Healthy Workplaces Campaign 2023-2025
Safe and healthy work in the digital age
Good Practice Exchange Event
Brussels, 20-21 May 2025

AI Regulations & OSH
The Challenge of Balance

Roberto Sammarchi, Attorney at Law
AIAS (IT) representative in ENSHPO

Abstract

This paper examines the delicate balance between leveraging Artificial Intelligence (AI) to enhance Occupational Safety and Health (OSH) and navigating the complex regulatory landscape, particularly the European Union's AI Act. It explores AI's transformative potential in preventing workplace accidents and promoting worker well-being through predictive analytics, real-time monitoring, and automated safety systems. The document highlights the need to address ethical considerations, data privacy concerns, and the potential for algorithmic bias. It analyzes the specifics of the EU AI Act, its risk-based approach, the definition of "AI system," and the implications of the "high-risk" classification for OSH applications. It explores pathways for OSH AI systems to potentially avoid high-risk classification and advocates for a nuanced regulatory approach that fosters innovation while safeguarding fundamental rights.

1. Introduction: The Transformative Potential of AI in Occupational Safety and Health (OSH)

The integration of Artificial Intelligence (AI) into the fabric of modern workplaces heralds a paradigm shift in Occupational Safety and Health (OSH). AI technologies offer a significant opportunity to move beyond traditional reactive safety measures towards proactive, and even predictive, strategies for safeguarding worker well-being. This transformative potential, however, is not without its complexities, particularly as regulatory frameworks endeavor to keep pace with rapid technological advancement. The central challenge lies in striking a delicate balance: harnessing the profound capabilities of AI to enhance workplace safety while simultaneously upholding fundamental rights and ethical principles through robust governance.

The Promise of AI in Enhancing Workplace Safety

AI's capacity to analyze vast datasets, identify subtle patterns, and operate with varying degrees of autonomy presents a suite of tools for revolutionizing OSH. Predictive analytics, for instance, can sift through historical incident data, equipment sensor readings, and even environmental factors to forecast potential hazards, such as impending machinery failures or conditions conducive to worker injury. Real-time monitoring systems, powered by AI, can continuously assess work environments for emerging risks, like the presence of toxic gases, or monitor worker physiological states to detect signs of fatigue or overexertion, thereby enabling timely interventions.

Furthermore, AI can drive automated risk prevention systems. Examples include sophisticated collision avoidance algorithms in robotic systems operating alongside human workers, or intelligent emergency shutdown protocols that activate when dangerous conditions are detected. Beyond direct intervention, AI can personalize OSH training, adapting educational content to individual worker needs and learning styles, and provide dynamic guidance in complex or hazardous tasks. The overarching ambition is clear: to significantly reduce, and ultimately eliminate, workplace accidents, injuries, and illnesses by creating smarter, safer working environments. This evolution aims to shift OSH management from a reactive stance, addressing incidents after they occur, to a predictive one, preventing them from happening in the first place.

Setting the Stage: The Dual Imperative of Innovation and Regulation

The enthusiasm for adopting AI in OSH is palpable across industries, driven by the promise of enhanced safety outcomes and operational efficiencies. Yet, this drive for innovation must be tempered by a thorough recognition of the ethical considerations and potential risks associated with AI deployment. Concerns surrounding data privacy, particularly with systems that monitor workers, the potential for algorithmic bias to perpetuate or even exacerbate existing inequalities, and the inherent opacity of some advanced AI models (often termed the "black box" problem) necessitate a careful and considered regulatory approach.

This introduces the core theme of this analysis: the critical need for balance. Regulations that are overly prescriptive or burdensome risk stifling innovation, potentially slowing the adoption of AI systems that could save lives and prevent harm. Conversely, insufficient or poorly conceived governance could lead to the misuse of AI, ethical breaches, or the erosion of worker trust and fundamental rights. The challenge, therefore, is to cultivate a regulatory ecosystem that fosters the responsible development and deployment of AI in OSH—one that encourages technological advancement while ensuring these powerful tools are used ethically and for the genuine benefit of worker safety and health.

A particularly nuanced aspect of regulating AI in OSH emerges when considering that many

such AI systems are inherently designed to specifically prevent harm. Regulatory frameworks, including the European Union's AI Act, tend to approach AI primarily through the lens of the risks *posed by* the technology itself. This perspective can create a situation where tools specifically developed to mitigate existing workplace hazards are subjected to stringent scrutiny and compliance burdens that may, paradoxically, hinder their widespread adoption. If the regulatory hurdles for implementing a safety-enhancing AI are perceived as too high or too costly, organizations might opt to continue with less effective, traditional safety measures. This potential outcome underscores a fundamental tension: a risk-centric regulatory approach, if not carefully calibrated, could inadvertently lead to suboptimal safety outcomes by discouraging the very innovations designed to improve them. This is particularly relevant given the observation that the AI Act tends to frame AI as a risk to be managed, rather than as a set of tools and technologies that can be proactively used to define, detect, evaluate, manage, control, reduce, and even eliminate risks in the workplace.

2. The EU AI Act: A New Regulatory Landscape for AI

The European Union has taken a pioneering step in addressing the challenges and opportunities presented by artificial intelligence through the enactment of Regulation (EU) 2024/1689, commonly known as the EU AI Act. This legislation represents the world's first comprehensive, horizontal legal framework for AI, aiming to establish harmonized rules for the development, marketing, and use of AI systems within the Union. Its overarching goals are to promote the uptake of human-centric and trustworthy AI while ensuring a high level of protection for health, safety, and fundamental rights, including those enshrined in the EU Charter.

Overview of the EU AI Act

The AI Act adopts a risk-based approach, a methodology central to its architecture, categorizing AI systems into distinct tiers based on their potential to cause harm. This tiered system includes:

- Unacceptable risk: AI systems posing a clear threat to the safety, livelihoods, and rights of people are deemed to contravene EU values and are therefore prohibited. Examples include social scoring by public authorities, real-time remote biometric identification in publicly accessible spaces for law enforcement (with limited exceptions), and AI manipulating human behavior to circumvent free will.
- High-risk: AI systems that have a high potential to adversely affect safety or fundamental rights are permitted but subject to a stringent set of mandatory requirements and conformity assessments before they can be placed on the market or put into service. Many AI applications relevant to OSH may fall into this category.
- Limited risk: AI systems such as chatbots or deepfakes are subject to specific transparency obligations, requiring users to be informed that they are interacting with an AI or that content is AI-generated.

- Minimal risk: The Act allows the free use of minimal-risk AI systems, such as AI-enabled video games or spam filters, which constitute the majority of AI systems currently in use. These systems are largely exempt from the Act's obligations, though voluntary codes of conduct are encouraged.

The AI Act's scope is notably broad, extending its obligations not only to providers and deployers of AI systems established within the EU but also to those located outside the Union if their AI systems are placed on the EU market, if the use of such systems affects persons located in the EU, or if the output produced by the system is used in the EU. This extraterritorial reach underscores the EU's ambition to set global standards for AI governance.

The Critical Definition of an "AI System" under Article 3(1)

Central to the AI Act's applicability is its definition of an "AI system." Article 3(1) defines an AI system as:

"A machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments".

The European Commission has issued non-binding guidelines to clarify this definition, breaking it down into seven key cumulative elements.

1. Machine-based system: This element signifies that AI systems operate on machines, encompassing both hardware (e.g., processors, memory) and software (e.g., code, algorithms). The approach is technology-neutral, covering traditional computing architectures as well as emerging technologies like quantum computing, provided they offer computational capacity.
2. Varying levels of autonomy: AI systems must be designed to operate with "some degree of independence of actions from human involvement and of capabilities to operate without human intervention". This excludes systems that are fully manually controlled. The degree of autonomy is a crucial determinant, particularly in OSH contexts where human oversight is often a fundamental safety principle.
3. May exhibit adaptiveness after deployment: This refers to self-learning capabilities that allow an AI system's behavior to change while in use, potentially leading to different outputs for the same inputs. Crucially, the guidelines clarify that adaptiveness is *not* a mandatory condition for a system to be classified as AI. This implies that even static, rule-based yet complex systems could fall under the definition if other criteria are met.
4. Explicit or implicit objectives: An AI system's objectives can be explicit (clearly stated goals encoded by developers, e.g., optimizing a cost function) or implicit (goals deduced from the system's behavior, training data, or interaction with its environment). This

distinction is vital for understanding the "intent" versus the "function" of an OSH AI system.

5. Infers...how to generate outputs: This capability is considered a key differentiator between AI systems and simpler traditional software. The system must be able to determine how to produce outputs based on the input it receives. This element will be explored in greater detail in the subsequent section.
6. Outputs (predictions, content, recommendations, decisions): These are the types of outputs an AI system can generate, reflecting the nature of its actions.
7. Can influence physical or virtual environments: The AI system's outputs must have the potential to affect tangible, physical objects or digital spaces, data flows, and software ecosystems. This is highly pertinent for OSH AI, which frequently interacts directly with and influences physical workplace environments.

The European Commission's guidelines emphasize that determining whether a system qualifies as AI requires a nuanced, case-by-case analysis of its specific architecture and functionality, rather than a rigid, mechanical checklist approach. Some elements may only be present during the pre-deployment (building) phase or the post-deployment (use) phase.

The breadth of this definition, while intended to be future-proof and technologically neutral, carries implications for existing technologies. While the Commission's guidelines aim to exclude "simpler traditional software systems or programming approaches", the cumulative nature of the seven elements and the fact that adaptiveness is optional mean that some sophisticated automated systems currently used in OSH, perhaps not previously categorized as "AI," might now fall under the Act's purview. For instance, an advanced industrial control system employing embedded logic to make autonomous safety adjustments based on multiple sensor inputs could potentially meet the criteria: it is machine-based, operates with a degree of autonomy, has implicit safety objectives, and infers how to generate decisions (e.g., slowing down a process) that influence the physical environment. This potential expanding scope could impose unforeseen regulatory obligations and costs on established OSH technologies, affecting industries that have long relied on them for safety and operational integrity.

Further complexity arises from the distinction drawn in the EC guidelines between an AI system's internal "objectives" (the goals of the tasks it is designed to perform) and its "intended purpose" (the external use for which the provider makes the system available). In the OSH domain, the *intended purpose* of an AI system might be unequivocally safety-enhancing, such as "to prevent a worker from entering an active robotic cell." However, its internal *objectives*, for example, "to optimize the robot's movement path while identifying and avoiding human-shaped obstacles based on learned patterns," could be more complex. If these internal objectives are poorly defined, or if the AI system learns unintended correlations or behaviors (especially in adaptive systems), it could lead to unforeseen risks, even if the fundamental objective remains

benign. This distinction is critical for conducting thorough risk assessments under the AI Act, as a system with a laudable safety purpose could still be classified as high-risk if its underlying operational objectives and inferential processes are opaque, unreliable, or prone to generating harmful outputs.

3. Decoding the AI Act's Language: "Infer," "Deduce," or "Derive"?

The precise terminology employed in legal definitions is paramount, as it delineates the boundaries of a regulation's scope. Within the EU AI Act's definition of an "AI system," the choice of the verb describing how a system generates outputs from inputs – specifically the English term "infers" – and its translations into other official EU languages, presents a critical point of analysis. The nuances between "infer," "deduce" (used in Italian and French translations), and "derive" (ableiten—used in the German translation) can have significant implications for determining which machine-based systems are subject to the Act's stringent requirements, particularly in the OSH domain.

The Significance of Terminology in Defining the Act's Scope

The above linguistic variance is a "key point." If the operative verb is interpreted narrowly, certain AI technologies might fall outside the Act's purview, while a broader interpretation would capture a wider array of systems. This directly affects AI developers, providers, and deployers in the OSH sector, influencing design choices, compliance strategies, and ultimately, the range of AI tools available for enhancing workplace safety.

Comparative Analysis of "Infer" (English) vs. "Deduce" (Italian/French) vs. "Derive" (German)

A comparative examination of these terms, drawing from general logic, legal usage, and the AI Act's context, reveals important distinctions:

- "Infer":
    - General Logic: To infer is to arrive at a conclusion by reasoning from evidence or premises. Inference can be deductive (reasoning from general principles to specific conclusions) or inductive (reasoning from specific observations to broader generalizations). It often implies arriving at a conclusion that is probable or plausible, rather than absolutely certain, especially in inductive or abductive reasoning. "Inference" is generally considered a broader term than "deduction".
    - Legal Context (US): In US legal parlance, an inference is a logical conclusion drawn from established facts or evidence during a judicial proceeding. The process of drawing such an inference is often described as "deduction" or "deductive reasoning" and can form persuasive circumstantial evidence.
    - AI Act (English Text): The AI Act, in its English version, uses "infers." The European

Commission's guidelines on the AI system definition clarify that "inferencing" capabilities distinguish AI systems from traditional rule-based software. This involves the system's capacity to "infer how to generate output based on input data," encompassing both the use phase (generating predictions, decisions) and the building phase (deriving models or algorithms using AI techniques like machine learning). The guidelines state this capacity allows AI to operate without being bound *solely* by human-defined rules. This interpretation leans towards a broad understanding of "infer," accommodating systems that learn from data and make probabilistic judgments.

- "Deduce":
  - General Logic: To deduce is to use logic or reason to form a conclusion from known facts or general principles, typically moving from the general to the specific. Deduction often implies drawing a particular inference from a generalization, aiming for a conclusion that is logically certain if the premises are true. It is generally seen as a more constrained form of reasoning than general inference.
  - Legal Context: While sometimes used interchangeably with "infer," "deduce" can carry the connotation of a more direct and logically necessary step from premises to conclusion.
  - AI Act (Italian "deduce," French "déduit"): The use of "deduce" in these language versions could suggest a narrower interpretation of the AI Act's scope. If interpreted strictly, it might imply that only AI systems operating on explicit logical rules and deriving certain conclusions would be covered, potentially excluding many contemporary AI systems based on statistical learning or neural networks which "infer" patterns rather than "deduce" from first principles.
- "Derive":
  - General Logic: "Derivation" is often associated with formal proof systems in logic, where a conclusion is reached by following a specific sequence of steps or rules of inference from a set of axioms or premises.
  - Legal Context: In legal contexts, "derive" frequently means to obtain, create, develop, or generate something from another source. For example, a "derivative work" in copyright law is a work based on or transformed from a pre-existing work. In data-related agreements, "derive" can mean to deduce or infer information (e.g., personally identifiable information from other data) or to generate "derived data" through processes like manipulation, calculation, or analysis of original data.
  - AI Act (German "ableitet"): The German term "ableitet" can translate to "derives," "deduces," or "infers." If interpreted in the sense of logical derivation or deduction, it might align with the narrower interpretation. However, if understood in the broader sense of generating or creating outputs from inputs, similar to how "derive" is used in

the context of derived data, it could align more closely with the encompassing nature of the English "infers" or even emphasize the generative aspect of some AI systems.

Implications of These Distinctions for AI Systems in OSH

These linguistic variations are not merely academic; they have tangible implications for AI systems deployed in OSH:

- If the narrower concept of "deduce" were universally applied, many advanced OSH AI systems that rely on machine learning to identify complex safety patterns from sensor data, or probabilistic models to predict accident risks, might fall outside the Act's scope. Such systems often "infer" correlations and likelihoods rather than "deducing" outcomes with logical certainty from predefined general principles.
- If "derive," in the sense of formal logical derivation, were the standard, it would similarly exclude many data-driven AI systems. However, if "derive" is interpreted as "to generate or create" (as in derived data), it would aptly cover generative AI but might be less intuitive for AI systems making predictive inferences for OSH risk assessment.
- The English term "infers," as clarified by the EC guidelines to include machine learning and the ability to operate beyond solely human-defined rules, appears to be the most encompassing. This interpretation is likely to capture a wider range of AI techniques currently used or envisioned for OSH applications, from sophisticated pattern recognition in real-time monitoring systems to complex decision support tools for risk management.

The divergence in these key terms across official language versions of the AI Act introduces a degree of legal uncertainty. While the EC guidelines attempt to provide a harmonized understanding by focusing on the system's capacity to go beyond explicitly programmed rules, the ultimate interpretation of these terms by national courts and the Court of Justice of the European Union (CJEU) will be crucial. This linguistic variance could potentially lead to differing applications of the Act across Member States, creating an uneven regulatory landscape. For developers and deployers of OSH AI systems, who often operate across multiple EU countries, such inconsistency could pose significant compliance challenges and may even influence where they choose to develop or market their technologies. This scenario underscores the importance of striving for a consistent interpretation that aligns with the Act's objective of creating a harmonized framework.

Addressing Ambiguities: "Virtual / Physical Environment" and "Can Influence"

Beyond the verb choice, the AI Act's definition contains other phrases that require careful interpretation for OSH applications:

- "Virtual or Physical Environment": The Act itself does not provide a formal definition for these terms. However, the EC guidelines offer clarification, stating that "physical

environments" refer to tangible, physical objects (e.g., a robot arm, machinery, a worker's physical workspace) and "virtual environments" encompass digital spaces, data flows, and software ecosystems. This definition is broad. For OSH, the "physical environment" is of paramount importance, as AI systems often interact with machinery, control industrial processes, or monitor the physical conditions of the workplace. The "virtual environment" is also relevant, potentially including AI-driven OSH training simulations, safety data dashboards, or the software systems that control physical safety mechanisms.

- "Can Influence": The phrase "can influence physical or virtual environments" appears quite generical. The critical question is what type or degree of influence is considered relevant for an AI system to fall under the Act. Does it require a direct, indirect, or substantial modification? The EC guidelines suggest that the AI system's output (predictions, content, recommendations, or decisions) must merely be capable of influencing these environments. It does not seem to necessitate a substantial modification or a direct causal link in all instances; the potential for influence appears sufficient. The guidelines also note that systems that only process or visualize data without influencing any outcome fall outside the definition.

  For OSH AI, this threshold is significant. An AI system that directly controls a safety barrier or an emergency stop mechanism clearly "influences" the physical environment. However, an AI-powered alert system that notifies a human supervisor of a potential hazard also "can influence" the environment by prompting human action, which then leads to a physical change or intervention. Similarly, an AI tool providing risk scores to an OSH manager "can influence" that manager's decisions regarding safety protocols, which in turn affect the workplace environment. If the threshold for "influence" is interpreted very broadly, many informational or advisory OSH AI systems could be captured by the Act, even if their impact is mediated by human judgment and their direct causal effect on the physical environment is indirect. This ambiguity could lead to uncertainty in classifying "soft" OSH AI tools, such as those used for decision support, risk analytics, or compliance checking, potentially subjecting them to stricter regulatory scrutiny than might be proportionate to their actual risk profile.

To provide a clearer understanding of these critical terms, the following table compares their definitions and implications:

Table 1: Comparative Definitions of Infer, Deduce, Derive and their Implications for the AI Act

Term	Definition in Logic/Reasoning	Definition/Use in Legal Context	AI Act Language Version & Implied Meaning (based on EC Guidelines)
Infer	Arriving at a conclusion by reasoning from evidence; can be deductive, inductive, or abductive. Often implies probabilistic conclusions. Broader than deduce.	A logical conclusion from established facts; process often termed "deduction."	English: "infers". Broad interpretation. Encompasses systems that learn from data and operate beyond solely human-defined rules, including ML.
Deduce	To use logic or reason to form a conclusion from known facts or general principles; typically general to specific, aiming for logical certainty if premises are true. More constrained than infer.	Often used for direct logical steps from premises.	Italian: "deducts" / French: "déduit". Potentially narrower scope, focusing on systems operating on more explicit logical steps.
Derive	Often used for formal proof systems, following specific steps.	To create, develop, or obtain something from another source (e.g., derivative works, derived data). Can mean to deduce or infer information from data.	German: "ableitet". Ambiguous; could mean logical derivation (narrower) or generation / creation of outputs (broader, aligning with "infer" or generative aspect).

This table highlights the potential for varied interpretations based on linguistic nuances, underscoring the need for consistent application of the AI Act's definitions to ensure legal certainty, particularly for technologies as critical as those used in OSH.

4. AI in OSH: Navigating the "High-Risk" Classification

The EU AI Act's framework, while aiming for a comprehensive governance of AI, does not

explicitly carve out a dedicated space for Occupational Safety and Health AI systems that frames them positively as tools for risk reduction. Instead, AI in the occupational domain is primarily viewed through the prism of potential risks it might introduce, often leading such systems to be categorized under the Act's "high-risk" classification. This perspective has significant ramifications for the development, deployment, and regulatory burden associated with AI solutions intended to make workplaces safer.

The AI Act's Implicit Treatment of OSH AI Systems

The AI Act's fundamental approach is risk-based. Consequently, AI systems relevant to OSH are typically captured under broad high-risk categories rather than being addressed through a lens that acknowledges their primary purpose of enhancing safety and preventing harm. This can foster a default perception that any AI system operating within the OSH domain is inherently problematic or carries an elevated level of danger, overshadowing its potential benefits.

The Default Assumption: OSH AI Systems as "High-Risk" (Article 6, Annex III)

Article 6 of the AI Act lays down the classification rules for high-risk AI systems. An AI system is generally designated as high-risk if it meets one of two main criteria:

1. It is intended to be used as a safety component of a product, or is itself a product, covered by existing EU harmonisation legislation listed in Annex I of the Act (e.g., machinery, medical devices), and that product is required to undergo a third-party conformity assessment for health and safety risks. Many AI systems embedded in industrial equipment or safety devices in OSH contexts could fall under this provision.
2. It falls within one of the specific areas and use cases listed in Annex III of the Act.

Several categories in Annex III are directly relevant to OSH and employment settings, making it likely that AI systems used in these contexts will be presumptively classified as high-risk:

- Employment, workers management and access to self-employment (Annex III, point 4): This is a broad category covering AI systems intended for recruitment, selection, making decisions on promotion or termination, task allocation, and monitoring or evaluating the performance and behavior of workers. AI tools used for OSH purposes within these worker management functions (e.g., AI monitoring for unsafe behaviors, AI allocating tasks based on fatigue levels) would likely be captured here.
- Management and operation of critical infrastructure (Annex III, point 2): AI systems used as safety components in the management and operation of critical digital infrastructure, road traffic, and the supply of water, gas, heating, and electricity. OSH AI systems ensuring the safety of these infrastructures could be included.
- Education and vocational training (Annex III, point 3): AI systems intended to determine access or assign persons to educational and vocational training institutions or to evaluate

learning outcomes, including for OSH certifications.

Consequences of High-Risk Classification

Designation as a high-risk AI system under the AI Act triggers a comprehensive and demanding set of obligations that apply throughout the system's life cycle. These requirements include, but are not limited to:

- Risk management systems: Establishing and maintaining a continuous iterative risk management process.
- Data and data governance: Ensuring high quality of training, validation, and testing datasets, particularly concerning relevance, representativeness, and freedom from biases.
- Technical documentation and record-keeping: Creating and maintaining extensive technical documentation and logs of the AI system's functioning to ensure traceability and allow for assessment of compliance.
- Transparency and provision of information to users: Providing deployers with clear and adequate information about the system's capabilities, limitations, and intended purpose, including instructions for use.
- Human oversight: Designing systems to allow for effective human oversight, with measures appropriate to the risks.
- Accuracy, robustness, and cybersecurity: Ensuring an appropriate level of accuracy, resilience against errors or inconsistencies, and robustness against attempts to alter its use or behavior by malicious third parties.
- Conformity assessment: Undergoing conformity assessment procedures, which may involve third-party notified bodies for certain systems, before being placed on the market or put into service.
- Registration: Registering the high-risk AI system in an EU-wide public database.
- Post-market monitoring: Implementing a post-market monitoring system to collect and analyze data about the performance of the AI system in real-world use and reporting serious incidents and malfunctions.

These obligations translate into significant development and compliance costs, complex administrative fulfillments, heightened legal risks, and the potential for substantial financial penalties for non-compliance (up to EUR 35 million or 7% of worldwide annual turnover for certain violations). Such a regulatory burden can act as a substantial barrier to entry and deployment, particularly for small and medium-sized enterprises (SMEs) or research institutions looking to innovate in the OSH AI space. This "chilling effect" on innovation is a serious concern, as it might deter the development of novel AI solutions that could offer significant improvements in workplace safety, simply because the upfront investment and ongoing compliance demands associated with high-risk classification are too prohibitive. This effect could be especially pronounced for AI systems that are genuinely groundbreaking and do not fit

neatly into predefined exceptions or established technological pathways.

The Derogation Provision for Annex III Systems (Article 6(3))

Recognizing that not all AI systems falling under the broad categories of Annex III will necessarily pose a high level of risk in every instance, the AI Act includes a derogation mechanism in Article 6(3). This provision allows an AI system listed in Annex III to *not* be considered high-risk if its provider can demonstrate that the system "does not pose a significant risk of harm to the health, safety or fundamental rights of natural persons, including by not materially influencing the outcome of decision making."

This derogation applies if the AI system fulfills one of the following conditions:

- It is intended to perform a narrow procedural task.
- It is intended to improve the result of a previously completed human activity.
- It is intended to detect decision-making patterns or deviations from prior decision-making patterns and is not meant to replace or influence the previously completed human assessment, without proper human review.
- It is intended to perform a preparatory task to an assessment relevant for the purposes of the use cases listed in Annex III.

However, an important caveat is that an AI system referred to in Annex III will *always* be considered high-risk if it performs profiling of natural persons.

Providers who believe their Annex III AI system qualifies for this derogation must document their assessment *before* the system is placed on the market or put into service and make this documentation available to national competent authorities upon request. They are also subject to registration obligations in the EU database. The European Commission is tasked with providing guidelines, including a comprehensive list of practical examples of high-risk and non-high-risk use cases of AI systems, by February 2026 to aid in the practical implementation of this article.

The ambiguity inherent in terms like "significant risk of harm" and "materially influencing the outcome of decision making" presents a challenge. While these phrases offer a degree of flexibility for providers to argue that their OSH AI system is not high-risk if its impact is genuinely low or indirect, this lack of precise definition also creates uncertainty. Providers may find it difficult to confidently assess whether their system meets these criteria, potentially leading to overly cautious self-classification as high-risk to avoid penalties, or, conversely, to attempts to downplay risks that national authorities might later deem significant. This uncertainty undermines legal predictability for OSH AI development and deployment. The burden of proof resting on the provider to meticulously document their non-high-risk assessment adds to this complexity.

Furthermore, the interaction between the AI Act and existing EU product safety legislation, such as the Machinery Directive, adds another layer of complexity. Article 6(1) and Annex I of the AI Act explicitly link high-risk classification to products covered by such legislation. Many AI systems in OSH will be integrated into machinery or function as safety components. These systems must then comply with both the essential health and safety requirements of the relevant product safety directive *and* the high-risk obligations of the AI Act. This dual compliance pathway can be intricate and costly, potentially leading to conflicting requirements or duplicative assessment efforts if the interplay between these regulatory frameworks is not managed carefully through clear guidance and harmonized standards.

The following table provides a structured overview of the AI Act's risk categories and their potential implications for OSH AI systems:

Table 2: EU AI Act Risk Categories & OSH Implications

Risk Category	**General AI Act Definition / Criteria**	**Examples of OSH AI Systems in this Category (with rationale)**	**Key Obligations/Consequences for OSH AI**	**Potential for "Non High-Risk Classification" (for High Risk Category)**
Unacceptable Risk (Prohibited)	Poses a clear threat to safety, livelihoods, fundamental rights; contravenes EU values.	- AI for social scoring of workers leading to detrimental OSH outcomes (e.g., denial of safety equipment based on score). - Emotion recognition in the workplace *unless* for strictly defined medical / safety reasons (e.g., fatigue detection for pilots, not general stress monitoring). -AI deploying subliminal techniques to make workers	Banned from the EU market. Severe penalties for violations.	N/A

		bypass safety protocols.		
High Risk	Potential adverse impact on safety or fundamental rights. Listed in Annex III or safety component under Annex I legislation requiring 3rd party conformity assessment.	- AI for monitoring worker compliance with safety procedures and making decisions on disciplinary actions (Annex III.4). - AI controlling safety-critical machinery (e.g., robotic arms in collaborative workspaces) (Annex I product or safety component). - AI used for allocating safety-critical tasks based on worker profiles (Annex III.4). - AI for evaluating OSH training effectiveness if it impacts employment status (Annex III.3 & III.4).	Strict obligations: risk management, data governance, technical documentation, transparency, human oversight, accuracy, robustness, cybersecurity, conformity assessment, registration, post-market monitoring. Significant costs and legal risks.	Yes, under Article 6(3) if the system does not pose significant risk of harm AND does not materially influence decision-making AND meets specific conditions (e.g., narrow procedural task, improves human activity, preparatory task). Does not apply if profiling natural persons.
Limited Risk	Systems interacting with humans (e.g., chatbots) or generating content (e.g., deepfakes).	- AI chatbots providing basic OSH information (users must be aware they are interacting with AI). - AI generating OSH training	Transparency obligations: users must be informed they are interacting with AI or that content is AI-generated.	N/A

		videos using synthetic characters (must be disclosed as AI-generated if it appears authentic).		
Minimal Risk	All other AI systems posing minimal or no risk.	- AI for predictive maintenance of non-safety-critical OSH equipment (e.g., scheduling checks for fire extinguishers based on usage data, where failure doesn't pose immediate high risk). - Simple AI tools for organizing OSH documentation without decision-making capabilities. - AI-powered spam filters for OSH communication channels.	Largely exempt from AI Act obligations, though voluntary codes of conduct and general principles (human oversight, fairness) encouraged.	N/A

This table illustrates the complex landscape OSH AI developers and deployers must navigate, emphasizing the critical importance of understanding the specific conditions under which an AI system might be classified and the pathways available to potentially avoid the most burdensome high-risk obligations.

5. Seeking Balance: Identifying Non High-Risk AI Pathways in OSH

The primary challenge for fostering the beneficial use of AI in Occupational Safety and Health lies in promoting its implementation and effective use without imposing undue regulatory burdens that could stifle innovation or make valuable safety tools inaccessible. A key aspect of

this endeavor is to understand and clearly define which AI systems intended for OSH applications could legitimately be considered outside the demanding "high-risk" classification under the EU AI Act. This requires a careful examination of the Act's provisions, particularly the derogations and nuanced interpretations that might allow for a more proportionate regulatory approach.

The Core Challenge: Promoting Beneficial AI in OSH without Undue Regulatory Burden

If the objective is to promote the implementation and effective use of risk management and risk reduction AI systems in the OSH domain, this necessitates identifying pathways for AI systems that, while contributing to safety, do not inherently carry the significant risks that the high-risk category is designed to address. The goal is to avoid a situation where the regulatory framework inadvertently discourages the adoption of AI that could prevent accidents and save lives.

Exploring Exemptions and Nuanced Interpretations within the AI Act

The primary mechanism for an AI system listed in Annex III to avoid high-risk classification is Article 6(3) of the AI Act. This provision allows for a derogation if the system "does not pose a significant risk of harm to the health, safety or fundamental rights of natural persons, including by not materially influencing the outcome of decision making." To qualify, the system must also meet one of several conditions related to its function:

- Performing a narrow procedural task: An example in an OSH context could be an AI system that automates the logging of safety equipment checks based on structured digital inputs from technicians. If the system merely records data according to predefined rules and does not make interpretive judgments or decisions affecting safety status, it might be considered to perform a narrow procedural task.
- Improving the result of a previously completed human activity: Consider an AI tool that reviews human-generated OSH incident reports for completeness or identifies potential inconsistencies against a checklist. If the AI flags these issues for human review and correction, and does not autonomously alter the report or determine its final assessment, it could be argued that it is improving a human activity without materially influencing the core decision (which remains with the human reviewer).
- Detecting decision-making patterns or deviations... not meant to replace or influence the previously completed human assessment, without proper human review: An AI system that analyzes patterns in safety inspection data over time to highlight anomalies or emerging trends (e.g., an increasing frequency of near-misses in a particular area) could fit this description. If its role is to bring these patterns to the attention of OSH professionals for their investigation and judgment, and human assessment remains the basis for any action, it may not be deemed high-risk.
- Performing a preparatory task to an assessment: An AI tool that collates relevant OSH

regulations, standards, and historical safety data for a specific workplace to assist a human OSH auditor in preparing for an audit could be seen as performing a preparatory task. If the AI organizes information but does not conduct the assessment itself or determine its outcome, it may qualify for the derogation.

Specific Categories for Potential Non-High-Risk Classification in OSH

Based on the provisions of the AI Act, several categories of OSH AI systems could potentially be classified as non-high-risk, provided they meet the stringent conditions of Article 6(3):

- AI for Education and Training in OSH (Article 4 considerations). Article 4 of the AI Act mandates that providers and deployers ensure a sufficient level of AI literacy among their staff and others dealing with AI systems. While this article focuses on competence and awareness, it does not directly classify AI systems used for training by risk level. AI systems used for general OSH awareness training, interactive simulations of non-critical tasks (e.g., practicing fire extinguisher use in a virtual environment), or providing access to safety manuals and procedures could potentially be considered non high-risk. This is plausible if the AI does not evaluate individuals in a way that determines their access to employment, promotion, or professional certification, as AI systems used for "evaluation of learning outcomes" or "steering the learning process" in education and vocational training are listed as high-risk under Annex III, point 3. The distinction would hinge on whether the AI is purely instructional versus evaluative with significant consequences.
  A critical consideration in this context is the potential of robust AI literacy programs, as mandated by Article 4, to serve as a significant risk mitigation measure. While not a direct path to declassifying an AI system, ensuring that workers and managers possess a thorough understanding of an OSH AI's capabilities, limitations, potential failure modes, and the context of its use can substantially reduce the actual risk posed by the system. If personnel are well-trained to use the AI safely, to recognize when its outputs might be erroneous, and to know when human intervention or override is necessary, this enhanced literacy could indirectly support an argument under Article 6(3) that the system, when deployed within such a competent human environment, poses a lower overall risk. This connects the procedural requirement of AI literacy initiatives and tools to a substantive reduction in operational risk.

- Information Systems Supporting Human Decision-Makers. This aligns with the conditions in Article 6(3) if the AI system's influence on the final decision is not material and human agency is preserved. Examples may include:
  - Risk evaluation support systems: AI that analyzes workplace data to identify potential OSH hazards or calculate preliminary risk scores, presenting these findings to a human expert who conducts the final risk assessment and determines mitigation strategies.
  - Compliance evaluation support systems: AI that scans OSH documentation or

procedures against regulatory checklists and flags potential areas of non-compliance for human review and verification.

- Alert systems monitoring variations in risks: AI that detects anomalies in environmental sensors (e.g., a sudden rise in temperature or gas concentration in a confined space) and alerts human personnel, who then assess the situation and decide on the appropriate response. The crucial factor for these systems to be considered non high-risk is that the AI genuinely *informs* or *supports* human judgment rather than *determining* the decision or action. The human decision-maker must retain the autonomy and capability to critically evaluate the AI's output and make an independent judgment. However, a common pitfall is the "human-in-the-loop" (HITL) fallacy. Simply having a human in the decision-making chain is not sufficient if that human is not genuinely empowered or equipped to override or critically assess the AI's recommendation. If the AI's output is presented with high authority, is too complex for the human to quickly verify, or if the human operator is subject to automation bias (an over-reliance on automated outputs) or undue time pressure, the AI may, in effect, be materially influencing the decision despite the nominal presence of a human. Regulators are likely to scrutinize the *meaningfulness* and *effectiveness* of human oversight, not just its procedural existence. Therefore, a robust HITL mechanism that ensures genuine human control is essential for such systems to potentially qualify as non-high-risk.

- AI Systems Aimed Purely at Risk Reduction without Introducing New Risks.
  This is a compelling argument from an OSH perspective. An AI system designed, for example, as an emergency stop mechanism that reliably halts machinery in a dangerous situation, or one that automatically reduces a system's operational parameters to a safer state upon detecting a hazard, is fundamentally risk-reducing. If such a system is highly reliable, its failure modes are well-understood and lead to a safe state (fail-safe design), and it does not introduce significant new AI-specific risks (e.g., from complex, unpredictable algorithms or biased data), one could argue that it does not pose a "significant risk of harm" in the sense targeted by the AI Act's high-risk provisions.
  However, this line of reasoning encounters a potential challenge with Article 6(1) of the AI Act. If such a risk-reducing AI is considered a "safety component" of a product (e.g., a machine) that falls under the scope of EU harmonisation legislation listed in Annex I (like the Machinery Directive) and that product requires a third-party conformity assessment for safety, then the AI system is presumptively classified as high-risk. The argument for non high-risk status would then need to be very nuanced: asserting that despite being a safety component, the AI-specific aspects of the system do not introduce the types of complex risks (e.g., opacity, unpredictability, bias) that the AI Act's high-risk regime is primarily designed to mitigate. The focus would be on demonstrating that its AI characteristics are simple, verifiable, and exceptionally reliable, distinguishing it from more complex

predictive or adaptive AI systems. This "safety component conundrum" highlights the tension between the functional safety criticality of a device and the specific risks introduced by the AI technology embedded within it. The AI Act's primary concern is with risks arising from AI, not necessarily all risks associated with systems that happen to use AI.

Successfully navigating these pathways to a non-high-risk classification requires meticulous documentation of the system's design, functionality, intended use, and a robust assessment demonstrating compliance with the conditions of Article 6(3). The forthcoming EC guidelines on this article will be critical in providing further clarity and practical examples.

6. The Path Forward: Fostering Ethical and Effective AI in OSH

The integration of AI into Occupational Safety and Health presents a journey that requires careful navigation. The ultimate goal is to harness AI's significant potential to create safer and healthier workplaces, but this must be achieved in a manner that respects fundamental rights, upholds ethical principles, and builds trust among all stakeholders. Achieving this necessitates a nuanced regulatory approach that avoids stifling innovation while ensuring robust governance.

The Imperative of Balancing Innovation with Fundamental Rights and Safety

The core message of this paper resonates throughout this challenge: leveraging AI's capabilities while safeguarding human rights and ethical principles is paramount. Overly restrictive regulations can indeed diminish AI’s effectiveness in protecting workers by hindering the development and adoption of beneficial safety technologies. Conversely, insufficient governance could pave the way for misuse, ethical breaches, and an erosion of worker trust, ultimately undermining the positive contributions AI could make. This balance is not a static point but a dynamic equilibrium that must be continuously sought and adjusted.

Arguments for a Nuanced Regulatory Approach

A regulatory framework that successfully fosters ethical and effective AI in OSH should be built on several key pillars:

- Transparency: This extends beyond the operational transparency required for high-risk AI systems (e.g., explainability of decisions). It also encompasses transparency in the regulatory process itself, with clear, accessible, and predictable guidelines and interpretations of the AI Act. For deployers, typically employers in the OSH context, transparency towards workers regarding the use of AI systems, their purpose, how they function, and what data they collect is crucial for building trust and ensuring compliance with data protection principles.
- Accountability: Clear lines of responsibility must be established for the entire lifecycle of an AI OSH system—from its design and development through to its deployment, operation,

and decommissioning. This includes defining who is accountable if an AI system errs, causes harm, or leads to discriminatory outcomes. Effective mechanisms for redress for affected individuals are also a vital component of accountability.

- Inclusivity: The development of AI solutions for OSH, as well as the shaping of regulatory interpretations and best practices, must involve a broad range of stakeholders. This includes AI developers, OSH professionals, employers, and, critically, workers and their representatives. The EU's OSH Framework Directive (89/391/EEC) already mandates the information and consultation of workers on the introduction of new technologies and their OSH implications, a principle that must be strongly upheld in the age of AI.

The Role of Ongoing Dialogue, Common Practices, and Guidelines

An ongoing dialogue between professionals and stakeholders is required for further investigation and the development of common practices. This dialogue should lead to the proposal of good practice guidelines, recommendations, and common interpretation frameworks.

- Interpreting Ambiguities: The AI Act, like any novel and comprehensive piece of legislation, contains terms and provisions that will require ongoing interpretation in specific contexts, such as OSH. Stakeholder dialogue can help develop consensus on how to apply these in practice.
- Developing Sector-Specific Guidance: Generic AI regulations may not fully address the unique challenges and opportunities of AI in OSH. Sector-specific guidelines, informed by expert input, can provide more tailored and practical advice.
- Fostering Best Practices: The sharing of experiences and the collective development of best practices can accelerate the adoption of safe and effective AI OSH solutions.
- Supporting Standardization: The AI Act itself anticipates the development of codes of conduct and harmonized standards, which can provide a presumption of conformity with its requirements. Bodies like EU-OSHA, which already conduct foresight exercises on digitalization and OSH, and national OSH institutes, can play a vital role in facilitating this dialogue and contributing to the development of such standards and guidelines. The European Commission's planned guidelines, for instance on the practical implementation of Article 6(3) concerning high-risk derogation, will be an important initial step in this process.

However, for such dialogue and guideline development to be truly effective, it must be substantive and lead to tangible improvements in safety and worker rights. There is a potential risk that these efforts could become performative, amounting to "compliance theatre," if they do not result in genuine commitments from all parties, particularly industry, to prioritize safety and ethical considerations over mere procedural compliance or reputational management. Meaningful dialogue requires a genuine willingness to share power and to ensure that the resulting frameworks are robust and enforceable.

Furthermore, the rapid evolution of AI technology means that governance frameworks for AI in OSH cannot be static. The "path forward" must incorporate mechanisms for continuous learning, active monitoring of the real-world impacts of AI in OSH, and the adaptation of regulations and guidelines as the technology and its applications mature. This implies a more agile and adaptive governance approach than has traditionally been associated with OSH legislation. This could involve, for example, regulatory sandboxes specifically for OSH AI innovations, regular reviews and updates to Annex III of the AI Act based on new evidence of risk or benefit, and proactive foresight studies to anticipate future developments.

A critical element in operationalizing this balance between innovation and safety will be the work of standardization bodies (e.g., CEN, CENELEC, ISO, IEC). Abstract principles of transparency, accountability, robustness, and human oversight need to be translated into concrete, verifiable technical and procedural standards for OSH AI systems. These harmonized standards can provide a practical route for providers to demonstrate conformity with the AI Act's requirements. The development of such standards must involve the active participation of OSH experts, AI ethicists, and worker representatives to ensure that they genuinely promote safety and ethical deployment, rather than primarily reflecting narrow industry interests or becoming a barrier to innovation for smaller players.

7. Conclusion: Striking the Balance for Trust and Sustainable Progress in AI for OSH

The journey of integrating Artificial Intelligence into Occupational Safety and Health is at a critical juncture. The transformative potential of AI to prevent accidents, mitigate risks, and enhance worker well-being is undeniable. Yet, this promise is intertwined with the profound challenge of ensuring its development and deployment are safe, ethical, and respectful of fundamental human rights. The EU AI Act represents a landmark attempt to navigate this complex terrain, but its successful application in the OSH domain hinges on achieving a delicate and dynamic balance – a balance that is central to building trust and enabling sustainable progress.

Reiterating the Central Challenge

The core tension lies between the drive to leverage AI's innovative capabilities for OSH improvement and the imperative to implement robust regulatory frameworks that address potential harms. As explored, overly restrictive measures risk stifling the very innovation that could save lives, while insufficient governance could lead to misuse, erosion of worker trust, and unintended negative consequences. The AI Act's definitions, particularly the nuances of terms like "infer," "deduce," and "derive," and the broad scope of "influence," create ambiguities that need careful and consistent interpretation. The default classification of many OSH AI systems as "high-risk," with its attendant compliance burdens, presents a significant hurdle that must be

navigated thoughtfully to avoid discouraging beneficial applications.

The Path to Trustworthy AI in OSH

Trust in AI systems, especially those impacting safety and health, is not an automatic byproduct of technological advancement. It must be meticulously earned and continually reinforced. This requires a multifaceted approach encompassing:

- Transparent Design and Operation: AI systems used in OSH must be understandable, with their capabilities, limitations, and decision-making processes (where feasible) made clear to those who develop, deploy, and are affected by them.
- Robust Validation and Verification: Rigorous testing and validation methodologies are essential to ensure that OSH AI systems perform reliably and safely in real-world conditions, and that they are free from harmful biases.
- Meaningful Human Oversight: While AI can augment human capabilities, effective human oversight remains crucial, particularly for high-risk applications. This means ensuring that humans can understand, intervene, and ultimately control AI systems when necessary.
- Clear Accountability Frameworks: Establishing who is responsible when AI systems err or cause harm is vital for building confidence and ensuring recourse.
- Inclusive Governance: Engaging all relevant stakeholders—developers, employers, OSH professionals, workers and their representatives, and regulators—in shaping the development, deployment, and governance of AI in OSH is fundamental.

The long-term economic implications of failing to strike the right balance are considerable. An overly restrictive regulatory environment could see the EU lag in the development and deployment of innovative OSH AI, potentially slowing safety improvements and creating economic disadvantages. Conversely, a permissive approach that leads to significant AI-related incidents, widespread bias, or a breakdown in worker trust could result in substantial societal and economic costs, including increased healthcare expenditures, legal liabilities, loss of productivity, and public resistance to technology adoption. Thus, achieving a well-calibrated balance is not merely an ethical or legal imperative but also an economic necessity for sustainable technological progress in OSH.

Call to Action for OSH Stakeholders

Realizing the promise of AI in OSH while navigating its challenges requires concerted effort from all parties:

- Regulators: Must provide clear, practical, and timely guidance on the AI Act's application to OSH. This includes clarifying ambiguous terms and the conditions for derogation from high-risk classification. Fostering international cooperation and ensuring that regulations can adapt to the rapid pace of technological evolution are also key.

- AI Developers and Providers: Should embrace "safety by design" and "ethics by design" principles from the outset. This involves investing in robust testing, validation, and bias mitigation, and being transparent about their systems' capabilities, limitations, and data usage.
- Employers/Deployers: Are responsible for conducting thorough risk assessments before deploying AI systems in the workplace. They must ensure meaningful consultation with workers and their representatives, as mandated by OSH law and the AI Act's principles, and invest in comprehensive AI literacy programs for their workforce as per Article 4 of the AI Act. Implementing robust human oversight mechanisms for AI systems is also critical.
- OSH Professionals: Need to develop expertise in AI-related risks and benefits to effectively advise organizations. Their involvement in developing best practices, standards, and risk assessment methodologies for AI in OSH is indispensable.
- Workers and Their Representatives: Must actively engage in consultations regarding the introduction and use of AI systems in the workplace. Advocating for their rights to safety, privacy, and fair treatment in the context of AI deployment is crucial.
- Researchers: Should continue to investigate the multifaceted impacts of AI on OSH, develop innovative methodologies for creating safe and reliable AI, and contribute to the evolution of ethical frameworks and governance models.

Final Thought

The EU AI Act, as a pioneering piece of legislation, will undoubtedly shape the trajectory of AI governance globally. Its success in fostering a balanced ecosystem for AI in Occupational Safety and Health will not only benefit workers within the Union but also set a precedent for other jurisdictions. A balanced, human-centric approach—one that prioritizes worker well-being, upholds fundamental rights, and fosters responsible innovation—is essential. Only through such an approach can we unlock the full potential of artificial intelligence to create truly safer and healthier workplaces for all, ensuring that AI systems remain powerful tools for enhancing OSH while steadfastly respecting societal values and human dignity. The challenge of balance is indeed central, not just to compliance, but to building trust and enabling the sustainable progress that will define the future of work in the age of AI.