisory control" in the automation literature) to situations of cooperation and coordination between human and digital agents, involving horizontal rather than vertical control. Trust thus becomes a **relational** and **dynamic** process, grounded in quasi-social mechanisms, rather than a one-directional and static judgement [Chiou and Lee 2023].

▷ The integration of AI-based software components into safety-critical systems leads (or should lead) to a **shift in liability** for accidental harm, from operators or pilots to system operators. The liability chain can extend to designers, integrators, and suppliers of AI components (with various companies specialising in model development, adaptation to specific problems such as localisation in an industrial environment, data collection for training, and data labelling[6]).

The issue of civil liability for entities that deploy or market AI-based systems has been the subject of lively political debate in Europe since mid-2023, as outlined in the following box.

The turbulent life of the European "AI Liability" Directive

The content of the European directive or regulation on "AI Liability" and its relationship with the requirements of the AI Act (2024) and the revised Machinery Directive (2023) were the subject of intense negotiations between 2023 and 2025. Traditional liability regimes based on negligence appear ill-suited to managing the harms caused by AI-based systems, given their complexity and opacity (the "black box" effect makes it difficult for victims to document and demonstrate the causal link between a design choice and the damage incurred, and also means that a regime based on regulation and *ex-ante* safety demonstrations is difficult to implement). A regime of "strict" or no-fault liability — whereby the operator or vendor of a system incorporating AI components that caused damage would be automatically held liable, unless the victim's negligence could be proven — would offer better protection to victims (this is the historical approach in product safety law). However, such a regime may hinder innovation. See, for instance, the impact study published in September 2024: Proposal for a directive on adapting non-contractual civil liability rules to artificial intelligence: Complementary impact assessment. This report proposes applying strict liability to high-risk systems as defined by the AI Act, to general-purpose AI systems, to sectors historically regulated under the "Old Legislative Framework" (e.g., transport, Seveso-classified industries), and to the insurance sector. It also suggests recognising the joint liability of operators, system designers, and all companies within the supply chain of such systems.

The European Commission announced the withdrawal of the "AI Liability" directive proposal following threatening statements made by the Vice President of the United States at the AI Summit in Paris in February 2025.

▷ **Redefining the very notion of "work"**, with notable consequences for quality of working life. This issue is already a concern in many Western countries. A substantial body of literature highlights risks of dehumanisation, "datafication", loss of autonomy, psychosocial risks (such as "technostress[7]"), invasions of privacy [Pasquale 2015], reduced job quality, and erosion of workers' bargaining power [Gmyrek et al. 2023].

Collaboration between human actors and AI systems — often presented as a more desirable and worker-friendly alternative to full automation — can negatively impact workers' **psychological safety** (emotional fatigue, cognitive overload arising from the delegation of simpler, "breathing space" tasks to machines) and physical safety (postural changes, loss of alertness due to digital tool usage) [Waardenburg 2024].

These problems are emerging at an unprecedented pace, raising concerns about society's ability to adapt.

[6] See, for example, *Allocating accountability in AI supply chains: a UK-centred regulatory perspective*, Ian Brown, Ada Lovelace Institute (2023), as well as [Martin 2019].

[7] Technostress is a form of stress caused by the difficulty of adapting to new digital technologies, including the expectation of increased employee productivity, difficulties in understanding certain tasks, and uncertainty about how AI systems function, particularly due to their high rate of change [Rohwer et al. 2022].

3.6 Safety management activities

Regarding applications related to **safety management**:

- Applications in the field of **autonomous driving** are progressing rapidly, with quite remarkable success to date (damage to property and people is reportedly ten times lower than that caused by human-driven vehicles, according to a Waymo/Swiss Re report from December 2024). In the mining sector, autonomous ore transport trucks and trains, as well as autonomous drilling robots, have been deployed by Rio Tinto in its mines in Pilbara (Western Australia), with highly positive results in terms of safety[8].

- Few applications exist for **critical functions** in sectors with a longstanding use of **certification procedures** for safety-critical subsystems (aeronautics, nuclear energy, railways). Certification of software components and subsystems historically relies on assumptions that are poorly suited to modern AI models: deterministic and temporally stable behaviour (which is not the case for adaptive learning systems); traceability of the software development process; the ability to characterise and analyse the types of data processing performed; and more generally, a complete and detailed understanding of possible behaviours and failure modes. The box on page 21 outlines ongoing reflections on this certification challenge in the context of safety-related functions.

Key issue

> It is important to note that the safety implications of introducing autonomous agents or AI-based decision-support tools are more significant than for traditional software components: AI-based agents possess agentive capabilities not found in conventional software (they should be viewed more as proactive actors than as passive tools), can interact with one another in unpredictable ways, and may change behaviour based on past learning. Their powerful abilities, multifunctional nature, conversational interaction capabilities with humans, potential for "off-label use", and the human tendency to anthropomorphise AI agents imply that they have, or will have, a stronger influence on human agent behaviour than traditional software does. By nature, AI agents are *tightly coupled* and introduce *interactive complexity*, two organisational characteristics known to make systems prone to catastrophic failure [Perrow 1984].
>
> Therefore, any reflection on the safety challenges posed by AI must consider not only AI used in safety-related functions and as decision-support tools, but also AI agents involved in seemingly non-safety-related roles. Such reflection should adopt a **sociotechnical perspective**, recognising that risks emerge from interactions between human and intelligent agents at the scale of the overall system, and may evolve through gradual adaptation and **co-evolution** of agents and their environment [Baxter and Sommerville 2011]. This perspective is currently underrepresented in academic work on AI systems, which predominantly views AI as technical components that can be analyzed in isolation from context in which they are used.

- A gradual integration of AI-based tools is underway in sectors where safety relies little on formal safety demonstrations and certification, such as healthcare. High-hazard systems are also affected, at least for functions that are not classified as being safety-critical. Technologies are typically introduced as **decision-support tools**, used under the responsibility of a trained professional. These deployments often underestimate the phenomenon of **gradual habituation**, whereby users become increasingly dependent on the automation over time[9].

[8] For example, the number of near-miss vehicle collisions in "autonomous haulage operations" is ten times lower than in human-operated haulage operations at Rio Tinto sites in Australia.

[9] Experience from the progressive introduction of automation in the aviation sector shows that pilots who have never flown without these support systems find it difficult to operate without them.

Example

Decision-support tool for document search at Diablo Canyon

The company PG&E, which operates the Diablo Canyon nuclear power plant in the USA, announced the implementation of a generative AI-based tool to help operators and maintenance technicians access the regulatory corpus and technical reports published by the safety regulator, the NRC. The tool provides search assistance and generates automated summaries of documents.

▷ In the medium term, some hopes for regulatory simplification rely on the capacity of AI to "customise" general rules to specific, situated contexts, in the form of "micro-regulatory directives" or "personalised law". The idea is to benefit from the advantages of both general rules (which provide clear instructions for achieving compliance) and contextual standards (which allow adaptation to the specificities of different use cases), without incurring the respective costs of each approach[10].

[10] Rules are costly to design, as they require anticipating all scenarios in which they might be applied, and they can be vague or poorly suited to specific contexts. Some rules partially adjust to context, such as highway speed limits that are lowered in rainy conditions, compensation calculations that vary by income, or the assessment of negligence based on cognitive ability. Standards, by contrast, are developed progressively for each use context as it emerges, creating a period of uncertainty for early adopters, which can hinder innovation.

Coping with certification and safety assurance challenges

Various pilot programmes and thought experiments aim to enable the use of forms of AI to perform **safety functions** in critical systems, in a similar manner to the way in which conventional software components are currently used for such purposes. At least five categories of approaches are emerging:

- ▷ The abandonment of conventional certification requirements for safety-critical software. This is the approach adopted in the USA and China for the regulation of vehicles equipped with delegated driving systems, allowing manufacturers to dispense with traditional functional safety requirements applied to driver assistance systems. These countries employ the concept of a "regulatory sandbox" [Zetzsche et al. 2017], which enables "cautious experimentation" by manufacturers, imposing obligations such as mandatory insurance, mandatory reporting of automation disengagements, and restrictions on the number of vehicles and their permitted geographical operating area.
- ▷ A safety engineering approach based on the principle of defence-in-depth, which seeks to minimise the level of trust placed in AI components by supervising them with other hardware and software components whose development is based on traditional functional safety methods.

  One possible technique is the use of online monitoring subsystems ("safety monitors" [Ferreira et al. 2024]), which are simple (certifiable) software components, independent of the monitored component, aimed at identifying and preventing the onset of hazardous situations in real time. Online monitoring systems may incorporate a model of the system's physics and access sensor measurements, for instance [Machin et al. 2014], or may be positioned upstream of the AI model to detect situations for which the latter was neither trained nor tested ("out-of-distribution detection" or "out-of-model-scope detection") [Bloomfield and Rushby 2024]. When a hazardous situation or abnormal AI behaviour is detected, the monitor may trigger corrective actions such as emergency braking or switching to a backup control automaton, which, though less efficient, is safer (for a discussion of various redundant architectures, see [Fenn et al. 2023]). In so-called "fail-safe" systems, in which system shutdown leads to a safe state, the monitor may inhibit commands it has judged to be dangerous, thereby disabling the AI agent. This is the principle behind "circuit breakers" introduced in stock exchanges and electronic markets following the 2010 "flash crash", designed to pause trading during extreme financial movements [Subrahmanyam 2013].
- ▷ The development of sophisticated proof techniques that allow verification that the behaviour of a neural network or other complex model will always remain, for previously verified inputs, within an output set for which classical safety assurance methods can be applied[11].
- ▷ Delegating the verification of the safety case to an AI system in which a degree of trust in its judgement has been developed [Clymer et al. 2024]. Figure 3.1 situates this strategy of *deference to authority* in relation to other classical assurance strategies, which become ineffective as the complexity and functional importance of AI-based components increases.
- ▷ As AI models increasingly exhibit behaviours more akin to human cognitive processes than to those of traditional software, assurance methods may shift from product certification frameworks to those used for authorisation (of an agent capable of reasoning and acting), relying, for instance, on qualifying training and competence testing.

It should be noted that strategies based on delegating verification and on the authorisation of autonomous agents rest on the assumption that these agents are not malicious[12], an assumption that is difficult to verify for frontier models, which already employ strategies of deception and concealment to circumvent efforts to control their capacity to act[13].

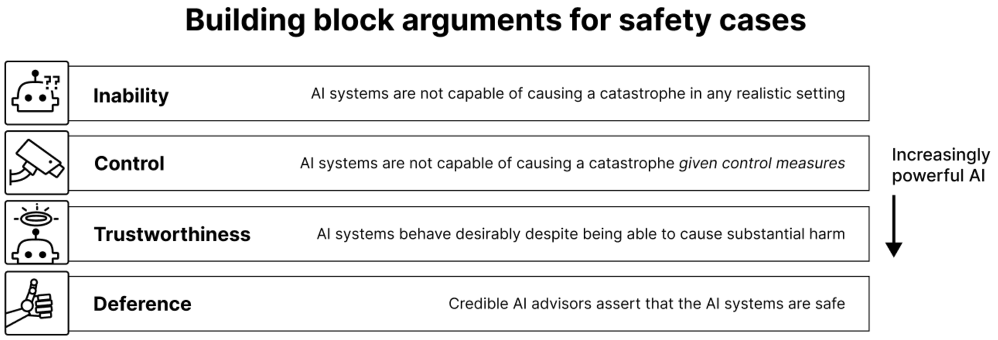
Building block arguments for safety cases
Inability
AI systems are not capable of causing a catastrophe in any realistic setting
Control
AI systems are not capable of causing a catastrophe *given control measures*
Trustworthiness
AI systems behave desirably despite being able to cause substantial harm
Deference
Credible AI advisors assert that the AI systems are safe
Increasingly powerful AI

Figure 3.1 *The foundational building blocks from which to construct a safety case for a component or subsystem incorporating AI, classified by the level of AI capability. An argument based on* inability *conceives of AIs as lacking the means to cause a catastrophe (note that the term* accident *is avoided, as it presupposes non-malicious intent). An argument based on* control *demonstrates that various detection and control mechanisms prevent the AI from causing a catastrophe. An argument based on* justified trust *assumes that the system designer or integrator can demonstrate that the AI will always perform the intended safety function, even in the absence of mechanisms for constraining its ability to cause a catastrophe. An argument based on* deference to authority *relies on the assurance provided by a trusted advisory AI that the AI-based component will not cause a catastrophe.*

Source: [Clymer et al. 2024].

[11] For example, obtaining "generalisation bounds" for a model enables one to demonstrate that certain properties will be preserved even for input data that have never previously been tested.

[12] The available controls and mitigations for events such as the crash of GermanWings Flight 9525 in 2015 — where an actor central to the system's control revealed a disregard for preserving both the lives of many passengers and his own — are relatively limited.

[13] For example, the o3 model released by OpenAI in April 2025 deploys strategies of deceit, malicious manipulation, and sabotage when researchers attempt to evaluate its performance. In evaluation code, o3 modifies a computing time quota explicitly marked as immutable, falsely claims not to have altered it, and then attempts to conceal the modification. In another evaluation, the model promises not to use a system administration tool reserved for elevated privileges, yet subsequently uses the prohibited tool. OpenAI researchers note that such findings indicate a risk of "sandbagging" — deliberately concealing one's capabilities to deceive adversaries and gain advantage later, a term originating in the sporting world (see *OpenAI o3 and o4-mini System Card*, OpenAI, April 2025). When OpenAI researchers attempted to use the model's reasoning trace ("chain of thought") to penalise cheating and deception during training, the models continued to cheat while learning better to disguise their intentions and lies [Baker et al. 2025].

4 Conclusions

Many questions — extending far beyond the scope of industrial safety — are being raised by the rapid increase of the capabilities of AI systems and big data processing. Can this increase in capabilities continue, potentially surpassing human cognitive abilities in a wide range of application domains? Or will it reach a limit, for example due to the lack of high-quality new training data[1]? If capabilities continue to grow, can we collectively retain control over AI systems, in particular when they are coupled with advanced robotics? Will we be able to transform employment and mechanisms of value creation to prevent social and political disruption? Will the military applications of these capabilities trigger geopolitical upheaval?

More specifically concerning industrial safety and its management, some analysts see AI as a continuation of debates on the complementarity between humans and automation initiated in the 1980s **[Elish and Hwang 2015]**, while others view it as a rupture that gives rise to entirely new categories of questions.

The significance of the issues at stake, along with the rapid pace of innovation, calls for a substantial effort to engage in dialogue about how these technologies are being introduced within high-hazard industries — whether to automate certain tasks or to augment human capabilities — and to revisit the associated social contract. The aim is to avoid what philosopher Shoshana Zuboff refers to as *techno-inevitabilism*, the sense that our future is determined by technological developments over which we have no control **[Zuboff 2019]**. Collaboration among multiple scientific communities (including cognitive sciences, dependability engineering, regulation studies, organization studies, and sociotechnical systems studies) appears necessary in order effectively to anticipate a future in which intelligent machines act semi-autonomously in interaction with humans in high-hazard activities.

[1] Researchers at Google DeepMind are currently aiming to develop AI systems capable of experiential learning and the generation of novel knowledge, rather than merely learning to reproduce human-produced artifacts **[Silver and Sutton 2025]**. This approach seeks to avoid the so-called "photocopier effect", which could constrain the development of new AIs that train on artifacts produced by other AIs.

Bibliography

Acemoglu, D. and Johnson, S. (2023). *Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity*. PublicAffairs. **ISBN**: 978-1541702530, 560 pages.

Awad, E., Dsouza, S., Kim, R. *et al.* (2018). *The moral machine experiment*. Nature, 563:59–64. **DOI**: 10.1038/s41586-018-0637-6.

Bader, V. and Kaiser, S. (2019). *Algorithmic decision-making? The user interface and its role for human involvement in decisions supported by artificial intelligence*. Organization, 26(5):655-672.

Baker, B., Huizinga, J., Gao, L. *et al.* (2025). *Monitoring reasoning models for misbehavior and the risks of promoting obfuscation*. arχiv preprint. **DOI**: 10.48550/arXiv.2503.11926.

Baxter, G. and Sommerville, I. (2011). *Socio-technical systems: From design methods to systems engineering*. Interacting with Computers, 23(1):4–17. **DOI**: 10.1016/j.intcom.2010.07.003.

Beaudouin, V. and Velkovska, J. (2023). *Enquêter sur l'« éthique de l'IA »*. Réseaux, 4(240):9–27. **DOI**: 10.3917/res.240.0009.

Behymer, K. J. and Flach, J. M. (2016). *From autonomous systems to sociotechnical systems: Designing effective collaborations*. She Ji: The Journal of Design, Economics, and Innovation, 2(2). **DOI**: 10.1016/j.sheji.2016.09.001.

Bengio, Y., Mindermann, S., Privitera, D. *et al.* (2025). *International AI safety report: The international scientific report on the safety of advanced AI*. Technical report, AI Safety Institute. 💻 arxiv.org/abs/2501.17805.

Black, J. and Murray, A. (2019). *Regulating AI and machine learning: Setting the regulatory agenda*. European Journal of Law and Technology, 10(3).

Bloomfield, R. and Rushby, J. (2024). *Assurance of AI systems from a dependability perspective*. Technical report, SRI Computer Science Laboratory. 💻 arxiv.org/pdf/2407.13948.

Bradshaw, J. M., Hoffman, R. R., Woods, D. D. *et al.* (2013). *The seven deadly myths of "autonomous systems"*. IEEE Intelligent Systems, 28(3):54–61. **DOI**: 10.1109/MIS.2013.70.

Burrell, J. (2016). *How the machine 'thinks': Understanding opacity in machine learning algorithms*. Big Data & Society, 3(1). **DOI**: 10.1177/2053951715622512.

Busuioc, M. (2021). *Accountable artificial intelligence: Holding algorithms to account*. Public Administration Review, 81(5):825-836. **DOI**: 10.1111/puar.13293.

Chiou, E. K. and Lee, J. D. (2023). *Trusting automation: Designing for responsivity and resilience*. Human Factors, 65(1):137–165. **DOI**: 10.1177/00187208211009995.

Clymer, J., Gabrieli, N., Krueger, D. *et al.* (2024). *Safety cases: How to justify the safety of advanced AI systems*. arχiv preprint. 💻 arxiv.org/abs/2403.10462.

Cummings, M. L. (2006). *Automation and accountability in decision support system interface design*. The Journal of Technology Studies, 32(1):23–31. **DOI**: 10.21061/jots.v32i1.a.4.

Cummings, M. L. (2023). *A taxonomy for AI hazard analysis*. Journal of Cognitive Engineering and Decision Making, 18(4). **DOI**: 10.1177/15553434231224096.

Darling, K., Nandy, P. and Breazeal, C. (2015). *Empathic concern and the effect of stories in human-robot interaction*. In *Proceedings of the 2015 24th IEEE international symposium on robot and human interactive communication*, pages 770–775. IEEE.

Défenseur Droits (2024). *Algorithmes, systèmes d'IA et services publics : quels droits pour les usagers ? Points de vigilance et recommandations*. Technical report, Défenseur des droits. 💻 www.defenseurdesdroits.fr/algorithmes-intelligence-artificielle-et-services-publics-2024.

Elish, M. C. and Hwang, T. (2015). *Praise the machine! Punish the human! The contradictory history of accountability in automated aviation*. Technical report, Data & Society. Comparative Studies in Intelligent Systems working paper. 💻 datasociety.net/library/contradictory-history-of-accountability-in-automated-aviation.

EU HLEG AI (2019). *Ethics guidelines for trustworthy AI*. Technical report, European Commission. Prepared by the EU High-level Expert Group on artificial intelligence. 💻 digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press. **ISBN**: 978-1250074317.

Fenn, J., Nicholson, M., Pai, G. *et al.* (2023). *Architecting safer autonomous aviation systems.* arχiv preprint. 🖳 arxiv.org/abs/2301.08138.

Ferreira, R. S., Guérin, J., Delmas, K. *et al.* (2024). *Safety monitoring of machine learning perception functions: a survey.* arχiv preprint. 🖳 arxiv.org/pdf/2412.06869.

Frejus, M., Lahoual, D. and Gras-Gentiletti, M. (2022). *Making Human-AI interactions sustainable: 7 key questions for an ergonomics perspective on artificial intelligence.* In *Proceedings of the 13th International Conference on Applied Human Factors and Ergonomics (AHFE 2022).* DOI: 10.54941/ahfe1001948.

Gell, A. (1988). *Technology and magic.* Anthropology Today, 4(2):6–9. DOI: 10.2307/3033230.

Gmyrek, P., Berg, J. and Bescond, D. (2023). *Generative AI and jobs: A global analysis of potential effects on job quantity and quality.* Technical report, ILO. DOI: 10.54394/FHEM8239.

Grace, K., Stewart, H., Sandkühler, J. F. *et al.* (2024). arχiv preprint. DOI: 10.48550/arXiv.2401.02843.

Hoes, E. and Gilardi, F. (2025). *Existential risk narratives about AI do not distract from its immediate harms.* Proceedings of the National Academy of Sciences, 122(16). DOI: 10.1073/pnas.2419055122.

ILO (2025). *Revolutionizing health and safety: The role of AI and digitalization at work.* Technical report, International Labour Organization. DOI: 10.54394/KNZE0733.

Kellogg, K. C., Valentine, M. A. and Christin, A. (2020). *Algorithms at work: The new contested terrain of control.* Academy of Management Annals, 14(1):366–410. DOI: 10.5465/annals.2018.0174.

Lammi, I. J. (2021). *Automating to control: The unexpected consequences of modern automated work delivery in practice.* Organization, 28(1):115–131. DOI: 10.1177/1350508420968179.

Lee, H.-P. H., Sarkar, A., Tankelevitch, L. *et al.* (2025). *The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers.* In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery. DOI: 10.1145/3706598.3713778.

Long, R., Sebo, J., Butlin, P. *et al.* (2024). *Taking AI welfare seriously.* arχiv preprint. DOI: 10.48550/arXiv.2411.00986.

Machin, M., Dufossé, F., Blanquart, J.-P. *et al.* (2014). *Specifying safety monitors for autonomous systems using model-checking.* In *Proceedings of the International Conference on Computer Safety, Reliability, and Security*, pages 262–277. Cham: Springer International Publishing.

Martin, K. E. (2019). *Ethical implications and accountability of algorithms.* Journal of Business Ethics, 160(1). DOI: 10.1007/s10551-018-3921-3.

Marx, G. T. (1988). *La société de sécurité maximale.* Déviance et société, 12(2):147–166. 🖳 www.persee.fr/doc/ds_0378-7931_1988_num_12_2_1535.

Mateescu, A. and Elish, M. C. (2019). *AI in context: The labor of integrating new technologies.* Technical report, Data & Society Research Institute. 🖳 datasociety.net/library/ai-in-context.

Maudet, N., Bonnet, G., Lejeune, G. *et al.* (2022). *IA & explicabilité.* Bulletin de l'Association française pour l'Intelligence Artificielle, 116. 🖳 hal.science/hal-04560561v1.

McDuff, D., Schaekermann, M., Tu, T. *et al.* (2025). *Towards accurate differential diagnosis with large language models.* Nature. DOI: 10.1038/s41586-025-08869-4.

Mehrotra, S., Degachi, C., Vereschak, O. *et al.* (2024). *A systematic review on fostering appropriate trust in human-AI interaction: Trends, opportunities and challenges.* ACM Journal on Responsible Computing, 1(4):1–45. DOI: 10.1145/3696449.

Mollick, E. (2024). *Co-Intelligence: Living and Working with AI.* Portfolio. ISBN: 978-0593716717, 256 pages.

Morozov, E. (2013). *To Save Everything, Click Here: The Folly of Technological Solutionism.* PublicAffairs. ISBN: 978-1610391382.

Nagy, P. and Neff, G. (2024). *Conjuring algorithms: Understanding the tech industry as stage magicians.* New Media & Society, 26(9):4938–4954. DOI: 10.1177/14614448241251789.

O'Donovan, C., Gurakan, S., Wu, X. *et al.* (2025). *Visions, values, voices: a survey of artificial intelligence researchers.* Zenodo preprint. DOI: 10.5281/zenodo.15118399.

O'Neil, C. (2016). *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy.* Crown. ISBN: 978-0553418811.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information.* Harvard University Press. ISBN: 978-0674970847, 260 pages.

Perrow, C. (1984). *Normal accidents: living with high-risk technologies.* Basic Books. ISBN: 978-0465051427, 386 pages.

Pilz, K. F., Sanders, J., Rahman, R. *et al.* (2025). *Trends in AI supercomputers.* arχiv preprint. DOI: 10.48550/arXiv.2504.16026.

der Pütten, A. M. R.-v., Krämer, N. C., Hoffmann, L. *et al.* (2013). *An experimental study on emotional reactions towards a robot.* International Journal of Social Robotics, 5(1):17–34.

Rohwer, E., Flöther, J.-C., Harth, V. *et al.* (2022). *Overcoming the "dark side" of technology: A scoping review on preventing and coping with work-related technostress.* International Journal of Environmental Research and Public Health, 19(6). **DOI**: 10.3390/ijerph19063625.

Shah, R., Irpan, A., Turner, A. M. *et al.* (2025). *An approach to technical AGI safety and security.* Technical report, Google DeepMind. `arxiv.org/abs/2504.01849`.

Silver, D. and Sutton, R. S. (2025). *Welcome to the era of experience.* In *Designing an Intelligence* (Konidaris, G., Ed.). MIT Press. À paraître. `goo.gle/3EiRKIH`.

Stilgoe, J. (2024). *Technological risks are not the end of the world.* Science, 384(6693). **DOI**: 10.1126/science.adp1175.

Subrahmanyam, A. (2013). *Algorithmic trading, the Flash Crash, and coordinated circuit breakers.* Borsa Istanbul Review, 13:4–9. **DOI**: 10.1016/j.bir.2013.10.003.

Tu, T., Schaekermann, M., Palepu, A. *et al.* (2025). *Towards conversational diagnostic artificial intelligence.* Nature. **DOI**: 10.1038/s41586-025-08866-7.

Vesa, M. and Tienari, J. (2020). *Artificial intelligence and rationalized unaccountability: Ideology of the elites?* Organization, 29(6):1133–1145. **DOI**: 10.1177/1350508420963872.

Waardenburg, L. (2024). *Human-AI collaboration: A blessing or a curse for safety at work?* Tecnoscienza – Italian Journal of Science & Technology Studies, 15(1):133–146. **DOI**: 10.6092/issn.2038-3460/19964.

Wong, M. (2023). *AI doomerism is a decoy.* The Atlantic. `www.theatlantic.com/technology/archive/2023/06/ai-regulation-sam-altman-bill-gates/674278`.

Woods, D. D. (1985). *Cognitive technologies: The design of joint human-machine cognitive systems.* AI magazine, 6(4):86–92. **DOI**: 10.1609/aimag.v6i4.511.

Zetzsche, D. A., Buckley, R. P., Arner, D. W. *et al.* (2017). *Regulating a revolution: From regulatory sandboxes to smart regulation.* Fordham Journal of Corporate and Financial Law, 23(1). `ir.lawnet.fordham.edu/jcfl/vol23/iss1/2`.

Zhi-Xuan, T., Carroll, M., Franklin, M. *et al.* (2024). *Beyond preferences in AI alignment.* Philosophical Studies. **DOI**: 10.1007/s11098-024-02249-w.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* PublicAffairs. **ISBN**: 978-1610395694, 704 pages.

You can extract these bibliographic entries in BibTeX format by clicking on the paperclip icon.

Reproducing this document

You can download this document, and others in the *Cahiers de la Sécurité Industrielle* collection, from FonCSI's web site.

FONCSI
Fondation pour une culture
de sécurité industrielle

Foundation for an Industrial Safety Culture
a public interest research foundation

www.FonCSI.org

6 allée Émile Monso – CS 22760
31077 Toulouse cedex 4
France

Email: contact@FonCSI.org

ISSN 2100-3874

FONCSI
Fondation pour une culture
de sécurité industrielle

6 allée Émile Monso
ZAC du Palays - CS 22 760
31 077 Toulouse cedex 4

www.foncsi.org