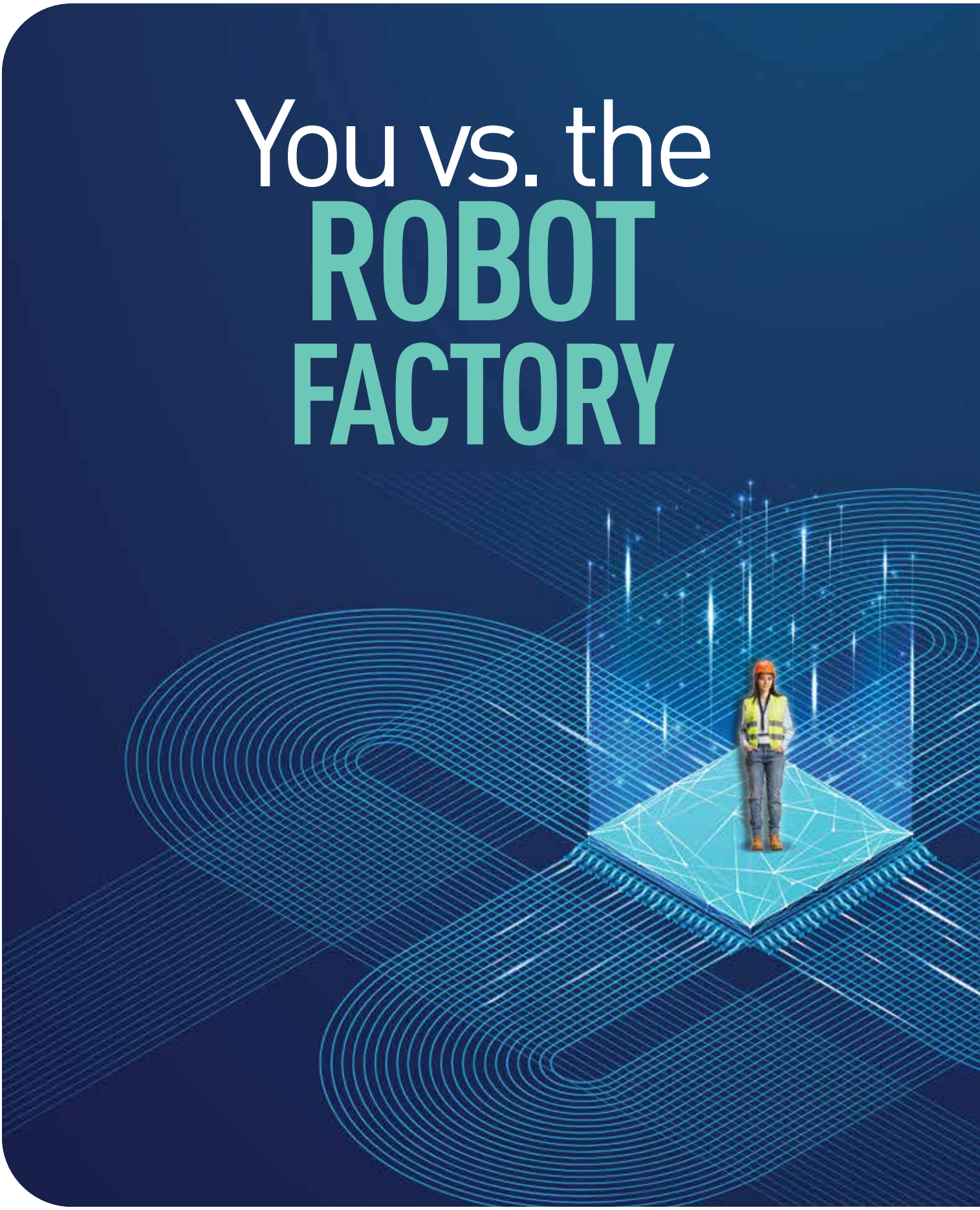
# You vs. the ROBOT FACTORY



# Some Principles for Understanding AI Hazards in the Workplace

**BY JOHN P. SADOWSKI**

What might artificial intelligence (AI) in the workplace look like in the near future? As an example, imagine an algorithmically controlled chemical factory. Perhaps it gathers data from distributed and wearable sensors and machine vision systems. Perhaps it controls chemical reaction parameters and hazardous machinery, operates the factory's ventilation system, and controls self-driving robotic vehicles. Perhaps it directs workers through messages on their smartphones to do certain tasks, and maybe even implements a chatbot to answer questions and issue clarifications to workers. How should an occupational and environmental health and safety (OEHS) professional go about hazard identification in this factory? What controls should they recommend implementing?

As AI-enabled systems are increasingly adopted in the workplace, OEHS practitioners and researchers must be prepared to consider their effects on worker health and safety. While AI may at first seem radically different from other potentially hazardous aspects of the workplace, it can indeed be understood using established OEHS principles. Further, although people might not initially think to ask an OEHS professional about AI safety, their presence throughout all industrial sectors gives them a responsibility and an opportunity to have a significant impact through education and assessment activities. This would allow end-users to improve their own practices, as well as request that software vendors consider health and safety impacts in their software design, without requiring

a wait for formal regulatory action. It would also foster the good governance and trust that would encourage adoption of AI in contexts where it is warranted.

The OEHS community has successfully dealt with emerging technologies before. Nanotechnology is one example where OEHS researchers, including those at NIOSH, have benefited from putting aside hype about extreme positive or negative implications and realizing that nanomaterials are, after all, chemicals. This enables an extension of existing chemical hazard frameworks to cover the novel properties of nanomaterials. Similarly, broadly established hazard identification and exposure assessment methodologies do not have to be discarded when considering AI systems in the workplace; these methods can be extended and adapted to cover the novel characteristics of AI systems. In fact, this framework was originally inspired by the question of what an AI parallel would look like for NIOSH's "10 Critical Nanotechnology Topic Areas" (**bit.ly/nano10topics**).

Nevertheless, OEHS research into AI is now in its infancy. While a few research groups are taking an OEHS-specific approach to AI safety, in the wider world much of the current discourse is often contextualized in terms of either ethics or macroeconomics. While these are important approaches to AI safety, ethics is a subdiscipline of philosophy and can sometimes be too abstract or forward-looking to provide the practical guidance needed for OEHS field activities, while macroeconomics focuses on impacts across sectors or regions rather than within single workplaces. A science of industrial hygiene for algorithms, which we might call "algorithmic hygiene," is needed, explicitly linking the characteristics of algorithmic systems to health and safety outcomes. This linking would guide research questions and provide a scientific basis for practical, actionable guidance for individual end-users, developers, and policymakers.

## FIVE PRINCIPLES FOR ALGORITHMIC HYGIENE

The framework described here identifies several core characteristics of algorithmic systems and links them to existing, well-known categories of occupational hazards and controls (see Figure 1). Devised through an informal literature review and conversations with multiple OEHS researchers and practitioners, the framework is intended as a starting point for discussions to guide field work on hazard identification and control, and to formulate research questions on the OEHS impacts of algorithms.

The following principles are proposed as a basis for conceptualizing what a science of algorithmic hygiene might be:

- **The term "trained algorithm," meaning use of a training data set to create an algorithm for a system's output or behavior, is a better conceptualization than "artificial intelligence."** The latter term is problematic due to its vagueness, broadly referring to any software that is perceived to mimic human capabilities, whereas the former is a practical description of a process. The OEHS professional's task is to frame these issues methodically in terms of risk assessment and

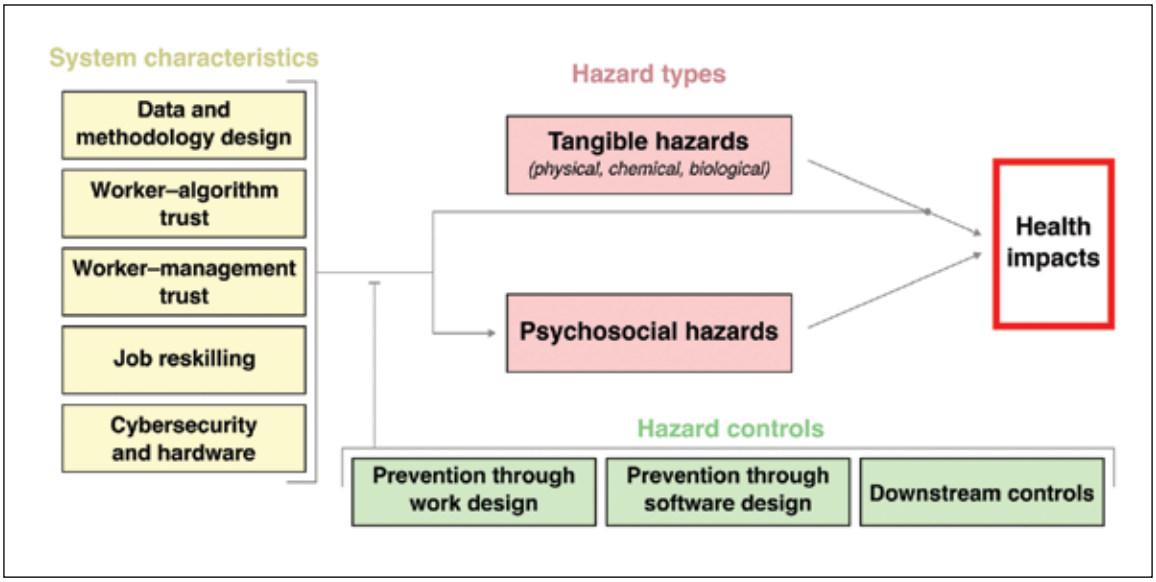



**Figure 1.** A framework for algorithmic hygiene that illustrates principles discussed in this article. Several categories of AI system characteristics can each cause any of the standard occupational hazard types. Because algorithms are software with no physical substance, they can mediate but not directly create any tangible hazards; they can, however, directly create psychosocial hazards. Exposure assessment can operate using familiar, existing "hazard-side" tools and methods already used for each hazard type, and more novel "system-side" methods that assess the AI system characteristics themselves. Hazard controls are divided between work-design controls that can be carried out within the end-user's organization, and software-design controls that must be applied by software developers who may work for a vendor. Figure copyright 2025 by John P. Sadowski.

management, which is in contrast to AI's historic origins of posing basic-science and philosophical questions about computation and cognition.

- **Algorithms must be distinguished from the physical platforms they may be embedded in, but platforms can help identify trained algorithms being used in a workplace.** Software capabilities using trained algorithms include generative platforms such as chatbots and image generators, machine vision, speech recognition, and people analytics for hiring and performance assessment. They may be embedded in physical platforms including collaborative robots, self-driving vehicles including forklifts and drones, smart HVAC systems, wearable or distributed sensors, or simply a computer interface. Note, however, that many of these platforms can be used without trained algorithms; a self-driving forklift, for example, could use rigidly programmed algorithms.
- **Algorithms are software with no physical substance; while they cannot directly create any new "tangible" hazards, they do alter the risk profile of physical platforms or substances they control or interact with.** Tangible hazards is an umbrella term for physical, chemical, and biological hazards. For example, a chemical exposure, an electric shock hazard from machinery, or a collision with a forklift can all occur in a workplace with no algorithms. When trained algorithms respectively control an HVAC system, instruct a worker on how to repair the machinery, or autonomously drive the forklift, the algorithms mediate but do not create tangible hazards. By contrast, algorithms can directly cause psychosocial hazards by changing work organization, skills, and relationships, leading to stress and its downstream health impacts.
- **Established "hazard-side" exposure assessment methods that are familiar to OEHS professionals can still be used even without AI subject matter expertise, although "system-side" methods will enhance the assessment.** OEHS professionals can still use familiar tools and methods that have been established for each hazard type, which can discern the impact of introducing algorithmic systems even if their characteristics are not known. Assessing the impact of specific system characteristics of algorithms may be more novel and challenging from an OEHS perspective, but research into their effects on health outcomes will be vital in establishing methods for practitioners and end-users.
- **There is a distinction between "prevention through work design" hazard controls that can be carried out within the end-user's organization and "prevention through software design" controls that must be applied by the software developers.** While these developers may be in-house, they are more often at an external vendor firm producing either a bespoke or off-the-shelf product, which can make communication a challenge. These prevention-through-design principles are paramount in designing hazard controls, as distinguished from less-desirable "downstream" controls implemented after a system is in use.


Hazards can arise from applying an existing algorithm to another problem with different requirements.

### SOME CATEGORIES OF SYSTEM CHARACTERISTICS

As mentioned, trained algorithms have several categories of inherent characteristics that may affect hazards of each type. While more research is needed to devise a detailed taxonomy of these system characteristics, a few broad, overlapping categories can be discerned.

*Data and methodology design* can cause hazards through use of non-representative or poorly assembled data, or through flawed methodology including overfitting or failing to account for outliers. Hazards can also arise from applying an existing algorithm to another problem with different requirements. Algorithmic bias is a subclass of poor data design that arises from not having a representative sample of people in the training set, or by inadvertently training algorithms on data that

Vanit Janthra/Getty Images

## Lack of a human-understandable explanation for an algorithm's behavior can influence the risk profile.

reflect past discriminatory practices. These can lead to hazardous unpredictable behavior, especially if a situation is encountered that was not part of the training dataset.

*Worker–algorithm trust* is sometimes posed in the AI safety literature as a behavioral question of whether a human chooses to override an algorithm's behavior or recommendation. By contrast, in research psychology and sociology, trust encompasses a rich collection of concepts regarding how one party acts in the interests of another. A worker may over-trust by not halting an adverse behavior or recommendation from a trained algorithm, or under-trust by ignoring or taking control from a trained algorithm's correct behavior. Lack of a human-understandable explanation for an algorithm's behavior or recommendation can influence the risk profile. Hazards can also arise from workers anthropomorphizing or becoming emotionally attached to robots, chatbots, and other systems, which may be either inadvertent or intended by the designers of algorithmic systems.

*Worker–management trust* is important because management may have greater control over or access to the algorithm, the training data set, or additional data generated from workers. This information asymmetry may lead to micromanagement, loss of worker autonomy, a perception of surveillance, or the assigning of blame to a human operator instead of system developers.

*Job reskilling* refers to changes in the tasks and skills required of workers. Reskilling requires retraining, flexibility, and openness to change from workers. Alternatively, there may be deskilling of some professions if an algorithm takes over functions previously performed by humans. Either can contribute to job precarity. Similarly, there is risk of work intensification through being forced to work at an algorithm's pace, or deintensification if situations requiring the worker to take control occur rarely. These are often approached in the literature from a macroeconomic viewpoint, but this framing can miss health hazards on the level of individual workers.

*Cybersecurity and hardware malfunctions* are typical problems for all computer systems. Data protection concerns can arise from internal misuse of data by management or employees, or from hackers stealing

hirun/Getty Images

data or maliciously altering the output or behavior of algorithms. A complication is that these problems may be dealt with as threats to the organization as a whole rather than to the health of individual workers. Hazards can also arise from routine hardware and sensor malfunctions.

## APPLYING THE PRINCIPLES

To make these principles more concrete, let's consider some broad observations about three example systems using trained algorithms that could be used in the previously described robot factory: a self-driving forklift using a machine vision system to avoid collisions, a smart HVAC system using distributed chemical sensors and a machine vision system to recognize the presence of people in each room, and an algorithm-driven management system that issues instructions to workers via a chatbot on a smartphone.

First consider how they might interact with tangible hazards. The self-driving forklift should be considered in the context of decades of extensive research on collisions involving hard-coded industrial robots; when their software is replaced with a trained algorithm, the physical robot still presents the same collision hazards, but the risk profile of a collision will change and will require different controls. Similarly, the smart HVAC system cannot itself directly create any new chemical hazards, but it could increase or decrease the risk of hazardous materials already present traveling through the building. Lastly, the algorithm-driven management system does not have a physical component at all, but it can still lead to tangible hazards, as it may give an instruction that results in a physical injury or chemical exposure through a worker's interactions with existing hazardous machinery or substances. For all of these, the novelty is not the hazards themselves but the new system characteristics that affect the hazards.

Trained algorithms more directly present psychosocial hazards, which stem from changes to work organization that can cause stress and downstream health impacts. For each of the three systems, examples include unexpected or unexplained behavior, lack of transparency on how machine vision data is being used and protected, and the need for retraining to operate a more complex system. Fear of a physical injury or chemical exposure caused by an algorithm's behavior is also a psychosocial hazard, even if the injury does not actually occur.

For exposure assessment, established and familiar "hazard-side" OEHS methods can be used. For example, inspecting incident reports and workers' compensation histories is still useful for assessing collision risks of a self-driving forklift. A smart HVAC system can still be compared to a traditional one using standard indoor air quality measurements, even without knowing the inner workings of the algorithm. The algorithm-directed

management system can still be assessed by inspecting how workers carry out work and interact with each other, having conversations with workers, and conducting surveys.

Of course, exposure assessment is enhanced by making use of the "system-side" aspects. Each of the system characteristics implies questions that the OEHS professional may ask to discover certain hazardous conditions during an assessment. To focus on the smart HVAC system specifically, an assessment may reveal conditions such as: the algorithm was trained in the dry climate of California but malfunctions because it is being used in the humid climate of Florida; the machine vision system was trained on an unrepresentative sample of people and does not recognize some workers; workers override unexpected and unexplained behavior that is actually mitigating a hazard; the chemical sensors fail due to normal fouling; or security flaws allow a hacker to alter the system to create a chemical hazard.

**There may be cases where an algorithm is used for its perceived "cool factor" but does not actually increase productivity or safety.**

Some hazard controls can be implemented through in-house work design, while others must be done through software design. For the self-driving forklift, work-design hazard controls may include proper training of workers on the forklift's behavior even if they are not operating it and transparency about how the footage from the machine vision cameras is to be used. Software-design controls may include the end-user guiding selection of the training and test data sets to ensure alignment with the conditions at that facility, and having the software provide human-understandable explanations for the forklift's behavior.

Elimination should not be neglected as a possible control if the trained algorithm is not necessary or does not perform better than hard-coded software or humans. There may be cases where an algorithm is used for its perceived "cool factor" but does not actually increase productivity or safety. For example, if the smart HVAC's machine vision system is used to increase ventilation or shut off hazardous machinery when a human enters a restricted area, one may ask whether it may be safer to use a laser curtain, or a door with a switch.

### A CALL TO ACTION

Although there is not yet any official guidance on workplace AI safety, OEHS professionals can take steps now to be aware of trained algorithms' effects on worker health and safety. Even asking during an inspection where AI is being used and what consideration has been given to its health and safety effects can have immediate impact. Educating workers and buyers of AI-enabled systems about their health and safety effects is even more impactful; vendors are sensitive to the requests of buyers, who can ask them what has been done to make the AI safe and free of negative effects on worker health.

Ultimately, for OEHS practitioners and researchers, the challenge is to determine in a structured way how each system characteristic interacts with each hazard and control type in a workplace. Everything from regulation to standards to guidance is downstream of the science of how hazards cause health impacts. OEHS researchers and practitioners should focus on advancing the science of algorithmic hygiene from its infancy into a solid body of evidence that can drive safe and healthy use of AI in workplaces throughout all sectors of the economy. S

**JOHN P. SADOWSKI, PhD,** is a scientist specializing in the health and safety impacts of emerging technologies.
This work was funded through a contract from NIOSH via Advanced Technologies & Laboratories International. The findings and conclusions in this report are those of the author and do not necessarily represent the official position of NIOSH.

Send feedback to **synergist@aiha.org**.

## RESOURCES

*American Journal of Industrial Medicine*: "Managing Workplace AI Risks and the Future of Work," **bit.ly/howard2411** (November 2024).

European Agency for Safety and Health at Work: "OSH and the Future of Work: Benefits and Risks of Artificial Intelligence Tools in Workplaces," **bit.ly/moore1905** (May 2019).

*International Journal of Environmental Research and Public Health*: "Occupational Safety and Health Equity Impacts of Artificial Intelligence: A Scoping Review," **bit.ly/fisher2307** (July 2023).

*International Journal of Environmental Research and Public Health*: "REDECA: A Novel Framework to Review Artificial Intelligence and Its Applications in Occupational Safety and Health," **bit.ly/pishgar2107** (July 2021).

NIOSH: "10 Critical Nanotechnology Topic Areas: An Overview," **bit.ly/nano10topics** (February 2024).

*Patterns*: "Calibrating Workers' Trust in Intelligent Automated Systems," **bit.ly/lucas2409** (September 2024).

*Proceedings of the National Academy of Sciences*: "Toward Understanding the Impact of Artificial Intelligence on Labor," **bit.ly/frank1904** (April 2019).

*Safety Science*: "Occupational Health and Safety in the Industry 4.0 Era: A Cause for Major Concern?" **bit.ly/badri1811** (November 2018).