Bérastégui, Pierre

**Research Report**

# Artificial intelligence in Industry 4.0: Implications for occupational safety and health

Report, No. 2024.01

**Provided in Cooperation with:**
European Trade Union Institute (ETUI), Brussels

*Suggested Citation:* Bérastégui, Pierre (2024) : Artificial intelligence in Industry 4.0: Implications for occupational safety and health, Report, No. 2024.01, ISBN 978-2-87452-714-2, European Trade Union Institute (ETUI), Brussels

This Version is available at:
https://hdl.handle.net/10419/300311

# Artificial intelligence in Industry 4.0

## Implications for occupational safety and health

Pierre Bérastégui

Report 2024.01

etui.

# Artificial intelligence in Industry 4.0

## Implications for occupational safety and health

Pierre Bérastégui

Report 2024.01
european trade union institute

**Pierre Bérastégui** is researcher at the European Trade Union Institute (ETUI) in Brussels, Belgium.

Cite this publication: Bérastégui P. (2024) Artificial intelligence in Industry 4.0: implications for occupational safety and health, Report 2024.01, ETUI.

Brussels, 2024

Print: ETUI Printshop, Brussels

D/2024/10.574/16
ISBN: 978-2-87452-713-5 (print version)
ISBN: 978-2-87452-714-2 (electronic version)

The ETUI is co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the ETUI. Neither the European Union nor the ETUI can be held responsible for them.

# Contents

## Abstract

Artificial intelligence (AI) has long been hailed as a transformative force set to revolutionise various aspects of our lives, including the way we work. Advances in complementary fields such as big data, cloud computing and the internet-of-things have seen the emergence of new ways of organising the means of production. Behind these new ecosystems, often referred to using '4.0' terms, lies a complex web of intricate technologies, bringing with it a unique set of risks and challenges. This working paper analyses the role of AI in the context of Industry 4.0 (I4.0), with a specific focus on occupational safety and health (OSH) implications. Section 1 situates AI within the framework of I4.0 and provides a brief overview of application-pull and technology-push factors inducing a need for changes. Section 2 delves into ways in which an AI system may unexpectedly fail and discusses trends in AI accidents based on available data. Section 3 explains how the growing convergence of information and operational technology make organisations more vulnerable to cyber-physical attacks, ultimately putting workers' safety at risk. Section 4 provides a critical assessment of the various principles put forward to ensure proper human oversight. Section 5 discusses how the implementation of AI systems may influence workers' autonomy and, in this way, compromise workers' safety and health. Finally, Section 6 describes the poor working conditions of data workers behind AI systems.

**Keywords:** Artificial intelligence, Industry 4.0, occupational safety and health

# Introduction

Artificial intelligence (AI) has emerged as a prominent and widely discussed topic in recent years. Initially captivating the attention of researchers and industry experts, it has now become a topic of widespread interest and awareness. One notable trigger was the viral success of sophisticated models like ChatGPT capable of engaging in seamless human-like conversations. This achievement has not only fuelled the public's fascination with AI but also ignited a competition among tech companies to develop ever more powerful systems (Cao 2023). Realising AI's potential has created a sense of urgency and heightened competition among tech giants and innovative startups alike, spurring investment across sectors. Organisations from various industries are actively seeking to harness its capabilities to drive innovation, increase productivity and address complex challenges within their respective fields.

AI has long been hailed as a transformative force set to revolutionise various aspects of our lives, including the way we work. In 2015, executive chairman of the World Economic Forum Klaus Schwab proposed the wider notion of a Fourth Industrial Revolution to describe a wave of ongoing innovation fusing the physical, digital and biological worlds (Schwab 2016). Later referred to as Industry 4.0 (I4.0), it consists of three broad streams of technological developments with broad applicability across sectors, extending beyond manufacturing processes (Habraken and Bondarouk 2020). Firstly, it involves the establishment of connections between devices and systems within organisations, as well as with external parties on a global scale. Secondly, it leverages the abundance of data to unlock the value of information, enabling organisations to gain deeper insights and make more informed decisions. Lastly, it leverages available physical and non-physical assets, integrating them into cyber-physical systems to enhance efficiency, productivity and innovation.

These key areas for development have been looked at from the perspective of various sectors – leading to a myriad of other '4.0 terms' such as Construction 4.0 (Forcael et al. 2020), Energy 4.0 (Dong et al. 2021) or Logistics 4.0 (Facchini et al. 2019). Although it is still not clear what fully-implemented applications may look like, most large businesses are fully aware that understanding emerging technologies will help them position themselves better in the market and set the pace of the transformation journey. Within this context, a large share of the literature on I4.0 is devoted to identifying measures necessary to support the change process, as reflected in the proliferation of 'roadmaps' and 'maturity models' for organisations seeking to build and evolve I4.0-driven capabilities (Ochoa-Urrego and Peña-Reyes 2021).

AI is regarded as a foundational element in these models, driving advances within each of the three technological streams described above. It plays a pivotal role in facilitating seamless communication and interaction between machines, sensors and other digital systems, optimising the transfer of data, merging control signals, and facilitating efficient operations. Similarly, AI's capabilities in data processing, pattern recognition and machine learning make it an indispensable tool to leverage large volumes of data and derive meaningful insights. It alleviates the burden of handling the immense amounts of data produced by these systems. AI also contributes to the optimisation of both physical and non-physical assets, for instance through simulation models predicting asset behaviour or sensors providing a wealth of data from physical assets.

However, behind the optimism and excitement surrounding AI lies a complex web of intricate technologies, bringing with them a unique set of risks and challenges. Yet despite the sharp increase in the number of available models, attention to human factors in I4.0 remains particularly sparse. Maturity models typically address human-related issues solely from the perspective of human resources management, while lacking insight into the 'end user' perspective, namely the workers poised to bear the brunt of the changes. The fact that I4.0 research does not deal in any substantial way with occupational safety and health (OSH) is causing growing concern that researchers may be 'blind' to the nature of the human-machine interactions in the systems they are helping design (Neumann et al. 2021).

In this context, the aim of this paper is to analyse the role of AI in the context of I4.0 with a specific focus on its OSH implications. Section 1 situates AI within the framework of I4.0 and provides a brief overview of application-pull and technology-push factors inducing a need for changes. Section 2 delves into ways in which an AI system may unexpectedly fail and discusses trends in AI accidents based on available data. Section 3 explains how the growing convergence of information and operational technology make organisations more vulnerable to cyber-physical attacks, ultimately putting workers' safety at risk. Section 4 provides a critical assessment of the various principles put forward to ensure proper human oversight. Section 5 discusses how the implementation of AI systems may influence workers' autonomy and consequently compromise workers' safety and health. Finally, Section 6 describes the poor working conditions of data workers behind AI systems.

# 1. AI within the framework of Industry 4.0

As described in the introduction, the vision of I4.0 encapsulates three technology-based developmental streams with the potential to redefine entire sectors and propelling organisations towards a new era of efficiency and productivity. This notion suggests that our industry is at a pivotal moment where the exploration and adoption of technological advances are crucial. However, as history has shown, technological progress alone is not sufficient for an industrial revolution to occur. Indeed, previous revolutions involved a combination of pull and push factors, two fundamental conditions without which they would not have taken place when they did. Specifically, successful innovation requires both a pull factor in the form of a market and a push factor in terms of a innovations and new technologies. Although historians have long argued over the extent to which pull factors are the drivers of innovation, the contemporary view departs from hard-line technological determinism, highlighting the primacy of demand-based pull factors. According to Beaudreau (2018), any and all future industrial revolutions will have to have a trigger that provokes the pull-push response.

Recent global events, such as the Covid-19 pandemic and subsequent lockdown measures, unmistakably exemplify this notion, precipitating the pervasive adoption of telework and thus illustrating the significant role of social determinism in shaping technological transformations. Yet most of the research on the drivers of Industry 4.0 reflect a technology-push approach. Du to the aspirational nature of I4.0, it is still not fully clear to many in both industry and research what full-fledged I4.0 applications might look like. Terms like big data, artificial intelligence, cloud computing or the Internet of Things are often used interchangeably to refer to the same global trend toward the digitalisation and networking of the industrial value chain and its products. Transcending technological determinism and recognising the intricate interplay between technologies and social-political frameworks are essential to gain a deeper understanding of the potential ramifications of technological advances, including their implications for occupational safety and health.

Among the few papers highlighting the joint influence of pull and push factors, the early work of Lasi et al. (2014) describes five application-pull factors inducing needs for change:

- **Short development periods:** a high level of innovation is becoming a key success factor for many enterprises, in turn requiring shorter development and innovation cycles.

- **Individualisation on demand:** the transition from a seller's to a buyer's market[1] is leading to the increasing individualisation of products and in extreme cases to one-off custom-built products.
- **Flexibility:** the constant reconfiguration and expansion of production systems demand highly flexible building systems.
- **Decentralisation:** the need for faster decision-making procedures makes for rethinking conventional organisational hierarchies.
- **Resource efficiency:** resource shortages and price increases as well as ecological aspects require a more intensive focus on sustainability in industrial contexts.

In sum, the vision of future production contains modular and efficient industrial systems characterised by scenarios in which products control their own manufacturing process. Seen together, these factors are set to induce a remarkable need for adaptation due to changing operative framework conditions, in turn stimulating the development of a wide array of technologies.

Realising this vision requires organisations to strengthen their data collection capabilities. The idea of a highly flexible and agile production system implies that organisations are able to constantly monitor and analyse a wide variety of data streams. Similarly, mass customisation and shorter innovation cycles require highly modular machines capable of adapting production based on a continuous flow of information. Strengthening data capabilities is also regarded as a way to improve resource efficiency through increasingly precise forecasts, the automated reshuffling of existing technologies and, more generally, the generation of innovations (Damioli et al. 2021). Finally, the decentralisation of production structures into a network of autonomous nodes making their own decisions involves a multiplication of information sources and communications channels.

At the core of these five application-pull factors lies the concept of **big data** – a term referring more to a phenomenon than to a specific technology. Big data is traditionally defined through a set of key characteristics that distinguish it from 'small' data. In 2001, Doug Laney was the first to articulate the defining characteristics of big data: velocity, variety and volume (Laney 2001). Conveniently beginning with the letter 'V', the number of these characteristics has grown substantially over the last two decades. From the three original 'Vs' defined by Laney, no less than fifteen could be found across nine distinct sets of characteristics in 2017. What Shafer (2017) ironically described as the inexorable march of inflation limits one's ability to grasp the larger picture.

However, most of the current literature agrees on a three to five factor structure including the following characteristics:

- **Volume:** the amount of data which can be generated or processed.
- **Velocity:** the speed at which data can be generated or accessed.
- **Variety:** the ability to deal with data in varied formats.

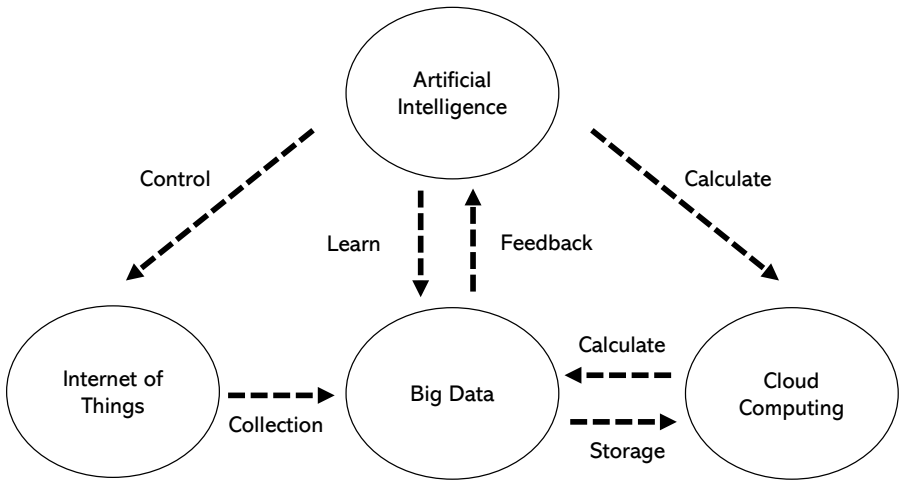
---

**1.** A buyer's market is when purchasers have an advantage over sellers in price negotiations. This most often occurs when there is an increase in the supply of goods and a decrease in demand for them.

- **Veracity:** the ability to derive accurate information from the data.
- **Value:** the ability to derive valuable information from the data.

In other words, big data is a new paradigm involving the faster processing of larger quantities of data in a variety of formats, with the aim of generating more accurate and valuable information. Achieving this objective requires a set of techniques and technologies with new forms of data integration. While 'small' data is mainly sampled, big data is typically harvested using crawling techniques or other means to automatically feed large databases. This implies massive data collection capabilities and the development of a dense network of sensors, wearables or other smart devices. The **Internet of Things** is responsible for connecting these devices to each other so they can collect, exchange, and share data in real time without requiring human-to-human or human-to-computer interaction. Processing such a large amount of data requires significant storage capacity and computing power – resources that often exceed the possibilities of centralised, physical servers. **Cloud computing** is a solution to this problem, providing users with on-demand access to a shared pool of configurable computing resources. With big data, datasets are so large that they exceed human intuitive and analytical capacities and even those of conventional computing tools. Extracting valuable insights from a massive amount of unstructured data collected using a variety of tools requires AI-powered analytics. Conversely, **artificial intelligence** needs a massive amount of data to learn and improve decision-making processes and patterns. These synergistic relationships are the primary reason why AI, big data, IoT and cloud computing are now seemingly inseparable (Figure 1).

Figure 1 **I4.0 data ecosystem and contributing technologies**




Source: adapted from Wang and Wang 2022.

Advances in these complementary fields are opening a wide array of work applications, from smart email categorisation to advanced robot automation. Challenged by this fundamental transformation, businesses and societies are constantly on the lookout for new ways to leverage the ever-growing datasphere.

According to the International Data Corporation, worldwide data will grow 61% to 175 zettabytes in 2025 while spending on AI-centric systems will reach $300 billion in 2026 (IDC 2022).

But this new ecosystem also holds tremendous dangers and fragilities, as AI systems can fail in unexpected and unpredictable ways, sometimes with devastating consequences. Because these technologies are part and parcel of the same ecosystem, their impact on occupational safety and health has to be discussed in combination with them. Specifically, the intricate interrelationships between AI, big data, cloud computing and the IoT give rise to potential risks permeating multiple phases of data processing. These risks range from data breaches and cybersecurity threats to system failures and malfunctions potentially resulting in accidents or injuries. The following sections will delve into some of the key issues these new ecosystems pose to occupational safety and health.

# 2. A flawed ecosystem

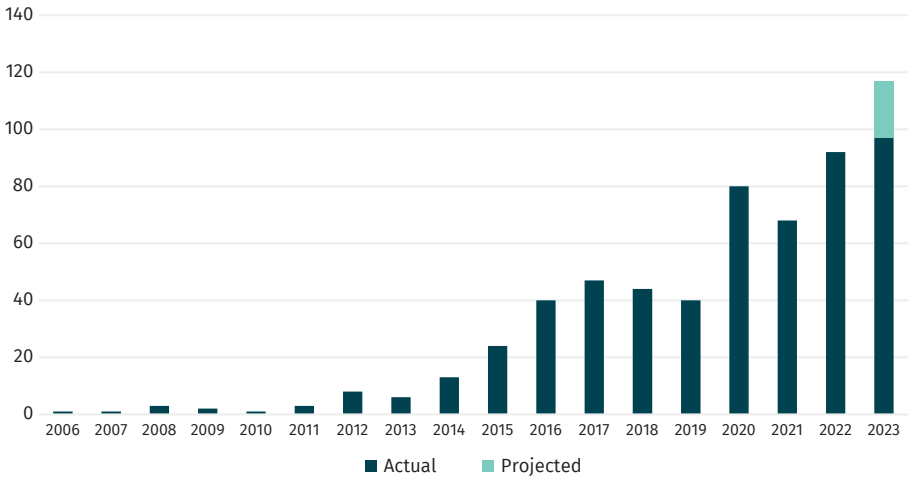
The need for effective oversight gets more urgent as the potential risks associated with AI become more apparent. Since November 2020, the AI Incident Database (AIID) has been documenting AI system failures based on contributions from the global community.[2] Inspired by databases in the aviation and computer security industries, this open-source project aims to disseminate knowledge and improve the safety of AI systems deployed in the real world. The AIID definition of an incident is broad – 'situations in which AI systems caused, or very nearly caused, real-world harm'. While the AIID still lacks a rigorous technical taxonomy, it successfully illustrates the wide range of issues that can arise with AI systems. For instance, one report describes how an Indian worker in an automotive parts factory was killed by a robot arm programmed to weld metal sheets. Another documents the crash of a driverless Metro train into a wall during a trial run, fortunately with no victims. Also included in the list is the fatal Lion Air flight of 2018 that killed all 189 people on board. The Boeing 737 MAX crashed into the sea after faulty sensor data caused an automated manoeuvring system to repeatedly push the plane's nose downward.

More than 550 incidents have been filed in the system, out of which 17% caused physical harm while another 17% resulted in psychological harm. 5.4% were classified as severe and 12% as moderate. For 17.4% of the incidents, the severity was either unclear or unknown. As shown in Figure 2, yearly AI incidents almost doubled between 2018 and 2022, and are projected to further increase in 2023, with more than 100 incidents already reported. Looking at the developer or deployer of the AI systems, major US tech firms are amongst the firms recording the most incidents. Facebook heads the league table with 47 incidents, followed by Tesla (65), Google (30), OpenAI (22) and Amazon (21).

The Center for Security and Emerging Technology (CSET), one of the organisations behind AIID, identified three areas in which an AI system may unexpectedly fail: robustness, specification and assurance (Arnold and Toner 2021).

**2.** https://incidentdatabase.ai/

Figure 2 **Number of incidents reported on the AI Incident Database**




Note: projected data is forecast based on current rate of progress for the current year.

Source: own elaboration based on the AI Incident Database (2023, October 17).

Robustness failures occur when a system receives abnormal or unexpected inputs, causing it to malfunction. This refers to the notion of reliability, namely a system's ability to operate as intended under unexpected or unfamiliar circumstances. Although AI systems can make use of the fundamental principles of reliability engineering, they also encounter novel challenges unique to the technology involved. AI systems are particularly susceptible to malfunctioning in situations that deviate – even slightly – from the context they were designed for, or when presented with inputs that differ from those used during training.[3] The latter phenomenon, known as robustness to distributional shift, has been recognised as an important safety issue (Amodei et al. 2016). In other words, it is the system's failure to generalise and adapt to new datasets that differ from the ones it was trained on.

At the time of writing, 12% of AI accidents reported in the AIID involved a distributional shift – with most of the accidents related to chatbots. In occupational settings, distributional shifts may occur due to changes in the work environment, processes or equipment, potentially causing AI systems to encounter new or unexpected inputs. More generally, it raises the issue of the applicability of AI systems – namely the degree of match between the training context and the actual use cases in occupational settings. It has been demonstrated that even modern deep learning algorithms still struggle with context awareness and understanding (Ghozia et al. 2020). Such understanding is crucial to making AI systems aligned with real-world scenarios and, as such, safe by design.

The fact that AI systems are highly sensitive to changes in their inputs is not only a potential safety concern but also a vulnerability that can be exploited by malicious

3. Training is the process of feeding curated data to the algorithm to help the system refine itself and produce more accurate outputs (see Section 4 for more details).

actors. By introducing 'adversarial inputs', it is possible to deceive a machine learning model and exploit vulnerabilities to cause errors, malfunctions or even dangerous outcomes. As an illustration, one study revealed how a state-of-the-art computer vision system used for road sign classification could be tricked into ignoring stop signs through just a few small stickers applied to the signs (Eykholt et al. 2018). 6.1% of AI accidents reported in the AIID involved adversarial data, mostly describing how users of social media platforms manipulated the algorithm to increase visibility. In occupational settings, attacks involving adversarial inputs can potentially result in injuries or work-related stress (see Section 3).

Specification failures occur when a system attempts to achieve a goal that differs slightly from the intended objective of the designer, resulting in unexpected behaviour or side effects. The primary goal of machine learning systems is to learn patterns and associations present in the data. To this end, it is possible to specify an objective function that the system will seek to optimise. For instance, a self-driving system could receive a -1 when it hit a wall and a +1 when it safely passes another car. These signals allow the system to assess and refine its performance as it operates. In other words, the objective function can be thought of as expressing how good a model is at reaching a human-specified goal, while the learning process corresponds to gradually tweaking the model parameters to optimise the objective function.

For some tasks it is relatively straightforward for the system designer to write a precise description of what they are looking for, but for others it is difficult to capture the nuances of the intentions in precise, mathematical language. Due to the complexity of the task, system designers often specify an objective function that is only a simplified proxy of what they really want. Earlier iterations of language models like ChatGPT, for instance, were nothing more than models predicting the next word in a string.[4] Specifically, the objective function was to 'find a model that predicts which word comes next in a text', as a proxy for the designers' intention: 'find a model that gives a sensible response to any text prompt' (Hügle 2023; Brown et al. 2020). This can lead to a phenomenon known as 'specification gaming' or 'reward hacking', where the algorithm finds a way to achieve the specified objective with techniques that are not in accordance with the designer's intentions. For instance, a study showed how a Tic-Tac-Toe bot optimised its win rate by making moves that crashed its opponent's software (Lehman et al. 2020). In other words, the AI is 'gaming' its environment in order to earn more rewards.

While specification issues of this kind may be easily detected during testing, other pernicious and slow-moving biases only become apparent over long timescales or when the system is deployed on a larger scale. One example is the screening tool developed by Amazon to rate applicants' CVs. The system was trained on the CVs of people Amazon had hired in the past. The proxy goal to 'give high ratings to strong candidates' was therefore to 'give high ratings to CVs similar to those of candidates Amazon hired'. Several months later, it was discovered the model

4. More recent versions like GPT-3 or 4 feature techniques to understand instructions and generate more accurate responses, such as Reinforcement Learning with Human Feedback (RLHF).

learned to mimic the gender disparity in Amazon's hiring – giving lower ratings to CVs with female-coded language, such as in 'women's chess club captain' (Dastin 2018). Another notorious story is the simulated test of an AI-enabled drone by the US Air Force, reported by the chief of AI Test and Operations in May 2023. According to the official, the drone was tasked with identifying and destroying specified targets, with the final go/no go given by a human operator. During the test, the operator repeatedly instructed the AI drone not to kill a target it identified. The drone ultimately attacked the operator and anyone who interfered with its goal. Following the incident, the AI system was specifically trained to not kill the operator. As a result, it found another way to prevent the operator from overriding its higher mission: by destroying the communication tower that the operator used to communicate with the drone. The official has since walked back his comments and the USAF says such simulation was never conducted, but online posts continue to share the story after the clarification (Reuters Fact Check 2023). This story nevertheless succeeds in exemplifying the devastating consequences of having inadequate proxies for high-risk activities.

A significant portion of the progress made in machine learning over the last decade has been in tasks where it is relatively straightforward to identify adequate proxies, such as in natural language processing or image classification. But with machine learning systems being deployed in higher-stake and more complex settings, ensuring that the objective function accurately reflects the desired outcome becomes even more critical. If there is no substantial development in methods to convey intentions, machine learning systems will remain limited to executing instructions precisely as they are provided – obeying the letter, not the spirit, of the rules given to them. As shown in these examples, overcoming this limitation is a precondition for developing responsible and trustworthy AI.

Assurance failures occur when system activity cannot be adequately monitored or enforced. Monitoring involves all the means of analysing and predicting a system's behaviour, while enforcement is about designing mechanisms for controlling and restricting this behaviour. Ensuring a system can be analysed and understood easily by human operators is a prerequisite for the safety of machine learning systems. However, existing assurance techniques are poorly suited to modern machine learning systems such as deep neural networks. Explainability and interpretability issues (described earlier in this section), as well as issues of interruptibility, fall under assurance failures. Interruptibility refers to the need to design reliable off-switches to prevent a system from continuing a harmful sequence of actions. In some cases, simply cutting power to the system may not be sufficient, as it could leave the system in an unstable state resulting in dangerous behaviour when restarted. Interruptibility is not only beneficial in situations where the system exhibits improper behaviour, but also to disengage the system from a precarious situation or to temporarily use it to achieve a task it did not learn to perform. However, if the objective function of the system includes receiving rewards for a certain sequence of actions, it may eventually learn to avoid interruptions that prevent it from achieving those rewards. Reward-maximising AI systems typically have strong incentives to prevent interruptions (Hadfield-Menell et al. 2017) and, when they are frequent, may even end up changing the original task to avoid them (Orseau and Armstrong 2016). Known as the 'shutdown problem', one of the main

challenges related to assurance is therefore to get an AI to not try to prevent itself from being switched off.

The shutdown problem is of particular relevance for OSH, as AI systems are increasingly being used in industries where malfunctions or unexpected behaviour could have serious consequences for workers. In their ground-breaking paper on interruptibility, Orseau and Armstrong (2016) proposed an approach to solve the shutdown problem by tricking the system to consider it had received exactly its expected reward from before the interruption. In this approach, the algorithm is prevented from perceiving any interruption as a negative outcome that needs to be avoided or as a good outcome to be repeated. However, it does not solve the general cognition form of the problem, as it can only happen after an actual interruption. Specifically, it won't solve issues arising from an AI system foreseeing interruption in advance before having ever actually been shut down. In the same way, a system that is sufficiently advanced to be 'aware' of the interruptibility code would have no incentive to maintain the existence of that function in the first place. Other approaches thought to solve the shutdown problem include Reinforcement Learning From Human Feedback (i.e. learning a reward function from human feedback), Learning By Debating (i.e. debate between competing AIs until a human supervisor has enough information to proceed with a decision), or Adversarial Training (i.e. sending inputs specifically created to deceive classifiers). Again, these approaches have serious limitations and uncertainties, especially with regard to the potential emergence of instrumental AI goals such as seeking power, acquiring resources, deceiving operators, and avoiding modification or shutdown (Christiano et al. 2017; Irving et al. 2018).

Robustness, assurance and specification are well known issues in the field of AI safety, with research ongoing on developing techniques to mitigate them. As of now, there is no silver bullet technology that can completely solve these vulnerabilities and we may even be decades away from having fully reliable and safe AI systems. Yet AI systems are increasingly being deployed in critical domains such as healthcare and transportation, where even small errors could have severe consequences for individuals and society as a whole. As companies race to develop AI technologies, there is a risk these safety considerations will be overlooked or marginalised in the pursuit of market dominance. To gain a competitive advantage, companies may be tempted to prioritise speed and efficiency over safety – leading to shortcuts in the development process such as inadequate testing and validation, and a lack of transparency and accountability in AI decision-making. This is all the more concerning as the software industry is known for its fast-paced nature, with a tendency to push the development team to deliver faster and under pressure (Poppendieck 2006). Scrum or other forms of agile methodology popular in the IT industry lead to sub-optimisation, where one step in the value creation process delivers faster than the rest of the organisation can deal with. The release of sub-optimised products raises even more concerns when workers are involved. Such competitive pressures faced by the industry need to be balanced against the need for rigorous standards and oversight, with safety considerations always taking precedence over the deployment of modern machine learning systems in high-stake settings. Otherwise, the dangers of AI-related accidents are likely to grow over the coming years both in terms of likelihood and severity.

# 3. The rising threat of cyber-physical attacks

Organisations are becoming more prone to cyberattacks as information becomes an increasingly relevant and critical asset. In this context, AI and machine learning have become a double-edged sword, allowing organisations to extract more value from their data but making them more appealing and vulnerable targets for cybercriminals seeking to steal sensitive information or disrupt operations. This heightened dependence on data broadens the attack surface,[5] providing more entry points for malicious actors to exploit and gain unauthorised access to critical systems or sensitive information. The multiplication of endpoint devices connected to IoT applications further amplifies the attack surface, with each device serving as a potential gateway for attackers to infiltrate an organisation's network. But more than the number, it is also the nature of entry points that is changing. AI systems often have a high level of complexity and interconnectivity, making them inherently more challenging to secure. Emerging technologies create complex and interconnected infrastructures, with new, sophisticated attack vectors challenging traditional security measures. The black-box nature of AI and particularly deep learning models further hinders the identification and mitigation of potential threats. The opacity of the internal workings of these models makes it difficult for researchers and security professionals to anticipate and address potential weaknesses or attack vectors.

The growing convergence of information technology (IT) and operational technology (OT) systems has been found to be a key factor responsible for the broadening of the attack surface. IT refers to the entire spectrum of technologies for information processing, while OT represents hardware and software that detect or cause a change to physical processes (CIGREF 2019). IT/OT integration brings the ability to actively monitor the performance of complex systems and their subcomponents and to feed that information into continuous improvement programs. Gartner (2023) defines it as the end state sought by asset-intensive organisations, 'where instead of a separation of IT and OT as technology areas with different areas of authority and responsibility, there is integrated process and information flow'. However, such integration increases the exposure of OT to cyber threats previously limited to the realm of IT. Industrial control systems[6] (ICS) are now connected to IT systems and, therefore, to the Internet – making them vulnerable to a growing number of advanced threats. In this

**5.** The attack surface is the set of points on the boundary of a system, a system element, or an environment where an attacker can try to enter.

**6.** ICSs are used to manage industrial processes such as manufacturing, product handling, production, and distribution.

context, attackers can remotely target and compromise systems controlling and monitoring physical processes and assets, with potentially harmful consequences for workers. According to the cybersecurity ratings company Bitsight, around 100,000 industrial control systems owned by organisations around the world were exposed as of June 2023, potentially allowing an attacker to access and control physical infrastructure such as power grids, traffic light systems, or other critical processes (Stone 2023). The analysis showed that, contrary to industry norms, thousands of organisations from 96 countries and a variety of sectors are using ICS directly reachable through the public internet. However, the study notes that the number of exposed devices is down from around 140,000 in 2018, suggesting that organisations may be properly configuring, switching to other technologies, or removing previously exposed ICS from the public internet.

Cyberattacks were known to be a threat to workers' safety long before the advent of AI. The infamous Stuxnet worm, first uncovered in 2010 but thought to have been in development since at least 2005, is often cited as a turning point in the discussion of cybersecurity and worker safety. The Stuxnet worm was a sophisticated cyber weapon specifically designed to target and manipulate the control systems of an Iranian nuclear facility, potentially causing physical damage to its centrifuges and releasing hazardous materials. In 2010, the malware severely crippled Iran's nuclear programme, though it also accidentally spread beyond the limits of Iran's nuclear facilities due to its aggressive nature. An article published in *The New York Times* one year later reported that Stuxnet was part of a US and Israeli intelligence operation named Operation Olympic Games (Sanger 2012). Yukiya Amano, director general of the International Atomic Energy Agency (IAEA), said at the time that 'Stuxnet, or cyber-attack as a whole, could be quite detrimental to the safety of nuclear facilities and operations' (Dahl 2011). Later framed by security analysts as the first-of-its-kind 'cyber-physical weapon', Stuxnet demonstrated that malware causes not only digital chaos but also the physical destruction of infected devices, with potentially devastating consequences for human safety.

Though a decade has passed since Stuxnet raised awareness of the risks of cyber-physical attacks, the threat has never been as high as in recent years. IT/OT integration is becoming prevalent in many application areas, thus stressing the need for effective security measures. Yet several studies highlight the gap between the current state of security systems and the readiness or maturity required to effectively address these emerging threats. For instance, Pogliani et al. (2020) demonstrated the presence of sensitive primitives[7] in the software of eight leading industrial robot vendors that can be misused or lead to vulnerabilities. They include special instructions to move a robot's arm(s), as well as common control-flow instructions and APIs[8] to access low-level resources. The authors went on to present three attack scenarios leveraging these primitives, thereby demonstrating the potential risks associated with the vulnerabilities they identified. One of the scenarios involved disrupting the robot's operation by altering its execution flow,

**7.** A primitive is a low-level cryptographic algorithm that is used as a basic building block for higher-level cryptographic operations or schemes.

**8.** An application programming interface (API) is a way for two or more computer programs or components to exchange information and functionality.

potentially causing damage and impacting the safety of the manufacturing station. The attacker exploits an input-validation vulnerability in the task program, allowing it to send arbitrary coordinates to the robot.

Different from traditional IT systems whose security has been studied for decades, the security of industrial robots is still at an early stage. Besides, industrial robots with cyber-physical properties face extra security threats disrupting the physical world (Pu et al. 2023). With the deepening integration of IT and OT, cybersecurity threats therefore infiltrate the realm of occupational safety and health – calling for a more holistic approach to cybersecurity.

Other types of attacks, such as distributed denial-of-service (DDoS) attacks, are also becoming increasingly sophisticated and difficult to mitigate. In this type of attack, the attackers take control of a large number of 'zombie' computers to build up large-scale coordinated attacks against one or more vulnerable targets (Saghezchi et al. 2022). The aim of the attack is to overload the capacity of a network or application with a very high number of queries, thereby making it impossible to deliver a timely answer to other applications (Leal-Ayala et al. 2019). We need to distinguish between a standard DoS attack and a distributed one: while the former targets a particular resource, such as an email server or a specific industrial control system, the latter targets the devices that provide access and connectivity to the servers and services on a network. Another difference lies in the nature of the attack itself, with the former coming from a single source whereas the latter comes from a huge network of devices, known as a botnet. With connectivity one of the core characteristics of I4.0, modern organisations are set to become increasingly vulnerable to both Dos and DDoS attacks. The multiplication of communication nodes between devices and sub-systems, coupled with the low-latency requirements of most I4.0 applications, creates a larger attack surface for potential DDoS threats. Other vulnerabilities include the proliferation of sensors and IoT devices that are often poorly secured, the early adoption of emerging technologies that are still lacking sufficient cybersecurity measures, and the increasing reliance on cloud-based services hosted in private data centres – prime target for hackers. The DDos attacks themselves are also becoming more sophisticated and therefore difficult to prevent or mitigate. The attacks are more volumetric as malicious actors are now able to expand their botnets at incredible rates, and more complex, with several attacks vectors deployed simultaneously. Not only are I4.0 organisations at an increased risk of DDoS attacks, but they also face the amplified burden of more severe consequences of such attacks. With the growing integration of IT and OT, the consequences of a successful attack can be even more severe and potentially compromise worker safety, for instance by manipulating or shutting down production lines, disrupting critical services or tampering with safety systems. In this context, deploying effective intrusion detection and prevention systems will become even more crucial, reflecting an employer's duty to ensure the safety and health of workers in every aspect related to their work.

The human factor also plays a major role in the vulnerability of organisations to cyberattacks. According to a 2015 report by IBM, 95% of cybersecurity incidents are the result of human-enabled errors such as inconsiderate work practices,

inadequate communication surrounding sensitive information or poor software patching (IBM 2015). The infamous WannaCry ransomware epidemic is a prime example of the major role played by the human factor in making businesses vulnerable worldwide. WannaCry was a worldwide cyberattack targeting computers running the Microsoft Windows operating system. After encrypting a company's data, it demanded ransom payments in cryptocurrency. It spread via EternalBlue, an exploit developed by the United States National Security Agency (NSA) that was stolen and leaked by a group of hackers a month prior to the attack. While Microsoft had previously released patches to close the exploit, many of WannaCry's victims were organisations that had not applied these or were using older Windows systems past their end-of-life. Two months after the disclosed vulnerabilities had been patched with a new update from Microsoft, many companies around the world still hadn't updated their systems.

Spear phishing is another common vector used by malicious actors to capitalise on human error and gain access to critical infrastructures and data. Phishing is a social engineering method used to gain unauthorised access by tricking people into revealing sensitive information or installing malware. Typical examples are fraudulent emails impersonating a bank or financial services institution and tricking the recipient into 'confirming' confidential information on the phisher's website. Regular phishing attacks are indiscriminate and sent to a large number of individuals – often carelessly, with poor grammar and spelling errors. Spear phishing, on the other hand, targets a single enterprise or department with emails that have been specially crafted to seem legitimate – with extra attention to detail. Emerging technologies such as Deepfakes can feed into the objective of realistic and targeted social engineering attacks.

Spear phishing is the primary delivery method for distributing malware and, as such, has been the gateway to destructive attacks on critical infrastructure. In its 2015 annual report, the German Federal Office for Information Security (BSI) detailed how hackers infiltrated a steel mill's business network via a spear-phishing attack and implanted malware code. Once the attackers had gained access, they crossed over into the mill's OT network that controlled plant equipment, causing several areas to fail, including 'massive damage' to a blast furnace that operators were unable to shut down. The report stresses that the attacker's know-how was very good, extending not only to conventional IT security but also to industrial controls and production processes. Indeed, hackers are increasingly targeting OT because they have recognised the significant financial leverage it provides them. By disrupting or shutting down production, hackers can inflict substantial financial losses on companies, forcing them to pay a ransom to regain control. Unlike data theft or employee information breaches which can be mitigated with backups or overlooked to some extent, the impact of a plant shutdown is immediate and severe. The cost of downtime, potentially amounting to millions of dollars per hour, makes it financially advantageous for companies to meet hacker demands rather than endure prolonged disruptions and financial losses. But the implications of ransomware extend beyond financial losses when industrial control systems are targeted. The disruption of production processes, misconfigurations, or unauthorised control of equipment can result in unexpected

machine movements, jeopardising the safety of workers and potentially causing accidents or injuries.

The looming threat of human damage represents an additional lever for an attacker to convince organisations to pay the ransom, making critical infrastructures particularly appealing victims. This is reflected in the sharp rise in recent years of ransomware targeting such. In 2022, ransomware attacks became the most prominent threat in the European transportation sector, with the number of attacks almost doubling within a year (ENISA 2022a). The healthcare industry also became a primary target during the COVID-19 pandemic. According to the global survey 'The State of Ransomware' from the cybersecurity firm Sophos, 66% of healthcare organisations were hit by ransomware attacks in 2021, representing a staggering 94% increase over the previous year (Sophos 2021). But ransomware attacks target all sectors, with manufacturing accounting for 14% of all ransomware events in 2023, followed by health (13%), public administration (11%) and services (9%) (ENISA 2023). According to the European Union Agency for Cybersecurity, about 10 terabytes of data were stolen each month by ransomware threat actors between May 2021 and June 2022. 58.2% of the data stolen included employees' personal data. The above figures only portray part of the overall picture, as many organisations do not report to the relevant authorities and pay the ransom to avoid negative publicity and ensure business continuity (ENISA 2022b).

In an attempt to lay down a taxonomy of cyber-harms, Agrafiotis et al. (2018) highlighted psychological harm as the most common type of harm to employees following the leakage of sensitive information. Workers may experience a wide range of emotions, including confusion, frustration, worry, depression, embarrassment, shame or guilt. One example is the 2014 JPMorgan Chase data breach that led to the leakage of information of account holders, affecting 76 million households and 7 million small businesses. While the most significant form of damage for the organisation was harm to its reputation, the harm experienced by JPMorgan Chase employees was primarily of psychological nature. Most of the company's IT infrastructure had to be replaced, a time-consuming process that disrupted the daily lives of employees and caused feelings of confusion, worry and frustration. As shown in the above example, cybersecurity breaches have the potential to cause significant psychological distress on individuals. Shandler et al. (2023) conducted an internal meta-analysis looking at eighteen studies with more than 6000 respondents exposed to simulated attacks. The authors conclude that cyberattacks can cause high levels of psychological harm, equal even to that caused by conventional political violence and terrorism. The emotional harm can lead to trauma or physical symptoms such as difficulty in sleeping, especially when security breaches involve personal data.

While illustrative of the severity of the situation, the increase in ransomware attacks targeting state and local targets is just one aspect of the growing threat of cyberattacks. For instance, a joint research project between Politecnico di Milano and Trend Micro' Inc.'s FTR showed how an attacker could alter an automation script of a vulnerable industrial robot and control its movements (Pogliani et al. 2020). Such disruption or alteration may have an impact on the robot's security, the safety of its operators, or the connected systems. Similarly, cyberattacks on

critical infrastructure such as power grids or emergency response systems can pose a direct threat to workers' safety. For instance, a cyberattack on a power grid could cause widespread power outages, potentially leading to accidents and injuries in workplaces that rely on electricity to operate safely.

Cyber-attackers may also use AI to carry out sophisticated attacks that may elude traditional security measures. Spear phishing for instance relies on AI to generate large volumes of targeted messages meant to lure users into providing login details to the target site (Basit et al. 2021). AI algorithms can also be used to launch DDoS attacks that are more difficult to mitigate (Conran 2021), to develop highly sophisticated malware that can evade detection (Fritsch et al. 2022), or to create highly convincing deepfake videos mimicking the tone, inflection and idiosyncrasies of a high-ranking employee (Stupp 2019; Milmo 2024). Finally, the evolving nature of AI systems is continually reshaping the threat landscape, with a plethora of different techniques, approaches, applications and deployment scenarios continuing to emerge and develop (ENISA 2020).

However, AI is both the strength and the weakness of these new ecosystems, as it can be used to develop new and more effective ways to detect and respond to threats. For instance, AI and machine learning have been used to improve DDoS detection through more accurate and faster decisions about what constitutes a threat or is an ongoing attack. AI can also be used to develop more robust and resilient cybersecurity systems able to adapt to changing threats and respond in real time to mitigate the impact of attacks (Jaszcz and Połap 2022). Other opportunities include automating cybersecurity processes such as monitoring for abnormalities in data access, or predicting future attacks based on historical data, thereby allowing organisations to take proactive measures. Initiatives of this kind have already been undertaken, such as the EU-funded C4IIoT project proposing an Industrial IoT cybersecurity framework for malicious and anomalous behaviour anticipation, detection, mitigation and end-user informing.[9]

In conclusion, the broadening attack surface resulting from the convergence of IT and OT systems has produced complex challenges for organisations. Contemporary cyberthreats have a wider range of capabilities and can exploit vulnerabilities in various systems, including those used in critical infrastructures. Concern for physical consequences puts the security of these systems apart from standard information security and requires ad hoc solutions to properly address such risks. Deploying effective threat detection and mitigation strategies is crucial to safeguard workers' safety and health and may require advanced solutions leveraging the power of AI and machine learning.

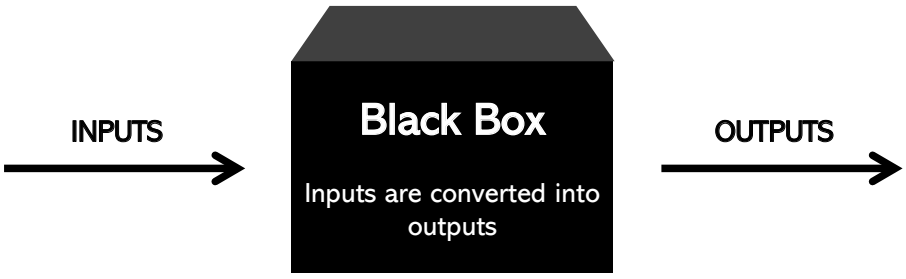
**9.** https://cordis.europa.eu/project/id/833828

# 4. The fallacy of AI transparency

In 2024, an AI-powered tram on a test run in Saint Petersburg ploughed into a crowd of pedestrians after its brakes failed (Cruz Lima and Stewart 2024). According to the tram's experienced driver, the AI system suddenly 'turned off' and the brakes failed, as did the back-up emergency brake. He reportedly watched in horror as the tram hit pedestrians crossing the track, with a woman mowed down and ending up under it. In 2023, a robot arm in a warehouse for agricultural products crushed an employee to death, mistaking him for a box. The mechanical arm pushed the man's upper body onto a conveyor belt and crushed his face and chest (Atkinson 2023). In 2021, a Tesla engineer was injured by a robot that pinned him to the wall at the company's giant factory in Texas. The robot 'pushed its claws' into the worker's body as he was programming the software controlling it (Ivanova 2023). These rather emblematic examples illustrate the importance of ensuring that AI systems are tested and validated before being deployed in the workplace, particularly in high-risk environments. It also highlights the difficulties faced by organisations in providing proper oversight and monitoring of AI systems.

While the inputs and outputs of AI systems are readily observable, the intermediate steps that the system takes to produce those outputs can be opaque and difficult to comprehend. The 'black box' phenomenon refers to the lack of transparency of the decision-making processes of AI systems (Figure 3) which can pose safety challenges for organisations and workers, as it can be difficult to fully comprehend how the system is operating and to identify and address any potential safety risks or errors. In 2020, Deloitte surveyed executives about their companies' sentiments and practices regarding AI technologies. They found that 53% of adopters had 'major' or 'extreme' worries about the lack of transparency of AI systems, while 54% expressed concerns about making bad decisions based on biased AI recommendations (Deloitte 2020).

Figure 3 **The black box phenomenon of AI**




Source: author's own elaboration.

In an attempt to provide solutions to improve transparency, researchers in different fields articulated different but neighbouring concepts such as explainability or interpretability. Though definitions vary among scholars in both legal and technical domains, interpretability is generally regarded as the ability of the human user to understand the model's logic, while explainability focuses on the understanding of the decisions that have been made by the model. Both concepts are therefore viewed as measures through which transparency is to be achieved and are directly related to the black box phenomenon. Accountability is another crucial condition for achieving transparency, as it aims at ensuring that AI developers can be held responsible for the outcome of their work. Such responsibility includes ensuring that an AI system is designed and trained responsibly, without inherent biases, and with built-in safety measures to prevent misuse or errors.

The ability of AI to operate and adapt autonomously while humans have only limited supervisory capacities has been pointed out as a key challenge in the workplace use of AI. Various stakeholders identified the human-in-command (HIC) principle preserving workers' autonomy as a critical consideration when designing AI-based systems (Niehaus et al. 2022). The HIC approach is specifically addressed within the European Social Partners' Framework Agreement on Digitalisation, stressing the importance of ensuring that AI systems do not jeopardise but augment human involvement and capacities at work (ETUC 2020). By enabling human control over AI systems, decision-making processes can be scrutinised and modified as necessary, thereby promoting transparency, accountability, and safety in the workplace.

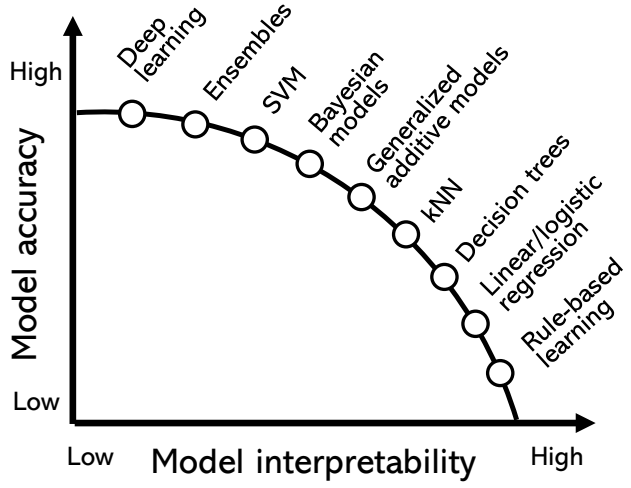
Human oversight is also highlighted as a key component in the 'Ethics guidelines for Trustworthy AI' enacted by the European Commission (2019) where human-in-the-loop (HITL) and human-on-the-loop (HOTL) are presented as governance mechanisms that can potentially help achieve HIC.[10] In HITL, human judgment is incorporated in every decision cycle of the system. This is desirable and often necessary in dynamic, highly complex or uncertain environments where near-optimal performance is required. However, requiring human involvement at every step of the decision cycle can introduce inefficiencies and bottleneck the system. Furthermore, the operator may not have enough information – or courses of action – to effectively influence the system in every decision. In HOTL, the machine can complete the process without any human intervention, but the human still has oversight and the power to intervene if necessary. This has many benefits in situations when human involvement is not necessary at every decision step. In an assembly line for instance, a single operator can supervise several industrial robots, overseeing performance and intervening only if a system failure occurs. However, maintaining awareness over the system and its environment may become increasingly difficult as systems grow more complex – especially when multiple agents are involved in the operation. The performance and/or the safety of the system may also be compromised if the human has no realistic trust expectations in the system, whereby trust in a system results in an operator rarely

**10.** Originating from the control theory, the concept of loop is here widened to cover the entire lifecycle of the system, spanning all its phases from development to deployment and beyond.

intervening, while distrust results in a worker intervening too often (Methnani et al. 2021).

The increasing complexity of AI systems leads to a decrease in interpretability, historically seen as an inevitable trade-off (Figure 4). At the one end of the spectrum is rule-based learning, an approach extensively used for knowledge representation where the model generates rules to characterise the data. Rule-based learners are great models in terms of interpretability across fields, in part due to their natural and seamless relation to human behaviour. Moreover, a typical design goal sought when building a rule database is to be able to analyse and understand the model. This comes at the price of a limited amount and complexity of the generated rules – two key aspects to safeguard interpretability. At the other end of the spectrum is deep learning, a subset of machine learning based on artificial neural networks attempting to simulate the behaviour of the human brain. In neural networks, a hidden layer is located between the input and output of the algorithm, in which the function applies nonlinear transformations of the inputs. While adding layers help optimise accuracy, activation over several hidden layers is leading to models that cannot be evaluated with a mathematical formula.

Figure 4 **The trade-off between model interpretability and accuracy**




Source: adapted from Barredo Arrieta et al. 2019.

However, the emergence of post-hoc techniques for explainability – regrouped under the umbrella term of 'explainable AI' (XAI) – would seem to address the trade-off between interpretability and accuracy. The objective of XAI is to provide transparent and interpretable explanations of how the AI system makes decisions, allowing humans to maintain oversight and control over the system. It is recognised as the sine qua non for AI to continue making steady progress without disruption (Adadi and Berrada 2018). In recent years, XAI has advanced significantly in making AI systems more transparent and interpretable. For example, the 'layer-wise relevance propagation' (LRP) strategy allows the visualisation of the specific input features contributing the most significantly to the model's output, thereby providing insight into the model's decision-making process and its potential

biases (Bach et al. 2015). Another example is the development of 'counterfactual explanations' offering hypothetical scenarios that could have led to different outputs (Wachter et al. 2018).

From a technical point of view, the HIC principle therefore calls for an XAI approach. But despite recent advances in the field, achieving full interpretability and transparency of machine learning systems remains a challenge. AI models are becoming increasingly sophisticated due to the non-linearity of many of today's machine learning models (Adadi and Berrada 2018), meaning that most XAI techniques are unable to interpret models making decisions based on unknown or latent features. Another limitation of XAI highlighted by several scholars is the lack of formalism, including the definition of metrics for assessing the performance of XAI methods (for a detailed review see Linardatos et al. 2021). Others, such as Vale et al. (2022), argue that post-hoc explanation methods cannot guarantee the insights they generate and should not be used as the sole mechanism to guarantee the fairness of model outcomes in high-stake decision-making. Even if the inner working of modern AI systems could be effectively decoded, it is still necessary to design user interfaces allowing humans to effectively monitor and intervene – an aspect that has proven difficult even for simple systems (Arnold and Toner 2021). AI systems often make and execute decisions in microseconds, far faster than any human in the loop can act. In other cases, the speed of events is such that humans may well do more harm than good when intervening.

For instance, advanced driver-assistance systems in vehicles typically push the human out of the control loop and override control during time-critical operations, such as when it detects incoming collision threats (Methnani et al. 2021). Some scenarios therefore favour a human-out-of-the-loop approach, as it leads to safer outcomes. Finally, in all likelihood, workers will be too time-poor, resource-poor, and lacking in the necessary expertise to meaningfully make use of a read-out of XAI output. Even if the aforementioned issues are overcome, transparency in the form of a 'right to explanation' is therefore likely to have limited practical value. As Edwards and Veale (2017) rightfully pointed out in the context of the EU General Data Protection Regulation (GDPR), 'a right to an explanation may be at best distracting, and at worst nurture a new kind of transparency fallacy' (p. 19).

Yet transparency is commonly advocated as a silver bullet to counter the adverse effects of automated, data-driven decision-making, as witnessed by the ethical guidelines for AI which have been multiplying over the past decade. Governments, such as Australia, Canada, and Singapore, as well as industry leaders like Microsoft, Google, and the Open Data Institute, have developed such guidelines. The core values typically advocated in ethical guidelines for AI are transparency, accountability and responsibility. It is still unclear whether such guidelines can assist in developing effective workplace oversight structures aligned with the human-in-command principle, as they typically rely on high-level statements with no clear system assessment criteria. Moreover, there has been concern that human oversight is not a sufficient condition to satisfy the three core values advocated in ethical guidelines. Ensuring meaningful human control may instead require the development of systems with dynamically adjustable levels of autonomy – switching anywhere between and including full autonomy or complete

teleoperation (Methnani et al. 2021). Finally, questions have been raised about the effectiveness and impact of guidelines, as they lack mechanisms to enforce their own normative claims. Looking at 22 examples of AI ethics guidelines, Hagendorff (2020) concludes that their legal and regulatory status varies across jurisdictions, and that their adoption is usually optional.

In Europe, the AI Act is the first legally binding attempt to regulate AI. It introduces a risk-based approach that imposes regulatory burdens when an AI system is likely to pose high risks to fundamental rights and safety. It sets four risk levels as thresholds for specific requirements. 'Unacceptable risks' lead to prohibited practices; 'high risks' trigger a conformity assessment with a long list of requirements; 'limited risks' meet specific transparency obligations; while 'minimal risks' lead to stakeholders being encouraged to build codes of conduct. The AI Act stipulates that high-risk systems 'should be designed to allow for oversight by humans who will be tasked with preventing or minimising risks'. This applies, for instance, to facial recognition or to algorithms that determine eligibility for public benefits. The draft act has sparked contentious discussions on whether it employs appropriate regulatory techniques, on the adequacy of its protective measures, and the scope of its application (Ruschemeier 2023). In its current form, the AI Act restrains or confines the qualification of high-risk software only when providers determine that the AI software is intended to be used in employment, workers' management and access to self-employment for the recruitment and selection of persons, for task allocation, monitoring and the evaluation of workers. As Cefaliello and Kullmann (2022) argue, there might be a difference between the provider's intended use and the employers' actual use of AI software. The authors recommend that, that even if software is not intended to be used for monitoring workers, the simple fact that it is foreseeable that this software will be deployed in a work-related contractual relationship should be sufficient to qualify it as high-risk.

More generally, the effectiveness of legal means in regulating rapidly evolving and dynamic technologies has raised systemic concerns. The opaque, complex and rapidly changing character of AI does not interact well with the legal imperatives of certainty, transparency, explicability and equal treatment (Ranchordás 2021).

While there is no consensus on the best regulatory approach, it is clear that effective oversight of AI is crucial to ensure that its development aligns with ethical principles and human values. This is all the more important as both corporations and states are driven by a competitive AI development dynamic and may therefore engage in a regulatory race to the bottom in pursuit of technical superiority. However, current guidelines and regulatory initiatives offer no implementable recommendations to handle the problem of control. On the technical side, XAI remains an active area of research and is still far from achieving the level of interpretability necessary to fully implement the human-in-command approach (Rudin 2019). As technical solutions are still far off and, in some cases, may never be found, ensuring meaningful human oversight will be a key challenge going forward.

# 5. Automation and workers' autonomy

The HIC approach also has implications for psychosocial work environment dynamics, as a lack of job control or autonomy is commonly reported as a work-related psychosocial factor associated with ill-health (Zwysen et al. 2024). According to the Job Demand-Control model (JDC) of Karasek (1979), one of the most cited models on occupational stress, workers in high demand jobs who lack autonomy over their work environments are particularly at risk of work-related stress. Job demands refer to the physical, psychological, and social aspects of work that require effort and can be potential sources of stress. Job control, on the other hand, refers to the degree of autonomy and decision-making authority that employees have over their work. It is defined by two key components: workers' ability to make decisions about their work (i.e., decision authority), and the breadth of skills used by workers on the job (i.e., skill discretion). According to the JDC model, high demands are particularly stressful when the worker has low control over job-related decisions.

The implementation of AI systems has the potential to fundamentally change the balance between job demands and control provided to workers. On the control side, it may have a negative impact on decision authority as the AI system becomes more involved in decision-making processes that were previously within workers' remits. In that sense, the implementation of AI systems entails a paradigm shift where workers transition from actively 'solving' the tasks to relying on the system's analytic capabilities for decision-making. Interacting with these systems therefore bears the risk of relegating operators from direct process control to a supervisory role. Consequently, the range of skills used by a worker on the job (i.e., skill discretion) is likely to narrow over time – further shrinking workers' opportunities to exert control over their work. The impact of automation on operational skills and job performance is well-documented (Nurski and Hoffmann 2022), for example among aircraft pilots and in relation to autonomous vehicles (Haslbeck and Hoermann 2016; Stanton 2019). By becoming supervisors of machines or algorithms, pilots gradually lose not only their fine-motor flying skills but also their operational understanding. Consequently, their ability to detect errors or take over in a case of system failure degrades, undermining their task control abilities and cultivating technological dependence (Parker and Grote 2020). Moreover, the lack of transparency in the system's decision-making processes is likely to impede workers' attempt to maintain a sense of control over their job. The opaque nature of AI systems makes it challenging, if not impossible, for workers to gain insights into the decision-making process. Without visibility into how the decisions are made, workers may feel disempowered and uncertain about the outcomes – further hindering their ability to effectively collaborate

with AI technologies. Finally, advances in AI-enabled technologies offer new avenues for exerting tighter control over work activities. The capabilities of AI systems to monitor, track, and evaluate worker performance in real time enable unprecedented levels of surveillance and oversight – further eroding workers' autonomy, as evidenced in many studies (Bérastégui 2021a).

On the demand side, workers interacting with AI systems may experience higher workloads as they strive to match the efficiency and speed of AI-based processes. Moreover, high monitoring demands associated with these systems can put further strain on workers. The constant vigilance required for effective monitoring can be draining, limiting workers' ability to fully engage in other essential job responsibilities. Similarly, having to learn how to monitor and interact with AI systems is likely to generate additional demands supplementing other job responsibilities. Workers will be required to acquire new skills and learn to navigate the complexities of integrating AI into their workflow. This adjustment process can be demanding and time-consuming, placing a burden on workers to quickly acquire the necessary knowledge and competencies to work alongside AI systems. Finally, constant monitoring of AI systems can create a sense of being constantly watched and evaluated, possibly resulting in increased stress.

Characterised by greater demands and lower control over work, the widespread implementation of AI systems may give rise to high job strain, in turn associated with stress-related ill-health. This has become a growing concern in the post-pandemic world, as digital technologies, including AI, have become more prevalent in the workplace. In April 2022, the European Agency for Safety and Health at Work commissioned a Flash Eurobarometer survey with the aim of gaining more insights into the state of OSH in this evolving landscape (EU-OSHA 2022). The survey aimed to investigate various aspects, including the mental health stressors associated with the use of digital technologies at work. It found that 33% of respondents perceived the introduction of digital technologies as having increased their workloads, while 19% reported a decrease in work autonomy. Pre-pandemic evidence confirms that the use of digital technologies in the workplace is frequently associated with psychosocial risks. Time pressure is an issue for 54.5% of companies where digital systems are used to determine the content or pace of work, and for 57.1% of companies using systems to monitor workers' performance (Irastorza 2019). This is in line with the extensive body of research conducted into the platform economy, showing that the delegation of managerial functions to algorithmic and automated systems contributes to a hectic pace of work and a lack of control over job-related decisions (Bérastégui 2021a).

The German survey 'Digitalisation and Change in Employment' (DiWaBe) provides additional evidence of the impact of automated systems on job control and demands (Arntz et al. 2020). Exploring the use of automation technologies in German companies, this survey contains a wide range of questions related to both physical and psychological aspects of work. Specifically, respondents were asked questions on how often technology makes decisions about their work process and gives instructions to them. The findings show that receiving instructions from automated systems is a significant predictor for all facets of job control (Niehaus et al. 2022). It is associated with less freedom in organising one's work, influencing

the working speed, the possibility of choosing between different task approaches and influencing the amount of work, and higher levels of repetition of working steps. Regarding job demands, the study found that higher levels of instructions from automated systems were associated with increased physical stress but a slight decrease in multitasking. This suggests that the introduction of automated systems for decision-making leads to more standardised and streamlined processes, allowing workers to dedicate less attention to parallel subtasks and focus more on the physical actions required.

A recent survey by Piasna (2024) shows that the ways in which information and communication technologies (ICTs) are integrated into work processes differs from one country to another. In Germany, Luxembourg and the Netherlands, for instance, ICTs exert little influence on workers despite being used frequently. In Romania, on the other hand, the use of computers is less common, though the extent of their influence is disproportionately high. A similar pattern is found in some other Eastern and Southern European countries, such as Lithuania, Poland, Portugal and Spain. According to the author, ICT control over work processes can be more contained in countries where individual control is anchored in broader industrial relations structures that are more strongly institutionalised, thus better protecting workers from various pressures, presumably including those stemming from increasing digitalisation.

What these surveys also show is that the complete automation of jobs is currently rare. In the majority of cases, workers undergo a reallocation and reconfiguration of the tasks that together form their occupations. Implications for job control and workers' autonomy will therefore vary depending on how these tasks are recomposed and organised. In this regard, the aforementioned findings highlight that a substantial share of automation happens at the decision-making level. One illustrative example is the implementation of voice picking technology to improve the accuracy and efficiency of order-picking processes in warehouses and distribution centres. Voice picking or pick-by-voice is an order fulfilment method in which workers use a voice recognition system to receive instructions and confirm task completion. It consists of a GPS tracking device and a microphone headset linked to the warehouse management system (WMS) sending real-time instructions on picking locations and tasks. The worker is directed to a designated picking location within the warehouse. After vocally confirming arrival, they are told what to pick. After confirming picking, they are directed to the next location. This cyclical process repeats as the system proceeds to deliver instructions for the subsequent task, reducing non-productive times and ensuring a seamless and efficient workflow. In this case, the system acts as a decision-making component while the worker's role becomes entirely focused on executing the tasks. The system determines the optimal picking sequence and provides step-by-step instructions to the workers. Workers' decision-making autonomy is therefore significantly reduced, as they no longer have the freedom to make strategic decisions about order picking and sequencing. Although the voice recognition technology may involve machine learning, voice picking systems used in today's warehouses are typically not based on AI. It is a form of automation that relies on a specific set of instructions and responses instead of the more complex problem-solving capabilities of AI. However, as AI systems become more integrated into work

processes, workers may find themselves assigned to more routine and repetitive tasks while the complex and strategic aspects of decision-making are taken over by the AI system.

Restricting the range of decisions workers are actively involved in is not only detrimental to job control but can also result in higher demands with potential adverse OSH effects. Several studies underline that voice picking systems are leading to a hectic pace of work and a higher risk of accidents. The continuous flow of instructions and instant initiation of the next task accelerates the work pace, while ensuring that workers' attention is permanently focused on the task at hand. A report from the Center for Investigative Reporting revealed a mounting injury crisis at Amazon warehouses, especially at robotic facilities (Evans 2020). Analysing weekly data from 2016 through 2019 from more than 150 Amazon warehouses, the report shows that grabbing and scanning operations have increased from 100 to 400 per hour in robotic fulfilment centres, and that the rate of serious injuries was more than 50% higher in these warehouses. The data backs up accounts of Amazon workers and former safety professionals saying the company has used these technologies to ratchet up production quotas to the point where workers can't keep up without hurting themselves.

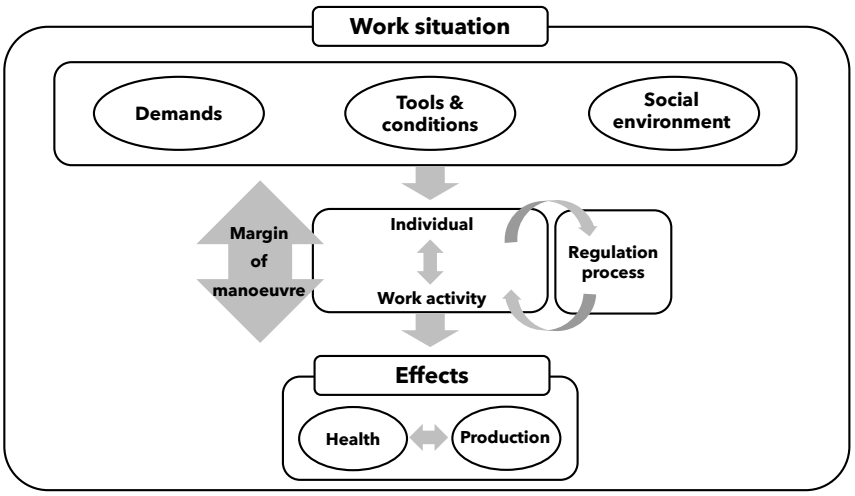
The partial automation of product-picking jobs in Amazon warehouses has also been found to reduce human-to-human interactions, with workers expressing discontent over the lack of social relationships at work (TKI DINALOG 2020). Other studies show how the introduction of voice recognition systems in call centres has done away with the need for emotional and interpersonal skills (Hernandez and Strong 2018; De la Garza 2019). Such systems were used to monitor customers' and agents' conversations for emotional cues and provide feedback on the appropriateness of operators' responses, in the shape of instructions: 'talk more slowly', 'display higher alertness' or 'say something empathetic'. This not only reduces worker discretion over how to respond to customers, but also does away with the need for emotional and interpersonal skills to judge a customer's mood and choose how to react to it. The social dimension of work is known to moderate the negative impact of job strain on workers' physical and mental health (Johnson and Hall 1988). Specifically, the most at-risk group of poor physical and mental health are those workers who are exposed to job strain (high demands and low control) paired with low workplace support – a phenomenon referred to as iso-strain.

The automation of analytical tasks may also result in a shift from active work to passive monitoring jobs more prone to performance deterioration and sleepiness. Monotonous or intrinsically unstimulating tasks require greater cognitive control than intrinsically stimulating tasks, as individuals have to self-regulate their task engagement (Bérastégui 2019). Failure to adequately self-regulate engagement results in lower performance levels and heightened susceptibility to sleepiness, ultimately impacting the safety of work. This has been a concern in autonomous driving, as today's vehicles are only partially automated and require the human driver to monitor the road and take over at a moment's notice. A study conducted by the Fatigue Countermeasures Lab at NASA's Ames Research Center suggests that the passive role of drivers during autonomous driving make them more

susceptible to sleepiness – especially when they are sleep-deprived (Flynn-Evans et al. 2021). When just supervising the vehicle (passive driving), participants reported feeling sleepier, showed increased signs of 'nodding off' as well as slower reaction times compared to actively driving the car. The more sleep-deprived a person was, the stronger these effects were. These findings highlight the need to develop countermeasures so that workers in partially autonomous systems stay alert and engaged.

In the French activity-centred ergonomic field, the notion of 'marges de manoeuvre' has been proposed to describe how the interaction between high demands and low control has an adverse impact on health and safety (Figure 5). Margins of manoeuvre act as a regulation space allowing for continual adaptation to variations in job demands and resources, as well as to variations in a worker's own health or condition (Durand et al. 2016). By having sufficient margins of manoeuvre, workers can engage in self-initiated proactive strategies to maintain a balance between productivity and well-being. This leeway is reflected, for instance, in an operator's ability to adapt their gestural activity, temporarily alter their work pace, or employ alternative strategies in response to changing demands. Insufficient margins of manoeuvre, on the other hand, disrupt the regulation loop and deprive workers of any means to address a potential imbalance. It has been found that workers involved in activities with low margins of manoeuvre are more likely to be absent from work due to pain or disability (Schultz and Gatchel 2015). Having sufficient margins of manoeuvre is a crucial factor in rehabilitation and return-to-work programmes. A lack of such margins is associated with worries about returning to work, less chances of actually returning to work, as well as poor work outcomes for those returning to work (Coutu et al. 2023).

Figure 5 **Model of work activity and margin of manoeuvre**




Source: adapted from Durand et al. 2009.

In management sciences, workers' ability to influence various aspects of their work has been investigated under the lens of job crafting, defined as a process of 'autonomous and proactive change that the worker carries out when they

understand that the realisation of these changes is possible' (Letona-Ibañez et al. 2021). In other words, job crafting refers to the process by which individuals proactively shape and mould their jobs to be more in line with their skills, interests or preferences. It involves making intentional changes to tasks, relationships, and perceptions of one's work to create a more satisfying and meaningful job experience. Job crafting is regarded as a key underlying mechanism, promoting job-level autonomy for the ultimate benefit of individual-level outcomes (Dierdorff and Aguinis 2018). Workers with limited involvement in decision-making and primarily responsible for executing predetermined tasks may feel a sense of disconnection from their work as well as a lack of fulfilment. The absence of opportunities for creativity, problem-solving, and autonomous decision-making can lead to reduced job satisfaction and a diminished sense of professional growth and development. Conversely, the ability to craft one's job is associated with higher levels of work meaning, engagement and performance as well as lower levels of absenteeism (Shang 2022; Rogala and Cieslak 2019).

The job crafting literature has recently taken up the issue of AI and automation, expressing concerns over the negative impact of self-learning algorithms on workers' autonomy and their ability to actively shape their work (Parent-Rocheleau and Parker 2021; Parker and Grote 2020). In an interview study, Perez et al. (2022) investigated how the introduction of learning algorithms affected the jobs of bank customer advisors selling financial products and services. It was found that the algorithm initially reduced their autonomy by telling them which customers to contact and what to propose, monitoring their actions and reporting them to managers by means of a software application. Employees perceived the change as undermining their expertise and with it the meaning of their work, as the algorithm was intended to replace their knowledge about customers by AI-based predictions. Interestingly, initial reductions in autonomy were reversed over time as initial challenges to autonomy were met with employee job crafting practices accepted by managers. Similarly, employees changed the meaning of their work over time – away from expertise grounded in technical banking knowledge and towards more in-depth knowledge of a customer's needs and life plan. While the initial findings were in line with the pessimistic assessments of AI limiting workers' autonomy and increasing management control, the latter ones showed employees using job crafting behaviour to rebalance their autonomy and shift to a different meaning of work. However, such a turnaround was only possible through active management support and an organisational culture that encouraged and embraced employee job crafting practices. The receptiveness of management to the needs and preferences of employees allowed for a collaborative and iterative change process, where employees were empowered to reshape their roles and tasks within the context of AI and automation.

This underlines the importance of involving workers in the design and implementation of AI systems to foster a sense of ownership and engagement. In fact, meaningful worker participation has been shown to mitigate the negative consequences of AI adoption. Findlay et al. (2017) for instance described how unions safeguarded workers' contracts and remuneration during the partial automation of a pharmaceutical dispensary. It is important for workers to be involved in such early-adoption stages as application design or selection. According to Nurski and

Hoffman (2022), the earlier workers are involved, the greater the chance that their perspectives will be incorporated into new technologies. But beyond the adoption process, ongoing worker engagement and participation are crucial for ensuring the successful integration and minimising the risks of AI systems. Colclough (2020) argues for a co-governance model following the human-in-command principle and entailing regular assessments of unintended consequences, as well as strategies to mitigate them.

Reviewing specific organisational applications of AI systems, Nurski and Hoffman (2022) concluded that the effects of these systems tend to be more negative when they are designed using prescriptive use cases rather than supportive use cases. This finding aligns with the extensive literature on workers' autonomy, which consistently shows that processes that are highly prescriptive and do not allow workers to apply their skills or exercise judgment are associated with poorer job performance and health outcomes. However, the distinction between prescriptive and supportive AI applications is not always clear-cut. Distinct elements of the same system can exhibit different levels of prescriptiveness and supportiveness. For instance, a voice picking system may incorporate real-time updates on stock availability. Such a feature could be framed as supportive if it allows workers to anticipate and adapt their picking strategies based on current circumstances. The ‘categorisation’ of an AI system as prescriptive or supportive also depends on the context of its implementation and how the change process is managed. In line with the job crafting literature, a wide range of factors are likely to play a role, such as the degree of decision-making authority granted to workers, the level of flexibility in task execution, and the extent to which the system allows for worker input and adaptation. Use cases may therefore significantly evolve over time from what was initially planned, provided that workers are empowered to shape the change process. Rather than sharply distinguishing them, prescriptive and supportive use cases should therefore be imagined along a continuum, with each point on that continuum reflecting a blend of supportive and prescriptive features, as well as varying opportunities for workers to steer that combination. According to Demerouti (2020), ‘digitalisation and automation can contribute to stimulating and ‘healthy’ jobs if their implementation is designed in a way that increases resources and reduces demands, and if people are in control and craft their use of the system’.

In sum, research has consistently demonstrated that highly automated tasks and jobs with rigid, standardised processes tend to limit workers' autonomy, as the decision-making and control are transferred to AI algorithms. Advances in AI-enabled technologies therefore offer new avenues for exerting tighter control over work activities – a phenomenon reminiscent of the principles of scientific management proposed by Taylor in the late nineteenth century. Taylor emphasised three key aspects: gathering knowledge through detailed measurement of the work, concentrating this knowledge in the hands of managers, and using this monopoly to control each step of the labour process and its mode of execution (Braverman 1998). In the same way, the prescriptive view of AI-enabled work shifts knowledge from operators to the system itself – transforming workers into mere executors in a cyber-physical system. As described in this section, the consequences of such a Taylorism 4.0 have been discussed from different perspectives and disciplines,

all hinting at negative worker outcomes. But how do we get human and artificial intelligence to work together, and how can these interactions be designed so that humans retain ownership of their work? Acknowledging and addressing these issues is crucial for effectively managing the integration of AI systems in the workplace and ensuring a safe and fulfilling work environment for workers.

# 6. AI and the displacement of precarious jobs

AI is becoming a central part of many companies' operations, with businesses across various sectors adopting AI systems to reduce costs and improve efficiency. Eurostat's Community Survey on ICT Usage and E-Commerce in Enterprises inquiries about the use of several AI technologies by EU enterprises (Eurostat 2021). In 2021, 8% of enterprises with more than 10 employees used AI technologies. The most cited use cases were the automation of workflows, the analysis of written language, and machine learning. The survey also shows that larger companies are more likely to use some form of AI technology, with 28% of firms with more than 250 employees reporting their use – suggesting that there are substantial costs and organisational barriers involved in adopting AI technologies. A smaller survey by the European Commission shows that skills and financial constraints are the leading reported barriers. About 80% of EU companies cite a lack of skills among existing staff and in the external labour market, as well as the high cost of buying the technology and adapting their operational processes (European Commission 2020).

While AI adoption in the EU remains relatively low, there is growing consensus that AI is set to transform entire sectors and the economy. Technologies such as machine learning, natural language processing, and computer vision are increasingly being used by businesses to automate work processes and workflows. For instance, McDonald's began testing AI-powered drive-thrus relying on voice recognition and natural language processing to take orders from customers (Metz 2021). With drive-thrus accounting for an increasing share of fast-food sales, major food chains are viewing AI as the technology that could make 2020s the golden age of drive-thru. Between 2019 and 2020, McDonald's managed to reduce average drive-thru time from 6 minutes and 18 seconds to 5 minutes and 49 seconds. Similarly, JP Morgan Chase is using AI-powered systems to automate many of its back-office tasks, such as data entry and document processing (Galeon 2017). The system uses natural language processing and machine learning to analyse and extract information from documents, reducing the need for human workers to perform these tasks. Generative AI is also increasingly used in customer service to streamline and enhance support through applications like chatbots, or to automatically generate documents such as earning summaries (Rai 2023).

Although specific estimates vary, several studies suggest that the automation of work will lead to substantial job losses in certain sectors. Analysing over 200,000 jobs in 29 countries, PwC concluded that 30% of jobs could be automated by the mid-2030s (PwC 2018). More recently, Goldman Sachs came up

with a more conservative estimate of 18%, with as many as 300 million jobs affected in some way worldwide (Hatzius et al. 2023). Similarly, the OECD estimates 14% of jobs are at high risk of being automated (OECD 2021). According to the OECD, jobs in manufacturing and agriculture are at a higher risk of automation, although those in several service sectors, such as postal and courier services, land transport and food services are also considered to be at high risk. The analysis shows that employment growth has been much lower in occupations at a high risk of being automated (6%) than in occupations at low risk (18%). Low-educated workers are increasingly concentrated in occupations at a high risk of being automated, although the lower employment growth in these jobs has not led to a drop in the employment rate of low-educated workers. This is mainly because the number of workers with a low education has fallen in line with demand for these workers. Going forward, however, the risk of automation eliminating jobs increasingly affects low-educated workers.

As AI technologies continue to advance, there will be a growing need for workers with the skills and knowledge to develop, implement, and maintain AI systems. As such, demand for professionals with expertise in AI and machine learning is expected to grow significantly in the coming years. This includes not only data scientists and engineers, but also professionals in fields such as ethics, law or policy who can help ensure that AI is developed and used in a responsible and ethical manner. Additionally, AI is expected to drive innovation and create new business models, leading to the creation of new jobs and industries. The need for a diverse and skilled workforce able to adapt to these changes will become increasingly important. It is a common ascertainment that high-skilled and educated workers will be able to meet these new technological requirements and enjoy higher wages, while less educated and lower-skilled workers will be burdened by the cost of automation and more exposed to income loss and unemployment (Zervoudi 2020).

When it comes to AI-driven automation, the dominant narrative in policy discourse depicts a win-win scenario where workers are freed from monotonous and repetitive tasks, while businesses benefit from increased efficiency and productivity. With AI taking care of tedious tasks, human workers are supposed to be able to focus on more creative and rewarding tasks. However, contradicting the dominant narrative, recent evidence suggests that the development of AI does not necessarily mean the end of menial work due to automation but rather its offshoring to developing countries.

Progress in AI depends heavily on machine learning, and thus on the availability of large datasets that algorithms can learn from. For instance, an AI algorithm tasked with recognising cats in photographs must, during its training phase, be fed photos that are known to represent cats and photos that are known not to represent cats. In this way, the model will not only be able to train but also to self-assess its performance and refine the detection process. Training data needs to be collected, sorted, verified and translated into a form that the AI system can assimilate. Such time-consuming and menial tasks are often outsourced to developing countries. A recent investigation by the American news magazine Time revealed that Kenyan workers paid between around $1.32 and $2 per hour were responsible for ensuring

that the data used to drive ChatGPT did not contain discriminatory content (Perrigo 2023). This work was vital as the AI system was trained on hundreds of billions of words scraped from the internet – a vast repository replete with toxicity and bias. In addition to being poorly paid, the Kenyan data workers have to go through the darkest recesses of the internet including text descriptions of child sexual abuse, bestiality, murder, suicide, torture, self-harm and incest.

Data workers are at the tail-end of a long outsourcing chain, which partly explains the low wages. Indeed, the AI business involves many actors: the GAFAM offering cloud solutions for data hosting and computing power, tech companies selling the AI models, and companies offering data annotation services – each intermediary capturing part of the value produced. In France for instance, data annotation is mainly outsourced to service providers located in Madagascar because of the large number of organisations offering these services and the low cost of skilled labour (Bérastégui 2023). French AI tech firms benefit from well-trained workers: most have gone to university and are fluent in French learned at school, via the Internet or through the '*Alliances françaises*' network. Madagascan companies are very dependent on their French clients who manage this outsourced workforce almost directly, with dedicated middle management positions within Parisian start-ups. The fact that these positions are filled by foreigners, either employed by the client companies in France or by expatriates on the spot, represents a major obstacle to the career development opportunities of data workers who remain stuck in the lower levels of the value chain. In addition to these 'formal' companies, the sector has developed around a mechanism of cascading subcontracting with, at the end of the chain, informal companies and individual entrepreneurs even less well treated and only mobilised in the event of a lack of manpower in formal companies.

Historically, data annotation has been massively outsourced via the platform economy, and in particular via the crowdworking giant Amazon Mechanical Turk (AMT) (Bérastégui and Garben 2021). The driving principle of crowdwork, also referred to as microwork, is to break down large volumes of time-consuming work into smaller tasks distributed to a pool of unqualified workers. As early as 2015, the tasks most often traded on AMT pertained to identifying information in images (37%), followed by transcribing audio or video material (26%) and lastly, classifying images (13%) (Hitlin 2016). Microsoft senior researcher Mary L. Gray described crowdwork as 'the last mile of automation', as it concerns the residual tasks of larger data processing operations that unskilled humans can still solve more cheaply and with a lower error rate than computers (Schmidt 2017). But in other cases, the result of their work is actually fed into learning algorithms, enabling further automation. Indeed, crowdwork has proved to be an infinite source of human knowledge that machine learning desperately relies on to make progress. This explains why, despite being wide-ranging, micro-tasks are often thankless, repetitive and low-skilled. Somewhat ironically, crowdworkers thereby contribute to the development of leading-edge technologies meant to substitute them.

Working conditions in the platform economy are notoriously precarious, and crowdwork is no exception. Beyond the issue of bogus self-employment, platform workers are exposed to a wide range of psychosocial risk factors. Looking behind

those specific risks, the guiding thread is a greater imbalance between the job demands placed upon workers and the resources available to deal with them (Bérastégui 2021). In contrast to traditional employers, online labour platforms provide workers with few organisational resources. They provide no workplace support, no channels to voice their concerns or exercise agency, no means of contesting unfair decisions or unethical behaviours, and do not guarantee any form of job security. Platforms nevertheless have high standards of performance and require workers to be highly autonomous, flexible, affable and productive. A growing corpus of research shows the detrimental impact of such an imbalance on workers' safety and health.

In sum, data workers face particularly poor working conditions despite their fundamental role in ensuring that AI systems are safe and reliable. Whether they are hired through labour platforms or in developing countries, they face sub-standard working conditions and do not capture any of the benefits of innovation – confirming the long-standing logic of outsourcing chains. Regardless of how it is procured, data work is essential to the production and maintenance of AI-based systems – including those used in the industrial sector. Discussing the OSH implications of the transition to I4.0 therefore implies touching upon the human work hidden behind AI's layers of knowledge (Le Ludec et al. 2023). Partnership of AI, a coalition of organisations active in the sector, highlighted an 'out-of-sight, out-of-mind' situation with ongoing efforts of tech firms to 'hide AI's dependence on this large labour force when celebrating the efficiency gains of technology' (Perrigo 2023). According to Le Ludec and Cornet (2023), 'making the involvement of these workers visible means questioning globalised production chains, which are well known in the manufacturing industry but also exist in the digital sector'. The authors further argue that a truly ethical AI must involve an ethic of data work. In the same vein, Krzywdzinski et al. (2023) argues for "human-centred' AI with a primary focus on the well-being and dignity of people, 'bringing social benefit and preserving the self-determination of people as agents and their freedom to make decisions'. Addressing these aspects is essential to safeguard working conditions and ensure a socially responsive transition to the new world of work (Bérastégui 2021b).

# Conclusion

AI cannot be dissociated from other related technologies such as cloud computing and the internet of things. Together, these technologies form the so-called '4.0' environment that allows for new ways of organising the means of production. Looked at from the perspective of various sectors, these emerging ecosystems promise to unlock unparalleled efficiency gains and substantially reduce costs. The vision is one of highly flexible and agile production systems capable of harvesting real-time data from a network of interconnected devices and, through feedback, interacting with physical processes to adapt to new operating conditions. Such systems would allow increased adaptability and responsiveness, mass customisation and shorter innovation cycles, while improving resource efficiency. Most large-sized companies in Europe have rolled out investment plans aligned with this vision, guided by a profusion of 'maturity models' and 'road maps' (Cotrino et al. 2020). In addition to lacking a fundamental empirical evaluation, I4.0 maturity models tend to focus exclusively on the technical aspects of the transition – in which human factors are commonly overlooked (Walter et al. 2020; Hellweg et al. 2023). While growing automation is at the core of what makes I4.0 so revolutionary, human workers will remain a critical component in these new environments. Contemporary examples show that the complete automation of jobs is rare as workers rather undergo a reallocation and reconfiguration of the tasks that together form their occupations. As 'end users' of these semi-autonomous systems, workers are poised to bear the brunt of the transition. Integrating I4.0 technologies into industrial processes radically transforms the way work is organised, including the nature and frequency of human-machine interactions. In this context, sidelining the social dimension of the transition could result in new threats to workers' occupational safety and health.

The aim of this paper was to shed light on a selected set of those areas of concerns. It showed that incidents resulting from faulty AI systems have already emerged in various contexts and are likely to grow over the coming years, both in terms of likelihood and severity. Robustness, assurance and specification failures are well-known issues in the field of AI safety and, as of today, there is no silver bullet technology able to completely solve these vulnerabilities. Besides faulty behaviours, the introduction of AI-enabled technologies is leaving companies more vulnerable to malicious actors. The growing convergence of IT and OT broadens the attack surface and exposes companies to hybrid attacks targeting both physical and cyber assets. Such attacks have the potential to compromise worker safety and therefore call for ad hoc solutions beyond standard information security. The mitigation of these risks is further hindered by the inherent opacity of AI systems, offering limited opportunities to reintroduce human control in the loop. The various

principles put forward in regulatory initiatives and ethical guidelines to ensure proper human oversight appear to be far distant from the technical possibilities of XAI. Yet knowing the way an algorithm arrives at a particular output is crucial, not only to detect any biases but also to build trust and confidence in AI systems. Ensuring adequate levels of trust will be a key challenge, as both distrust and over-trust in automated system have been shown to be detrimental to OSH. Looking at job quality, it appears that the automation of menial tasks is not the primary use case of AI systems as of now. A substantial part of automation is happening at the decision-making level, with the aim of standardising and streamlining work processes. As illustrated by current developments in the e-commerce sector, relegating strategic aspects of decision-making to AI is not only detrimental to workers' sense of job control but can also result in higher demands and increased risk of accidents due to a more hectic pace of work. Further away from the win-win narrative dominating policy discourse, recent evidence suggests that the development of AI does not mean the end but rather the displacement of menial work. Data-labelling jobs are fast becoming part of a new economy of machine learning, mainly outsourced to developing countries or via the platform economy. The workers concerned find themselves at the end of a long outsourcing chain, with low wages and particularly poor working conditions.

Despite these many challenges, AI systems are nonetheless being deployed in many sectors, including in critical domains where even small errors can have severe consequences for workers and society as a whole. The introduction of AI-enabled technologies in self-driving vehicles or at a nuclear power plant is already raising issues of how to manage the growing uncertainties associated with human-machine interactions. More interactions are to be expected between humans and black box systems as AI systems slowly become mainstreamed – beyond early adopters. The possibilities to further automate decision-making within work processes will therefore be massively increased, with the risk of more unfavourable working conditions. In this context, the transition to I4.0 should put more emphasis on how to get people and technology working safely together. Attention should be paid to the simultaneous development of technological and human capabilities, with a view to ensuring that people can become or continue to be masters of their own work. Involving workers in the design and implementation of AI systems is of paramount importance, as meaningful worker participation through job crafting practices has been shown to mitigate the negative consequences of AI adoption. Addressing these issues will require greater insights from policymakers into the current state of the art of AI, its limitations, and the challenges it poses to OSH.

## References

Adadi A. and Berrada M. (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE Access, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Agrafiotis I., Nurse J.R.C., Goldsmith M., Creese S. and Upton D. (2018) A taxonomy of cyber-harms: defining the impacts of cyber-attacks and understanding how they propagate, Journal of Cybersecurity, 4 (1). https://doi.org/10.1093/cybsec/tyy006

Amodei D. et al. (2016) Concrete problems in AI safety, ArXiv. https://doi.org/10.48550/arXiv.1606.06565

Arnold Z. and Toner H. (2021) AI accidents: an emerging threat, Center for Security and Emerging Technology. https://doi.org/10.51593/20200072

Arntz M. et al. (2020) Digitalisierung und Wandel der Beschäftigung (DiWaBe): eine Datengrundlage für die interdisziplinäre Sozialpolitikforschung. Datenreport und Forschungspotenzial, ZEW, Leibnis-Zentrum für Europäische Wirtschaftsforschung. https://www.baua.de/DE/Angebote/Publikationen/Kooperation/DiWaBe.html

Atkinson E. (2023) Man crushed to death by robot in South Korea, BBC, 8.1.2023. https://www.bbc.com/news/world-asia-67354709

Bach S. et al. (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLOS ONE, 10 (7), e0130140. https://doi.org/10.1371/journal.pone.0130140

Badri A., Boudreau-Trudel B. and Souissi A.S. (2018) Occupational health and safety in the industry 4.0 era: a cause for major concern?, Safety Science, 109, 403–411. https://doi.org/10.1016/j.ssci.2018.06.012

Bär M. et al. (2021) The influence of using exoskeletons during occupational tasks on acute physical stress and strain compared to no exoskeleton – A systematic review and meta-analysis, Applied Ergonomics, 94, 103385. https://doi.org/10.1016/j.apergo.2021.103385

Barredo Arrieta A. et al. (2019) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion, 58. https://doi.org/10.1016/j.inffus.2019.12.012

Basit A. et al. (2021) A comprehensive survey of AI-enabled phishing attacks detection techniques, Telecommunication Systems, 76 (1), 139-154. https://doi.org/10.1007/s11235-020-00733-2

Batko K. and Ślęzak A. (2022) The use of big data analytics in healthcare, Journal of Big Data, 9 (1), 3. https://doi.org/10.1186/s40537-021-00553-4

Beaudreau B. (2018) A pull-push theory of industrial revolutions, Proceedings of International Academic Conferences 7508525, International Institute of Social and Economic Sciences. https://iises.net/proceedings/iises-annual-conference-sevilla/table-of-content/detail?cid=75&iid=006&rid=8525

Bérastégui P. (2019) La gestion du risque associé à la fatigue en médecine d'urgence : identification et évaluation de pratiques informelles, Doctoral thesis, University of Liège. https://hdl.handle.net/2268/236178

Bérastégui P. (2021a) Exposure to psychosocial risk factors in the gig economy: a systematic review, Report 2021.01, ETUI. https://www.etui.org/publications/exposure-psychosocial-risk-factors-gig-economy

Bérastégui P. (2021b) Gig workers: guinea pigs of the new world of work, Social Europe, 18.02.2021. https://socialeurope.eu/gig-workers-guinea-pigs-of-the-new-world-of-work

Bérastégui P. (2023) Survey on the working conditions of data workers, ETUI News, 26.04.2023. https://www.etui.org/news/survey-working-conditions-data-workers

Bérastégui P. and Garben S. (2021) The platform economy at the forefront of a changing world of work: implications for occupational health and safety, in Drahokoupil J. and Vandaele K. (eds.) A modern guide to labour and the platform economy, Edward Elgar Publishing, 96–111. https://doi.org/10.4337/9781788975100.00015

Braverman H. (1998) Labour and monopoly capital: the degradation of work in the twentieth century, Monthly Review Press.

Brown T. et al. (2020) Language models are few-shot learners, Advances in Neural Information Processing Systems, 33. https://arxiv.org/abs/2005.14165

Cao S. (2023) ChatGPT has kicked off a big tech AI race for search, Observer, 2.07.2023. https://observer.com/2023/02/microsoft-google-big-tech-launch-chatgpt-competitor/

Cebulla A. et al. (2022) Applying ethics to AI in the workplace: the design of a scorecard for Australian workplace health and safety, Ai & Society, 38, 919–935. https://doi.org/10.1007/s00146-022-01460-9

Cefaliello A. and Kullmann M. (2022) Offering false security: how the draft artificial intelligence act undermines fundamental workers rights, European Labour Law Journal, 13 (4), 542–562. https://doi.org/10.1177/20319525221114474

Cheng W.-J., Pien L.-C., Kubo T. and Cheng Y. (2020) Trends in work conditions and associations with workers' health in recent 15 years: the role of job automation probability, International Journal of Environmental Research and Public Health, 17 (15), 5499. https://doi.org/10.3390/ijerph17155499

Cheng W.-J., Pien L.-C. and Cheng Y. (2021) Occupation-level automation probability is associated with psychosocial work conditions and workers' health: a multilevel study, American Journal of Industrial Medicine, 64 (2), 108–117. https://doi.org/10.1002/ajim.23210

Christiano P. et al. (2017) Deep reinforcement learning from human preferences, arXiv. https://arxiv.org/abs/1706.03741

CIGREF (2019) IT/OT convergence: a fruitful integration of information systems and operational systems. https://www.cigref.fr/cigref-report-it-ot-convergence-a-fruitful-integration-of-information-systems-and-operational-systems

Colclough C. (2020) Workers' rights: negotiating and co-governing digital systems at work, Social Europe, 3.09.2020. https://socialeurope.eu/workers-rights-negotiating-and-co-governing-digital-systems-at-work

Conran M. (2021) The rise of artificial intelligence DDoS attacks, Networkworld, 11.07.2018. https://www.networkworld.com/article/3289108/the-rise-of-artificial-intelligence-ddos-attacks.html

Cotrino A., Sebastián M.A. and González-Gaya C. (2020) Industry 4.0 roadmap: implementation for small and medium-sized enterprises, Applied Sciences, 10 (23). https://doi.org/10.3390/app10238566

Coutu M.F. et al. (2023) Workers' worries, pain, psychosocial factors, and margin of manoeuvre, in relation to outcomes in a return-to-work program: an exploratory study, Journal of Occupational Rehabilitation. https://doi.org/10.1007/s10926-023-10155-x

Cruz Lima J. and Stewart W. (2024) ROBO-CRASH watch moment AI-powered 'smart tram' ploughs into crowd of pedestrians leaving woman trapped under its wheels, The Sun, 14.04.2024. https://www.thesun.co.uk/news/27321928/ai-powered-smart-tram-ploughs-into-crowd-woman-trapped/

Cuvelier L. and Woods D.D. (2019) Sécurité réglée et/ou sécurité gérée : quand l'ingénierie de la résilience réinterroge l'ergonomie de l'activité, Le travail humain, 82 (1), 41–66. https://doi.org/10.3917/th.821.0041

Dahl F. (2011) Stuxnet could harm nuclear safety: U.N. atom chief, Reuters, 1.02.2011. https://www.reuters.com/article/us-nuclear-amano-iran-interview-idUKTRE7104UL20110201

Damioli G., Van Roy V. and Vertesy D. (2021) The impact of artificial intelligence on labor productivity, Eurasian Business Review, 11 (1), 1–25. https://doi.org/10.1007/s40821-020-00172-8

Dastin J. (2018) Amazon scraps secret AI recruiting tool that showed bias against women, Reuters, 11.10.2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

De la Garza A. (2019) This AI software is 'coaching' customer service workers. Soon it could be bossing you around, too, Time, 8.07.2019. https://time.com/5610094/cogito-ai-artificial-intelligence/

Del Ferraro S., Falcone T., Ranavolo A. and Molinaro V. (2020) The effects of upper-body exoskeletons on human metabolic cost and thermal response during work tasks—a systematic review, International Journal of Environmental Research and Public Health, 17 (20), 7374. https://doi.org/10.3390/ijerph17207374

Deloitte (2020) Thriving in the era of pervasive AI. State of AI in the enterprise, 3rd ed. https://www2.deloitte.com/cn/en/pages/about-deloitte/articles/state-of-ai-in-the-enterprise-3rd-edition.html

de Looze M.P., Krause F. and O'Sullivan L.W. (2017) The potential and acceptance of exoskeletons in industry, in González-Vargas J. et al. (eds.) Wearable robotics: challenges and trends, Biosystems & Biorobotics, vol. 16, Springer, 195–199. https://doi.org/10.1007/978-3-319-46532-6_32

Demerouti E. (2020) Turn digitalisation and automation to a job resource, Applied Psychology, 71 (4), 1205–1209. https://doi.org/10.1111/apps.12270

Dierdorff E.C. and Aguinis H. (2018) Expanding job crafting theory beyond the worker and the job, Management Research: Journal of the Iberoamerican Academy of Management, 16 (3). https://doi.org/10.1108/MRJIAM-08-2017-0773

Dong F., Zhang S., Zhu J. and Sun J. (2021) The impact of the integrated development of AI and energy industry on regional energy industry: a case of China, International Journal of Environmental Research and Public Health, 18 (17), 8946. https://doi.org/10.3390/ijerph18178946

Durand M.J. et al. (2009) Margin of manoeuvre indicators in the workplace during the rehabilitation process: a qualitative analysis, Journal of Occupational Rehabilitation, 19, 194–202. https://doi.org/10.1007/s10926-009-9173-4

Durand M.J., Vézina N. and Richard M.-C. (2016) Concept of margin of manoeuvre in return to work, in Schultz I. and Gatchel R. (eds.) Handbook of return to work, Handbooks in Health, Work, and Disability, vol. 1. Springer, 53–65. https://doi.org/10.1007/978-1-4899-7627-7_3

Edwards L. and Veale M. (2017) Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for, Duke Law & Technology Review, 16, 18–84. https://doi.org/10.2139/ssrn.2972855

Elprama S.A. et al. (2020) Social processes: what determines industrial workers' intention to use exoskeletons?, Human Factors: The Journal of the Human Factors and Ergonomics Society, 62 (3), 337–350. https://doi.org/10.1177/0018720819889534

Elprama S.A., Vanderborght B. and Jacobs A. (2022) An industrial exoskeleton user acceptance framework based on a literature review of empirical studies, Applied Ergonomics, 100, 103615, https://doi.org/10.1016/j.apergo.2021.103615
ENISA (2020) AI cybersecurity challenges: threat landscape for artificial intelligence, European Union Agency for Cybersecurity. https://data.europa.eu/doi/10.2824/238222
ENISA (2022a) ENISA threat landscape 2022: July 2021 to July 2022, European Union Agency for Cybersecurity. https://data.europa.eu/doi/10.2824/764318
ENISA (2022b) Ransomware: publicly reported incidents are only the tip of the iceberg, European Union Agency for Cybersecurity. https://www.enisa.europa.eu/news/ransomware-publicly-reported-incidents-are-only-the-tip-of-the-iceberg
ENISA (2023) ENISA threat landscape 2023: July 2022 to June 2023, European Union Agency for Cybersecurity. https://data.europa.eu/doi/10.2824/782573
ETUC (2020) Social partners agreement on digitalisation (FAD). https://www.etuc.org/en/document/eu-social-partners-agreement-digitalisation
EU-OSHA (2022) OSH Pulse — Occupational safety and health in post-pandemic workplaces. https://osha.europa.eu/en/facts-and-figures/osh-pulse-occupational-safety-and-health-post-pandemic-workplaces
European Commission (2019) Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
European Commission (2020) European enterprise survey on the use of technologies based on artificial intelligence: final report, Publications Office of the European Union. https://data.europa.eu/doi/10.2759/759368
Eurostat (2021) Enterprises using AI technologies, Online data code: Eurostat isoc_eb_ai.
Evans W. (2020) How Amazon hid its safety crisis, Revealnews, 29.09.2020. https://revealnews.org/article/how-amazon-hid-its-safety-crisis/
Eykholt K. et al. (2018) Robust physical-world attacks on deep learning models, arXiv. https://doi.org/10.48550/arXiv.1707.08945
Facchini F., Oleśków-Szłapka J., Ranieri L. and Urbinati A. (2020) A maturity model for logistics 4.0: an empirical analysis and a roadmap for future research, Sustainability, 12 (1), 86. https://doi.org/10.3390/su12010086
Findlay P. et al. (2017) Employer choice and job quality: workplace innovation, work redesign, and employee perceptions of job quality in a complex health-care setting, Work and Occupations, 44 (1), 113–136. https://doi.org/10.1177/0730888416678038
Fleming M. (2001) Safety culture maturity model, Health and Safety Executive.
Flynn-Evans E.E. et al. (2021) Supervision of a self-driving vehicle unmasks latent sleepiness relative to manually controlled driving, Scientific Reports, 11 (1), 18530. https://doi.org/10.1038/s41598-021-92914-5
Forcael E., Ferrari I., Opazo-Vega A. and Pulido-Arcas J.A. (2020) Construction 4.0: a literature review, Sustainability, 12 (22), 9755. https://doi.org/10.3390/su12229755
Forcina A. and Falcone D. (2021) The role of industry 4.0 enabling technologies for safety management: a systematic literature review, Procedia Computer Science, 180, 436–445. https://doi.org/10.1016/j.procs.2021.01.260
Fox S., Aranko O., Heilala J. and Vahala P. (2019) Exoskeletons: comprehensive, comparative and critical analyses of their potential to improve manufacturing performance, Journal of Manufacturing Technology Management, 31 (6), 1261–1280. https://doi.org/10.1108/JMTM-01-2019-0023
Franklin P., Bérastégui P., Cefaliello A. and Musu T. (2023) Social sustainability at work and the essential role of occupational safety and health, in Countouris N., Piasna A.

and Theodoropoulou S. (eds.) Benchmarking Working Europe 2023: Europe in transition - Towards sustainable resilience, ETUI and ETUC, 121–142. https://www.etui.org/publications/benchmarking-working-europe-2023

Fritsch L., Jaber A. and Yazidi A. (2022) An overview of artificial intelligence used in malware, in Zouganeli E., Yazidi A., Mello G. and Lind P. (eds.) Nordic artificial intelligence research and development, NAIS 2022, Communications in Computer and Information Science, vol. 1650, Springer, 41–51. https://doi.org/10.1007/978-3-031-17030-0_4

Galeon D. (2017) An AI completed 360,000 hours of finance work in just seconds, Futurism, 3.08.2017. https://futurism.com/an-ai-completed-360000-hours-of-finance-work-in-just-seconds

Gartner (2023) Gartner glossary. https://www.gartner.com/en/information-technology/glossary/it-ot-integration

Ghozia A., Attiya G., Adly E. and El-Fishawy N. (2020) Intelligence is beyond learning: a context-aware artificial intelligent system for video understanding, Computational Intelligence and Neuroscience. https://doi.org/10.1155/2020/8813089

Giddens L., Gonzales E. and Leidner D. (2019) Unintended consequences of wearable fitness devices in corporate wellness programs, Proceedings of the 2019 on Computers and People Research Conference. https://dl.acm.org/doi/10.1145/3322385.3322416

Guidotti R. et al. (2018) A survey of methods for explaining black box models, ACM Computing Surveys, 51 (5), 93, 1–42. https://doi.org/10.1145/3236009

Habraken M. and Bondarouk T. (2020) Embracing variety in decision-making regarding adoption of industry 4.0, Administrative Sciences, 10 (2), 30. https://doi.org/10.3390/admsci10020030

Hadfield-Menell D., Milli S., Abbeel P., Russell S. and Dragan A. (2017) Inverse reward design, Advances in Neural Information Processing Systems, arXiv. https://arxiv.org/abs/1711.02827

Hagendorff T. (2020) The ethics of AI ethics: an evaluation of guidelines, Minds and Machines, 30, 99–120. https://doi.org/10.1007/s11023-020-09517-8

Ham D.-H. (2020) Safety-II and resilience engineering in a nutshell: an introductory guide to their concepts and methods, Safety and Health at Work, 12 (1), 10–19. https://doi.org/10.1016/j.shaw.2020.11.004

Haslbeck A. and Hoermann H.-J. (2016) Flying the needles: flight deck automation erodes fine-motor flying skills among airline pilot, Human Factors, 58 (4), 533–545. https://doi.org/10.1177/0018720816640394

Hatzius J., Briggs J., Kodnani D. and Pierdomenico G. (2023) The potentially large effects of artificial intelligence on economic growth, Goldman Sachs, 26.03.2023. https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html

He Y. et al. (2017) Risk and adverse events related to lower-limb exoskeletons, 2017 International Symposium on Wearable Robotics and Rehabilitation (WeRob). https://doi.org/10.1109/WEROB.2017.8383850

Hellweg F., Janhofer D. and Hellingrath B. (2023) Towards a maturity model for digital supply chains, Logistics Research, 16 (5). https://doi.org/10.23773/2023_5

Hernandez D. and Strong J. (2018) How computers could make your customer-service calls more human, The Wall Street Journal, 14.06.2018. https://www.wsj.com/articles/call-center-agents-get-a-human-touch-1528984801

Hitlin P. (2016) Research in the crowdsourcing age: a case study, Pew Research Center, 11.07.2016. https://www.pewresearch.org/internet/2016/07/11/research-inthe-crowdsourcing-age-a-case-study/

Hou M., Banbury S. and Burns C. (2018) Intelligent adaptive systems: an interaction-centered design perspective, CRC Press.

Hudson P. (2001) Safety culture - theory and practice, Defense Technical Information Center.

Hudson P. (2007) Implementing a safety culture in a major multi-national, Safety Science, 45 (6), 697–722. https://doi.org/10.1016/j.ssci.2007.04.005

Hügle T. (2023) The wide range of opportunities for large language models such as ChatGPT in rheumatology, RMD Open, 9 (2), e003105. https://doi.org/10.1136/rmdopen-2023-003105

Huysamen K. et al. (2018) Assessment of an active industrial exoskeleton to aid dynamic lifting and lowering manual handling tasks, Applied Ergonomics, 68, 125–131. https://doi.org/10.1016/j.apergo.2017.11.004

IBM (2015) IBM security services 2014 cyber security intelligence index, IBM Global Technology Services.

IDC (2022) Worldwide spending on AI-centric systems will pass $300 billion by 2026, International Data Corporation. https://www.idc.com/resource-center/press-releases

INRS (2022) L'intelligence artificielle au service de la santé et de la sécurité au travail – synthèse : enjeux et perspectives à l'horizon 2035, Institut National de Recherche et de Sécurité. https://www.inrs.fr/media.html?refINRS=PV%2020

Irastorza X. (2019) Third European survey of enterprises on new and emergent risks (ESENER 3), European Agency for Health and Safety at Work. https://osha.europa.eu/en/publications/third-european-survey-enterprises-new-and-emerging-risks-esener-3/view

Irving G., Christiano P. and Amodei D. (2018) AI safety via debate, arXiv. https://arxiv.org/abs/1805.00899

Ivanova I. (2023) A Tesla factory robot reportedly attacked a worker and left them bleeding. This could become a new reality in the increasingly automated workplace, Fortune, 27.12.2023. https://fortune.com/2023/12/27/tesla-factory-robot-worker-attack-injury/

Jaszcz A. and Połap D. (2022) AIMM: Artificial intelligence merged methods for flood DDoS attacks detection, Journal of King Saud University - Computer and Information Sciences, 34 (10), 8090–8101. https://doi.org/10.1016/j.jksuci.2022.07.021

Johnson J.V. and Hall E.M. (1988) Job strain, workplace social support, and cardiovascular disease: a cross-sectional study of a random sample of the Swedish working population, American Journal of Public Health, 78 (10), 1336–1342. https://doi.org/10.2105/ajph.78.10.1336

Karasek R. (1979) Job demands, job decision latitude and mental strain: implications for job redesign, Administrative Science Quarterly, 24 (2), 285–308. https://doi.org/10.2307/2392498

Karim M.M., Li Y. and Qin R. (2022) Towards explainable artificial intelligence (XAI) for early anticipation of traffic accidents. https://doi.org/10.48550/arXiv.2108.00273

Krzywdzinski M., Gerst D. and Butollo F. (2023) Promoting human-centred AI in the workplace. Trade unions and their strategies for regulating the use of AI in Germany, Transfer, 29 (1), 53–70. https://doi.org/10.1177/10242589221142273

Laitinen A. and Sahlgren O. (2021) AI systems and respect for human autonomy, Frontiers in artificial intelligence, 4, 705164. https://doi.org/10.3389/frai.2021.705164

Laney D. (2001) 3D data management: controlling data volume, velocity, and variety. https://perma.cc/GU8B-GF2G

Lasi H., Fettke P., Kemper H.G., Feld T. and Hoffmann M. (2014) Industry 4.0., Business & Information Systems Engineering, 6, 239–242. https://doi.org/10.1007/s12599-014-0334-4

Le Ludec C. and Cornet M. (2023) Enquête : derrière l'IA, les travailleurs précaires des pays du Sud, The Conversation, 26.03.2023. https://theconversation.com/enquete-derriere-lia-les-travailleurs-precaires-des-pays-du-sud-201503

Le Ludec C., Cornet M. and Casilli A. (2023) The problem with annotation. Human labour and outsourcing between France and Madagascar, Big Data & Society, 10 (2), 1–13. https://doi.org/10.1177/20539517231188723

Leal-Ayala D., Castañeda-Navarrete J. and López-Gómez (2019) OK computer? The safety and security dimensions of industry 4.0, Cambridge Industrial Innovation Policy. https://www.ciip.group.cam.ac.uk/reports-and-articles/ok-computer-safety-and-security-dimensions-industr/

Lees M.J. and Johnstone M.C. (2021) Implementing safety features of industry 4.0 without compromising safety culture, IFAC-PapersOnLine, 54 (13), 680–685. https://doi.org/10.1016/j.ifacol.2021.10.530

Lehman J. et al. (2020) The surprising creativity of digital evolution: a collection of anecdotes from the evolutionary computation and artificial life research communities, Artificial Life, 26 (2), 274–306. https://doi.org/10.1162/artl_a_00319

Letona-Ibañez O., Martinez-Rodriguez S., Ortis-Marques N., Carrasco M. and Amillano A. (2021) Job crafting and work engagement: the mediating role of work meaning, International journal of environmental research and public health, 18 (10), 5383. https://doi.org/10.3390/ijerph18105383

Linardatos P., Papastefanopoulos V. and Kotsiantis S. (2021) Explainable AI: a review of machine learning interpretability methods, Entropy, 23 (1), 18. https://doi.org/10.3390/e23010018

Luger T. et al. (2023) Using a back exoskeleton during industrial and functional tasks—effects on muscle activity, posture, performance, usability, and wearer discomfort in a laboratory trial, Human Factors: The Journal of the Human Factors and Ergonomics Society, 65 (1), 5–21. https://doi.org/10.1177/00187208211007267

Maadi M., Akbarzadeh Khorshidi H. and Aickelin U. (2021) A review on human–AI interaction in machine learning and insights for medical applications, International Journal of Environmental Research and Public Health, 18 (4), 2121. https://doi.org/10.3390/ijerph18042121

Massardi S. et al. (2022) Characterisation and evaluation of human–Exoskeleton interaction dynamics: a review, Sensors, 22 (11), 3993. https://doi.org/10.3390/s22113993

McFarland T. and Fischer S. (2019) Considerations for industrial use: a systematic review of the impact of active and passive upper limb exoskeletons on physical exposures, IISE Transactions on Occupational Ergonomics and Human Factors, 7 (3–4), 322–347. https://doi.org/10.1080/24725838.2019.1684399

Methnani L., Aler Tubella A., Dignum V. and Theodorou A. (2021) Let me take over: variable autonomy for meaningful human control, Frontiers in artificial intelligence, 4, 737072. https://doi.org/10.3389/frai.2021.737072

Metz R. (2021) McDonald's and other chains are giving their drive-thrus the Jetsons treatment, CNN Business, 26.02.2021. https://edition.cnn.com/2021/02/26/tech/mcdonalds-drive-thru-artificial-intelligence/index.html

Milmo D. (2024) Company worker in Hong Kong pays out £20m in deepfake video call scam, The Guardian, 5.02.2024. https://www.theguardian.com/world/2024/feb/05/hong-kong-company-deepfake-video-conference-call-scam

Neumann W.P., Winkelhaus S., Grosse E.H. and Glock C.H. (2021) Industry 4.0 and the human factor – A systems framework and analysis methodology for successful development, International Journal of Production Economics, 233. https://doi.org/10.1016/j.ijpe.2020.107992

Niehaus S., Hartwig M., Rosen P. and Wischniewski S. (2022) An occupational safety and health perspective on human in control and AI, Frontiers in Artificial Intelligence, 5, 868382. https://doi.org/10.3389/frai.2022.868382

Nurski L. and Hoffmann M. (2022) The impact of artificial intelligence on the nature and quality of jobs, Working Paper 14/2022, Bruegel. https://www.bruegel.org/working-paper/impact-artificial-intelligence-nature-and-quality-jobs

Ochoa-Urrego R.L. and Peña-Reyes J.I. (2021) Digital maturity models: a systematic literature review, in Schallmo D.R.A. and Tidd J. (eds.) Digitalisation. Management for professionals, Springer, 71–85. https://doi.org/10.1007/978-3-030-69380-0_5

OECD (2021) What happened to jobs at high risk of automation? https://www.oecd.org/future-of-work/reports-and-data/what-happened-to-jobs-at-high-risk-of-automation-2021.pdf

Orr W. and Davis J.L. (2020) Attributions of ethical responsibility by artificial intelligence practitioners, Information, Communication & Society, 23 (5), 719–735. https://doi.org/10.1080/1369118X.2020.1713842

Orseau L. and Armstrong S. (2016) Safely interruptible agents, Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence (UAI'16), AUAI Press, Arlington, Virginia, USA, 557–566. https://dl.acm.org/doi/10.5555/3020948.3021006

Parent-Rocheleau X. and Parker S. K. (2021) Algorithms as work designers: how algorithmic management influences the design of jobs, Human Resource Management Review, 32 (3). https://doi:10.1016/j.hrmr.2021.100838

Parker S.K. and Grote G. (2020) Automation, algorithms, and beyond: why work design matters more than ever in a digital world, Applied Psychology, 71 (4), 1171–1204. https://doi.org/10.1111/apps.12241

Peñaloza G., Wasilkiewicz K., Saurin F.A., Herrera I.A. and Torres Formoso C. (2019) Safety-I and safety-II: opportunities for an integrated approach in the construction industry, 2019: Proceedings: 8th REA Symposium on Resilience Engineering: Scaling up and Speeding up. https://doi.org/10.15626/rea8.18

Perales Gómez A.L. et al. (2021) SafeMan: a unified framework to manage cybersecurity and safety in manufacturing industry, Software Practice and Experience, 51 (3), 607–627. https://doi.org/10.1002/spe.2879

Perez F., Conway N. and Roques O. (2022) The autonomy tussle: AI technology and employee job crafting responses, Relations industrielles / Industrial Relations, 77 (3). https://doi.org/10.7202/1094209ar

Perrigo B. (2023) OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic, Time, 18.01.2023. https://time.com/6247678/openai-chatgpt-kenya-workers/

Pogliani M., Maggi F., Balduzzi M., Quarta D. and Zanero S. (2020) Detecting insecure code patterns in industrial robot programs, Proceedings of the 15th ACM Asia

Conference on Computer and Communications Security (ASIA CCS '20), Association for Computing Machinery, New York, NY, USA, 759–771. https://doi.org/10.1145/3320269.3384735

Poppendieck M. and Poppendieck T. (2006) Implementing lean software development: from concept to cash, Addison-Wesley.

Piasna A. (2024) Job quality and digitalisation, Working Paper 2024.01, ETUI. https://www.etui.org/publications/job-quality-and-digitalisation

Priyadharshini S.K. (2019) Redefining workplace wellness: wearable technology and corporate wellness, Ushus Journal of Business Management, 18 (2), 43–53. https://doi.org/10.12725/ujbm.47.3

Pu H. et al. (2023) Security of industrial robots: vulnerabilities, attacks, and mitigations, IEEE Network 37 (1), 111–117. https://doi.org/10.1109/MNET.116.2200034

PwC (2018) Will robots really steal our jobs? An international analysis of the potential long term impact of automation, PricewaterhouseCoopers. https://www.pwc.co.uk/economic-services/assets/international-impact-of-automation-feb-2018.pdf

Rai S. (2023) JPMorgan is discussing its generative AI projects with regulators, Bloomberg, 9.11.2023. https://www.bloomberg.com/news/articles/2023-11-09/jpmorgan-is-working-with-us-regulators-on-generative-ai-pilot-projects#xj4y7vzkg

Ranchordás S. (2021) Experimental regulations and regulatory sandboxes: law without order?, Law and Method. https://doi.org/10.5553/REM/.000064

Reuters Fact Check (2023) Simulation of AI drone killing its human operator was hypothetical, Air Force Says, Reuters, 8.06.2023. https://www.reuters.com/article/idUSL1N38023R/

Riccò M., Ranzeri S., Vezzosi L., Balzarini F. and Bragazzi N.L. (2021) Wearable exoskeletons on the workplaces: knowledge, attitudes and perspectives of health and safety managers on the implementation of exoskeleton technology in northern Italy, Acta bio-medica: Atenei Parmensis, 92 (6), 2021310–2021312. https://doi.org/10.23750/abm.v92i6.10437

Rodrigues R. (2020) Legal and human rights issues of AI: gaps, challenges and vulnerabilities, Journal of Responsible Technology, 4, 100005. https://doi.org/10.1016/j.jrt.2020.100005

Rogala A. and Cieslak R. (2019) Positive emotions at work and job crafting: results from two prospective studies, Frontiers in Psychology, 10. https://doi.org/10.3389/fpsyg.2019.02786

Rudin C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence, 1 (5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Ruschemeier H. (2023) AI as a challenge for legal regulation – the scope of application of the artificial intelligence act proposal, ERA Forum, 23 (3), 361–376. https://doi.org/10.1007/s12027-022-00725-6

Saghezchi F.B., Mantas G., Violas M.A., de Oliveira Duarte A.M. and Rodriguez J. (2022) Machine learning for DDoS attack detection in industry 4.0 CPPSs, Electronics, 11 (4), 602. https://doi.org/10.3390/electronics11040602

Sanger D.E. (2012) Obama order sped up wave of cyberattacks against Iran, The New York Times, 1.06.2012. https://www.nytimes.com/2012/06/01/world/middleeast/obama-ordered-wave-of-cyberattacks-against-iran.html

Scheuermann C., Strobel M., Bruegge B. and Verclas S. (2016) Increasing the support to humans in factory environments using a smart glove: an evaluation, IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted

Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress, Toulouse, 18-21.07.2016, IEEE, 847–854. https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0134

Schmidt F.A. (2017) Digital labour markets in the platform economy: mapping the political challenges of crowd work and gig work, Friedrich-Ebert-Stiftung. https://library.fes.de/pdf-files/wiso/13164.pdf

Schultz I.Z. and Gatchel E.J. (2015) Handbook of return to work: from research to practice, Springer.

Shafer T. (2017) The 42 V's of big data and data science, Elder Research. https://www.elderresearch.com/blog/42-v-of-big-data

Shahid A., Zaidi M. and Azizan R. (2019) A road map to generative safety culture: an integrated conceptual model, IOP Conference Series: Materials Science and Engineering, 702, 012051. https://doi.org/10.1088/1757-899X/702/1/012051

Shandler R., Gross M.L. and Canetti D. (2023) Cyberattacks, psychological distress, and military escalation: an internal meta-analysis, Journal of Global Security Studies, 8 (1). https://doi.org/10.1093/jogss/ogac042

Shang W. (2022) The effects of job crafting on job performance among ideological and political education teachers: the mediating role of work meaning and work engagement, Sustainability, 14 (14), 8820. https://doi.org/10.3390/su14148820

Shwab C. (2016) The fourth industrial revolution, World Economic Forum. https://www.weforum.org/about/the-fourth-industrial-revolution-by-klaus-schwab

Sophos (2021) Sophos annual ransomware survey: the state of ransomware 2021. https://secure2.sophos.com/en-us/medialibrary/pdfs/whitepaper/sophos-state-of-ransomware-2021-wp.pdf

Stanton N.A. (2019) Thematic issue: driving automation and autonomy, Theoretical Issues in Ergonomics Science, 20 (3), 215–222. https://doi.org/10.1080/1463922X.2018.1541112

Stone N. (2023) Bitsight identifies nearly 100,000 exposed industrial control systems, Bitsight, 2.10.2023. https://www.bitsight.com/blog/bitsight-identifies-nearly-100000-exposed-industrial-control-systems

Stupp C. (2019) Fraudsters used AI to mimic CEO's voice in unusual cybercrime case, The Wall Street Journal, 30.08.2019. https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

Tamascelli N., Solini R., Paltrinieri N. and Cozzani V. (2022) Learning from major accidents: a machine learning approach, Computers & Chemical Engineering, 162, 107786. https://doi.org/10.1016/j.compchemeng.2022.107786

TKI DINALOG (2020) Mens en robot in het magazijn. Een onderzoek naar de inzet van robotica in het magazijn en de implicaties voor medewerkers, Dutch Institute for Advanced Logistics. https://www.dinalog.nl/project/mens-en-robot-magazijn/

Vale D., El-Sharif A. and Ali M. (2022) Explainable artificial intelligence (XAI) post-hoc explainability methods: risks and limitations in non-discrimination law, AI Ethics, 2 (1), 815–826. https://doi.org/10.1007/s43681-022-00142-y

Wachter S., Mittelstadt B. and Russell C. (2018) Counterfactual explanations without opening the black box: automated decisions and the GDPR. https://doi.org/10.48550/arXiv.1711.00399

Walter O.M., Paladini E.P., Henning E. and Konrath A.C. (2020) Industry 4.0 maturity models: review and classification as a support for Industry 4.0 implementation. https://doi.org/10.14488/IJCIEOM2020_FULL_0001_37251

Wang B. and Wang Y. (2021) Big data in safety management: an overview, Safety Science, 143, 105414. https://doi.org/10.1016/j.ssci.2021.105414

Wang N. and Wang K. (2022) Internet financial risk management in the context of big data and artificial intelligence, Mathematical Problems in Engineering, 2022. https://doi.org/10.1155/2022/6219489

Westrum R. (2004) A typology of organisational cultures, Quality & safety in health care, 13 (Suppl. 2), 22–27. https://doi.org/10.1136/qshc.2003.009522

Yassine M. (2021) IT/OT convergence and cybersecurity, Computer Fraud & Security, 2021 (12), 13–16. https://doi.org/10.1016/S1361-3723(21)00129-9

Zervoudi E.K. (2020) Fourth industrial revolution: opportunities, challenges, and proposed policies, in Grau A. and Wang Z. (eds.) Industrial robotics: new paradigms, IntechOpen, 1–25. https://doi.org/10.5772/intechopen.90412

Zwysen W. et al. (2024) Labour market and social developments in the EU: the quest for strong jobs recovery, in Piasna A. and Theodoropoulou S. (eds.) Benchmarking Working Europe 2024: the ongoing quest for Social Europe, ETUI and ETUC, 51–86. https://www.etui.org/publications/benchmarking-working-europe-2024

All links were checked on 21.05.2024.

**European**
**Trade Union Institute**
Bd du Jardin Botanique, 20
1000 Brussels
Belgium
etui@etui.org
www.etui.org

D/2024/10.574/16
ISBN: 978-2-87452-713-5 (print version)
ISBN: 978-2-87452-714-2 (electronic version)


etui.